scientific data

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of *Mylabris sibirica* Fischer von Waldheim, 1823 (Coleoptera, Meloidae)

Chenhui Shen[1], Guofeng Yang[1], Min Tang[1], Xiaofei Li[1], Li Zhu[1], Wei Li[2], Lin Jin[3], Pan Deng[4], Huanhuan Zhang[5], Qing Zhai[6], Gang Wu[1] ✉ & Xiaohong Yan[1] ✉

Mylabris sibirica **is a hypermetamorphic insect that primarily feeds on oilseed rape during the adult stage. However, the limited availability of genomic resources hinders our understanding of the gene function, medical use, and ecological adaptation in *M. sibirica*. Here, a high-quality chromosome-level genome of *M. sibirica* was generated by PacBio, Illumina, and Hi-C technologies. Its genome size was 138.45 Mb, with a scaffold N50 of 13.84 Mb and 99.85% (138.25 Mb) of the assembly anchors onto 10 pseudo-chromosomes. BUSCO analysis showed this genome assembly had a high-level completeness of 100% (n = 1,367), containing 1,358 (99.4%) single-copy BUSCOs and 8 (0.6%) duplicated BUSCOs. In addition, a total of 11,687 protein-coding genes and 35.46% (49.10 Mb) repetitive elements were identified. The high-quality genome assembly offers valuable genomic resources for exploring gene function, medical use, and ecology.**

Background & Summary

Brassica napus is one of the most important oil crops around the globe, and rapeseed oil is among the most nutritious edible vegetable oils[1]. Insect pests remain a significant challenge to rapeseed yield, such as *Phyllotreta striolata*, *Plutella xylostella*, and *Mylabris sibirica*[2–4]. Chemical insecticides are the primary method of managing pests. However, improper use and overuse of pesticides result in resistance development, environmental contamination, and negative impacts on beneficial insects like honey bees, parasites, and insect predators[5,6]. Exploring innovative methods based on RNA interference (RNAi) will have promising applications in pest management[3,7,8]. At present, the genomes of *P. xylostella* and *P. striolatah* have been published[9], but there are no reports on the genome of *M. sibirica*, which limits the development of novel technologies for controlling *M. sibirica*.

M. sibirica (Coleoptera, Meloidae) is a hypermetamorphic pest that poses significant threats to oilseed rape production, resulting in substantial economic losses[4]. The adults of this species primarily feed on flowers of *B. napus*[4]. Previous studies on *M. sibirica* have traditionally placed a strong emphasis on classification[10,11], phylogenetics[12], and medical value[4]. However, there is currently no research on gene functions in *M. sibirica*, except for one study that selected optimal reference genes[13]. Therefore, whether for studying molecular mechanisms of development, reproduction, and physiology, or screening novel target genes, assembling the high-quality genome of *M. sibirica* is imperative.

[1]Key Laboratory of Agricultural Genetically Modified Organisms Traceability, Ministry of Agriculture and Rural Affairs, Oil Crops Research Institute of Chinese Academy of Agricultural Science/Supervision and Test Center (Wuhan) for Plant Ecological Environment Safety, Ministry of Agriculture and Rural Affairs, Wuhan, 430062, China. [2]Northern Propagation Experiment Station, Center for Science and Technology Dissemination and Industrial Development, Oil Crops Research Institute of Chinese Academy of Agricultural Science, Wuhan, 430062, China. [3]Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, 210095, China. [4]Institute of Leisure Agriculture, Jiangsu Academy of Agricultural Sciences, Nanjing, 210014, China. [5]Institute of Vegetable, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, 850032, China. [6]College of Plant Protection, Henan Agricultural University, Zhengzhou, 450002, China. ✉e-mail: wugang@caas.cn; yanxiaohong@caas.cn

Libraries	Number	Total length (bp)	Average length (bp)	N50 length (kb)
Survey	185,232,438	27,784,865,700	150	—
RNA	119,695,954	17,954,393,100	150	—
Hi-C	139,254,850	20,888,227,500	150	—
PacBio Hifi	2,501,010	15,444,413,507	6,175.27	6.75

Table 1. Sequencing raw data of the *Mylabris sibirica* assembly.

Chromosomes	length (bp)	PacBio HiFi sequencing coverage (X)
Msib_ChrX	8,466,837	18.48
Msib_Chr1	13,496,263	34.57
Msib_Chr2	16,795,200	34.98
Msib_Chr3	13,750,752	34.34
Msib_Chr4	15,291,861	34.03
Msib_Chr5	13,838,200	33.99
Msib_Chr6	15,138,339	34.46
Msib_Chr7	16,384,000	34.09
Msib_Chr8	13,740,500	34.79
Msib_Chr9	11,346,979	34.22

Table 2. Chromosome length and sequencing depth statistics.

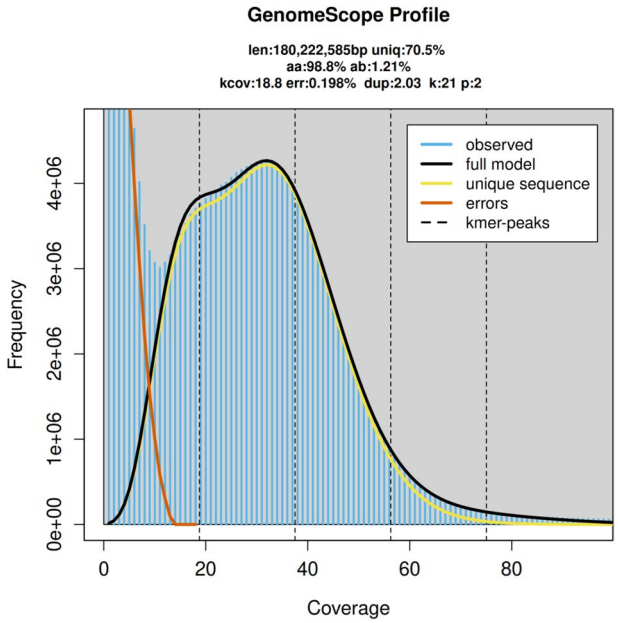

Fig. 1 Genome survey at 21-mer of *Mylabris sibirica* estimated by GenomeScope. The vertical dotted lines represent the peaks of different coverages for the heterozygous, the homozygous, and the duplicated sequences, separately.

Cantharidin, produced by *M. sibirica*, is a medicinal compound that has been found to have significant therapeutic effects on liver cancer, stomach cancer, esophageal cancer, breast cancer, and cervical cancer[14–16]. It has also a certain attractant effect on certain insects in the Diptera, specifically Ceratopogonidae, and Anthomyiidae[17]. Moreover, it is reported that cantharidin is a feeding deterrent to insects[18]. Hence, obtaining the high-quality genome of *M. sibirica* will have some recommendations for medical treatment, antifeedants, and attractants.

In this study, we generated a chromosome-level genome assembly of *M. sibirica* by utilizing PacBio, Illumina, and Hi-C technologies. We annotated protein-coding genes (PCGs) and repeat elements. This study provides the first genome assembly for *M. sibirica*, which will facilitate studies on gene function, medical use, and ecological adaptation and also greatly benefit efforts to manage the pest.

Content	Msib
Genome assembly	
Assembly size (bp)	138,452,083
Number of pseudo-chromosomes (sizes)	10 (138,248,931 bp)
Number of scaffolds/contigs	21/33
Longest scaffold/contig (Mb)	16.80/16.78
N50 scaffold/contig length (Mb)	13.84/11.35
GC content (%)	31.15
BUSCO completeness (%)[1]	100
S	99.4
D	0.6
F	0.0
M	0.0
Mapping ratio of BGI reads (%)	
BGI	95.19
HIFI	95.23
RNA-sr	97.79

Table 3. Genome assemblies results of *Mylabris sibirica*. BUSCO: Benchmarking Universal Single-Copy Orthologs; C, complete BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; M, missing BUSCOs.

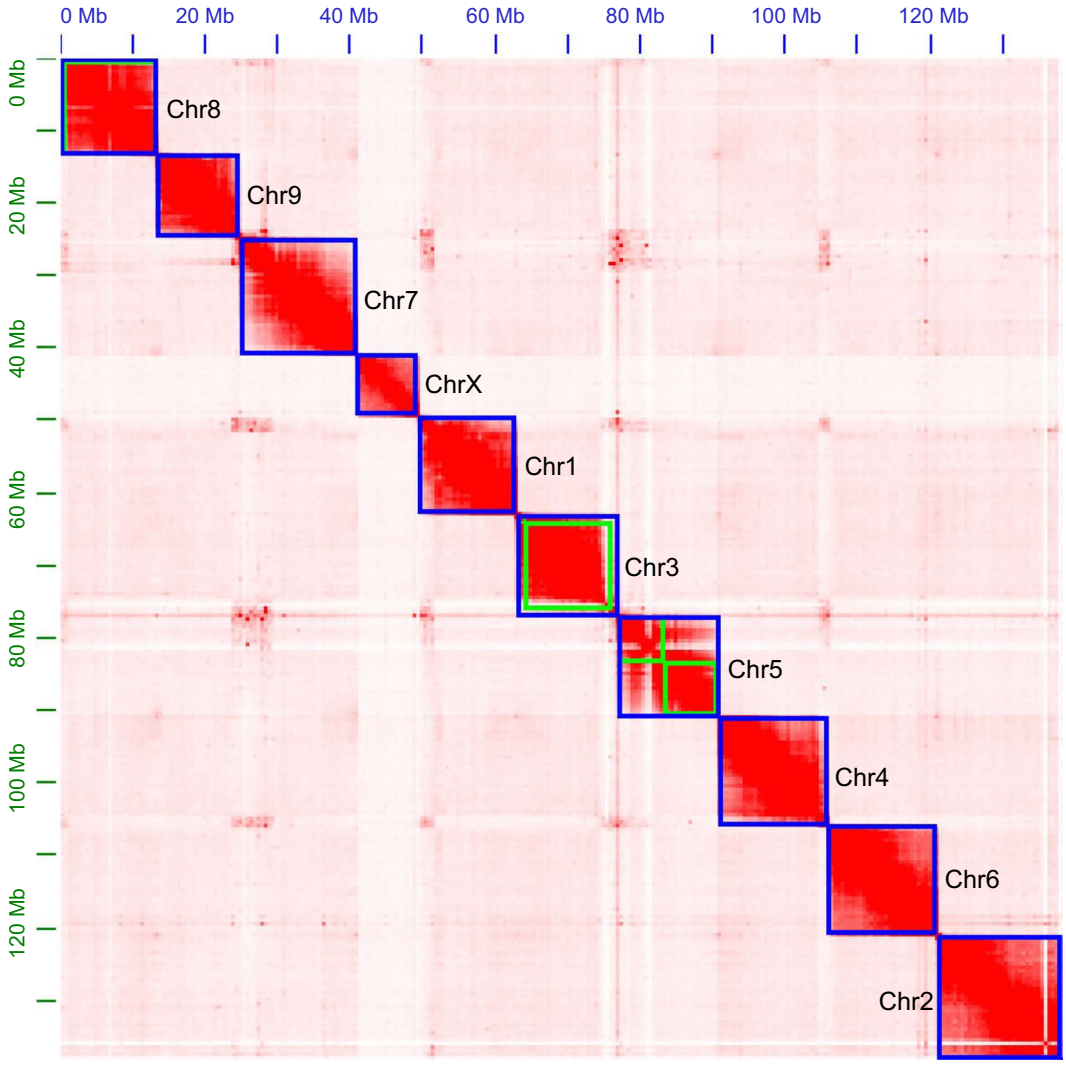

Fig. 2 Genomic heatmap. Genome-scale chromosome heatmap of *Mylabris sibirica*, with individual chromosomes outlined in blue.

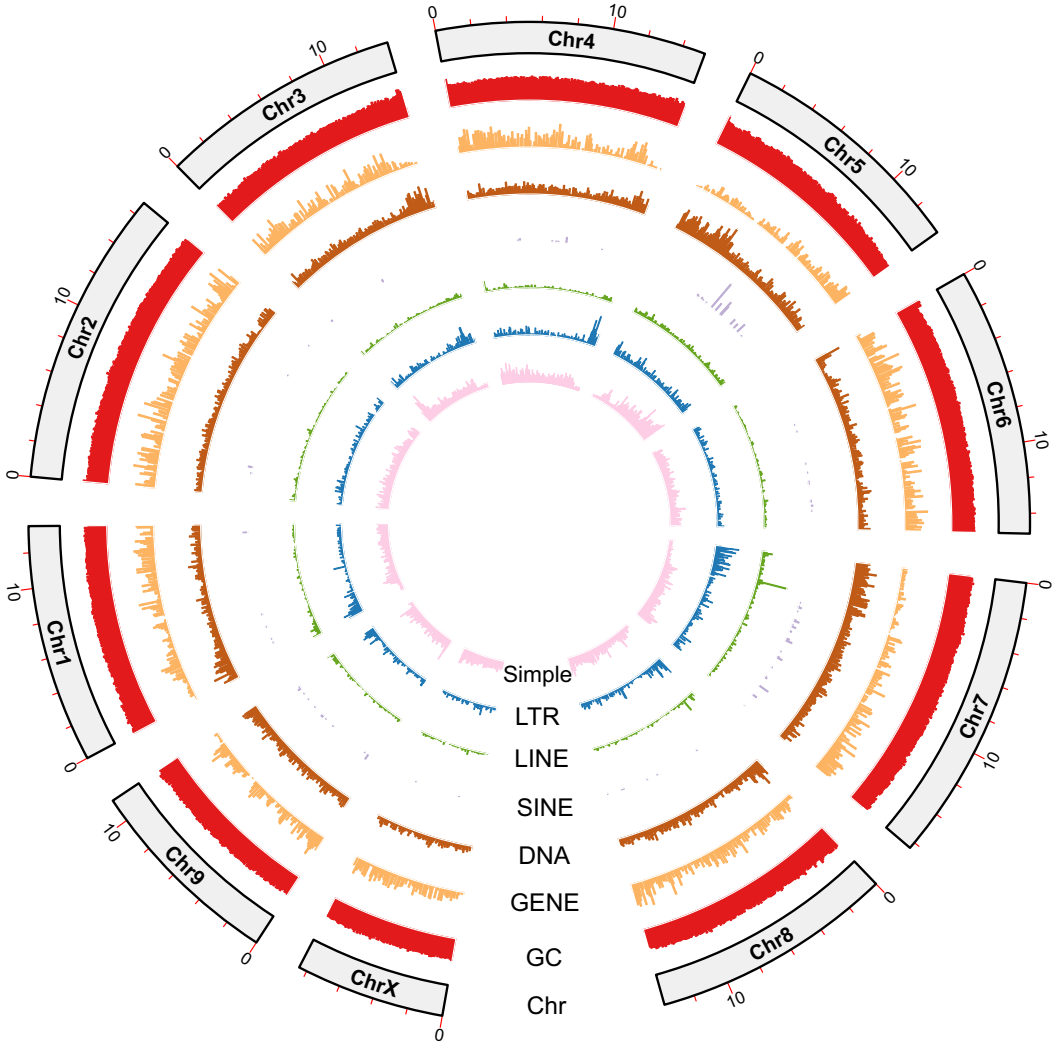

Fig. 3 Genomic features. Circos plot with a window size of 100 Kb. Each circle from inside to outside represents simple repeats, long terminal repeats (LTR), long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE), DNA, gene density, GC content, and chromosome length.

Methods

Sample collection and sequencing. *M. sibirica* adults used in the study were obtained in July 2023 from *B. napus* in Haidong City, Qinghai Province, China (coordinates: 36°31′10.47″N, 101°59′7.40″E). The adults were reared under suitable conditions (25 ± 2 °C, 16 h:8 h photoperiod, and 50 ~ 60% relative humidity) in the insectary for one week. The adult tissues (two males after removing the gut) were used to perform PacBio, Hi-C, genome survey, and transcriptome sequencing.

Following the manufacturer's methods, genomic DNA and RNA from samples were extracted using the FastPure® Blood/Cell/Tissue/Bacteria DNA Isolation Mini Kit (Vazyme Biotech Co., Ltd, Nanjing, China) and TRIzol reagent (YiFeiXue Tech, Nanjing, China), respectively. The concentration was determined by Qubit, and the final amount of DNA or RNA was not less than 1000 ng. High molecular weight (HMW) gDNA was sheared into 15 kb with the MegaruptorTM device (Diagenode, Liege, Belgium) and was enriched using the AMPurePB Beads. For transcriptome sequencing, we used the TruSeq DNA PCR-free kit to generate a short library with 150 bp paired-end reads and a 350 bp insert size. All short-read libraries were executed for sequencing by the Illumina NovaSeq6000 platform. The Hi-C experiment was performed on a previously published protocol[19], which contained numerous procedures: DNA cross-linking, chromatin digestion using the restriction enzyme MboI, end repair, and DNA purification. For PacBio HiFi mode, High-quality Pacbio HiFi reads were generated by Pacific Biosciences' (PacBio) SMRT (pbccs v6.4.0) sequencing technology (https://github.com/PacificBiosciences/ccs). All libraries were created and sequenced by Berry Genomics (Beijing, China). Finally, we obtained a total of 80.92. Gb of clean data, consisting of 15.44 Gb (112×) of PacBio, 26.64 Gb from Illumina (201×), 20.89 Gb from Hi-C (151×), and 17.95 Gb from transcriptome data (Table 1).

Genome survey. The raw Illumina data underwent quality control and removal of duplicate reads using BBTools v38.82[20]. The "bbduk.sh" tool was conducted to control specific quality, including removing base positions

	Msib
Structure annotation	
Number of protein-coding genes	11,687
Number of predicted protein sequences	17,756
Mean protein length (aa)	568
Mean gene length (bp)	5,959.6
Mean exon length (bp)	360.2
Mean CDS length (bp)	306.9
Mean intron length (bp)	881.0
BUSCO completeness (%)	99.6%
S	99.6%
D	31.5%
F	0.0%
M	0.4%
Repeat annotation	
DNA transposons (Mb)	5.89 (4.21%)
SINEs (kb)	34.195 (0.02%)
LINEs (Mb)	1.50 (1.07%)
LTRs (Mb)	2.23 (1.60%)
Simple repeats	3.35 (2.42%)
Low complexity	0.56 (0.40%)
Unclassified (Mb)	35.22 (25.44%)
ncRNA annotation	
Number of ncRNA	673
rRNA	187
miRNA	57
snRNA	34
tRNA	256

Table 4. Annotation statistics of the *Mylabris sibirica* genome assembly.

with quality scores below 20 (>Q20), filtering out sequences shorter than 15 bp, eliminating polymer A/G/C tails longer than 10 bp, and correcting overlapping paired reads. Subsequently, GenomeScope v2.0.1[21] was utilized for k-mer analysis, with a maximum k-mer depth threshold set at 1,000. The frequency of k-mers was assessed using "khist.sh" (BBTools), with a k-mer length of 21. A k-mer analysis showed that the genome assembly size of 180.22 Mb, with a heterozygosity of 1.21% and a repeat content proportion of approximately 29.52% (Fig. 1).

Genome assembly. The primary assembly of PacBio HiFi long reads was generated by utilizing Hifiasm v0.16.1[22]. Redundant regions were eliminated by Purge_Dups v1.2.5[23] with a 70% cut-off for identifying contigs as haplotig. We used Minimap2 v2.17[24] to remove redundancy for read mapping. Juicer v1.6.2[25] was used to align to the assembly for Hi-C reads. Subsequently, 3D-DNA v180922[26] was used to anchor the contigs onto the chromosomes. Moreover, we used Juicebox v1.11.08[26] to review and correct any errors to ensure accuracy.

According to the NCBI nucleotide and UniVec databases with a sequence identity of 0.8, possible contaminants were detected by MMseqs. 2 v11[27], which implements BLASTN-like searches. For further assessing vector contaminants, blastn (BLAST + v2.11.0[28]) was utilized against the UniVec database. We considered that sequences with over 90% hits in the aforementioned database likely contained contaminants. We used online BLASTN analysis in the NCBI nucleotide database to recheck sequences with over 80% hits. Then, any possible bacterial contamination from the assembled scaffolds was eliminated. According to previous publications[29,30], the sex-determination mechanism of this species is Xy_p. The final assembly was remapped using raw HiFi data, and each chromosome length was determined using MiniMap2[24], with the parameter "-ax map-hifi", to identify the autosomes and sex chromosomes. After that, chromosomal coverage was computed by dividing raw data, based on chromosome length using SAMtools v1.9[31], with the parameter "flagstat". The X chromosome was distinguished from other chromosomes by having around half the chromosomal coverage (18.48) (Table 2). The size of the final genome assembly of *M. sibirica* was 138.45 Mb, consisting of 21 scaffolds and 33 contigs, with the scaffold and contig N50 sizes of 13.84 Mb and 11.35 Mb, respectively. Among them, 22 contigs (99.85%, 138.25 Mb) were anchored onto ten pseudo-chromosomes with lengths ranging from 11.35 to 13.84 Mb (Table 3; Figs. 2, 3).

Genome annotation. RepeatModeler v2.0.4[32] was performed to construct a *de novo* repeat library of the genome with the parameter "-LTRStruct". Subsequently, Dfam 3.5[33] and RepBase-20181026 database[34] were employed to create a custom library. The repeated sequences were identified using RepeatMasker v4.1.4[35]. The results showed that the genome of *M. sibirica* was 35.46% repetitive reads (49.10 Mb). The transposable elements mainly include DNA elements (4.21%), LTR elements (1.60%), and LINE elements (1.07%) (Table 4).

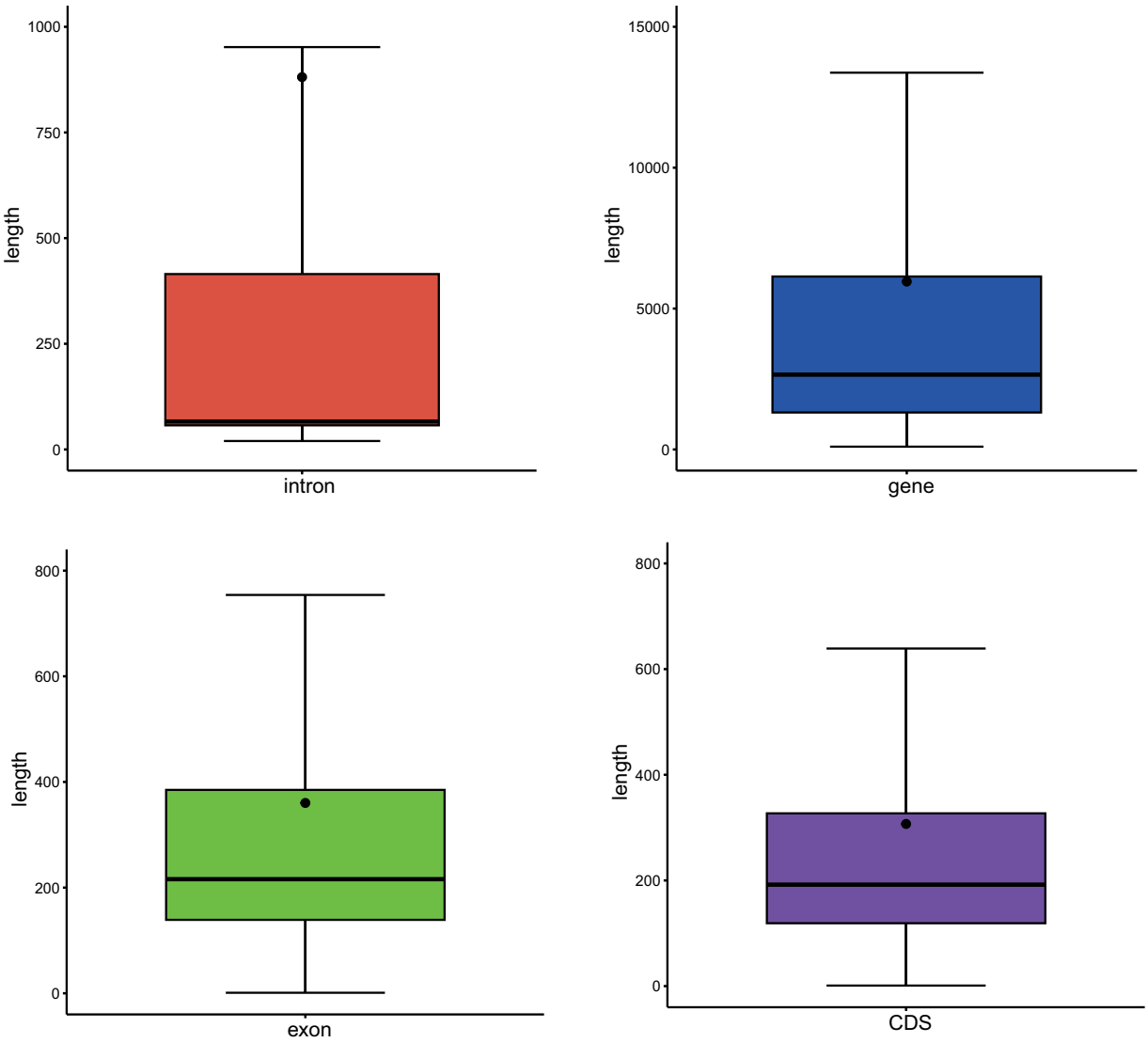

Fig. 4 Length distribution map. The average length of exons, introns, and CDS of each gene.

Non-coding RNAs (ncRNAs) in *M. sibirica* were detected using Infernal v1.1.4[36] against the Rfam v14.10 database[37], while transfer RNAs (tRNAs) were identified using tRNAscan-SE v2.0.9[38]. The analysis revealed a total of 673 ncRNAs in the *M. sibirica* genome, including 1 long non-coding RNAs, 2 ribozymes, 34 small nuclear RNAs, 57 microRNAs, 256 tRNAs, and 187 ribosomal RNAs (Table 4).

We utilized MAKER v3.01.03[39] to integrate three strategies: ab initio prediction, homology search, and transcriptomic data. For the *ab initio* prediction, BRAKER v2.1.6[40] with the parameter "-etpmode" and "-softmasking", and GeMoMa v1.7.1[41] were employed for MAKER. We used HISAT2 v2.2.0[42] with the parameters "-q" and "-x" to generate transcriptome alignments. BRAKER, combining the results of the two (Augustus v3.3.4[43] and GeneMark-ES/ET/EP v4.68_lic[44], with their default parameters), was employed to automatically train prediction models, which was based on RNA-seq alignments and reference proteins mined from the OrthoDB10 v1 database[45]. According to the parameters "GeMoMa.c = 0.4 GeMoMa.p = 10" and protein sequences from five species, including *Anoplophora glabripennis*[46] (GCF_000390285.2), *Dendroctonus ponderosae*[47] (GCF_000355655.1), *Tribolium castaneum*[48] (GCF_000002335.3), *Drosophila melanogaster*[49] (GCF_000001215.4) and *Bombyx mori*[50] (GCF_014905235.1), GeMoMa predicted genes by employing protein homology and intron conservation position. Additionally, the protein sequences required from the same set of five species used in GeMoMa were included as evidence for protein homology in MAKER. Finally, 11,687 protein-coding genes were annotated with an average length of 5,959.6 bp in the *M. sibirica* genome (Table 4). The average number of exons, introns, and CDS of each gene were 5.9, 4.8, and 5.5, whose corresponding mean length was 360.2, 881.0, and 306.9 bp, respectively (Fig. 4). Based on the parameters "–more -sensitive -e 1e-5", we used Diamond v2.0.11.1[51] to annotate the gene function against the UniProtKB database. Moreover, eggNOG-mapper v2.1.5[52] and InterProScan 5.53–87.0[53] were applied to assign Gene Ontology (GO), enzyme codes (EC), Kyoto Encyclopedia of Genes and Genomes (KEGG), clusters of orthologous groups (COG), KEGG

pathway annotations and protein domains. The InterProScan analyses consisted of five databases: Pfam[54], SMART[55], Superfamily[56], Gene3D[57], and CDD[58] (Conserved Domain Database). In summary, Genes with 9,253 GO terms, 4,403 KEGG pathways, 2,462 Enzyme Codes, 4,403 Reactome pathways, and 10,269 COG categories were identified by combining the eggNOG and InterProScan annotation results.

Data Records

The raw sequencing data and genome assembly of *M. sibirica* have been uploaded to NCBI. The Illumina, Hi-C, PacBio, and RNA-seq data can be accessed under the following accession numbers SRR29001182[59], SRR29001180[60], SRR29001183[61], SRR29001181[62], respectively. In addition, the assembled genome has been deposited in the GeneBank (GCA_039776855.1) in NCBI[63]. We have deposited the genome annotation results in Figshare[64].

Technical Validation

Two independent approaches were adopted to examine the completeness and the genome assembly quality. BUSCO v5.0.4[65] with the "insecta_odb10" database (n = 1,367) was first used to assess the completeness of the genome. To investigate the quality of the *de novo* assembly, Merqury v1.3[66] was performed to identify possible assembly sequence errors based on efficient k-mer set operations and QV score calculation. Consequently, the k-mer completeness value of *M. sibirica* is 94.2%, and the QV score is 59.336. BUSCO analysis showed a commendable 100.0% of complete BUSCOs, including 99.4% of single-copy genes and 0.6% of duplicated BUSCOs (Table 2). To evaluate mapping accuracy, Minimap2 and SAMtools v1.9[65] were conducted to align the clean reads required from both Illumina and PacBio sequencing with the final assembly. Finally, the mapping rates were 95.23% for Illumina reads, 95.19% for PacBio reads, and 97.79% for RNA-seq reads, respectively. In conclusion, these values show the high level of genomic assembly in our study.

Code availability

No specific script was used in this work. All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatic software.

Received: 17 May 2024; Accepted: 24 January 2025;
Published online: 14 February 2025

References

1. Tan, Z. *et al*. Functional genomics of *Brassica napus*: progresses, challenges, and perspectives. *J. Integr. Plant Biol.* **66**, 484–509 (2024).
2. Li, Z., Costamagna, A. C., Beran, F. & You, M. Biology, ecology, and management of flea beetles in *Brassica* crops. *Annu. Rev. Entomol.* **69**, 199–217 (2024).
3. Deng, P., Peng, Y., Sheng, Z., Li, W. & Liu, Y. RNAi silencing *CHS1* gene shortens the mortality time of *Plutella xylostella* feeding Bt-transgenic *Brassica napus*. *Pest Manag. Sci*., (2024).
4. Tian, Yu. bo., R. Y., Liu, Hai. bo. & Mao, Yu. Distribution characteristics of the main medicinal Insects of Meloidae in the Hasi Mountain, Gansu province. *Sichuan Journal of Zoology*. **40**, 383–393 (2021).
5. Zhang, G., Olsson, R. L. & Hopkins, B. K. Strategies and techniques to mitigate the negative impacts of pesticide exposure to honey bees. *Environ. Pollut.* **318**, 120915 (2023).
6. Ahmad, M. F. *et al*. Pesticides impacts on human health and the environment with their mechanisms of action and possible countermeasures. *Heliyon*. **10**, e29128 (2024).
7. Zhao, Y. Y., Liu, F., Yang, G. & You, M. S. PsOr1, a potential target for RNA interference-based pest management. *Insect Mol. Biol.* **20**, 97–104 (2011).
8. Guo, M. *et al*. RNAi assays in the striped flea beetle (Phyllotreta striolata) suggest Psγ-COPI and PsArf1COPI as potential molecular targets for pest control. *Pestic. Biochem. Physiol.* **193**, 105428 (2023).
9. Xie, C. *et al*. Genome-wide analysis of *V-ATPase* genes in *Plutella xylostella* (L.) and the potential role of PxVHA-G1 in resistance to *Bacillus thuringiensis* Cry1Ac toxin. *Int. J. Biol. Macromol.* **194**, 74–83 (2022).
10. ČernÝ, L. & Vrabec, V. Mylabris (Mylabris) snizeki sp. nov. from Jordan, with a key to the Jordanian species of the nominotypical subgenus (Coleoptera: Meloidae). *Zootaxa*. **4555**, 146–150 (2019).
11. Pan, Z. & Ren, G. D. New synonyms, combinations and status in the Chinese species of the family Meloidae Gyllenhal, 1810 (Coleoptera: Tenebrionoidea) with additional faunistic records. *Zootaxa*. **4820**, zootaxa.4820.4822.4823 (2020).
12. Bologna, M. A., D'Inzillo, B., Cervelli, M., Oliverio, M. & Mariottini, P. Molecular phylogenetic studies of the Mylabrini blister beetles (Coleoptera, Meloidae). *Mol. Phylogenet. Evol.* **37**, 306–311 (2005).
13. Shen, C. H. *et al*. Evaluation of reference genes for quantitative expression analysis in *Mylabris sibirica* (Coleoptera, Meloidae). *Front. Physiol.* **15**, 1345836 (2024).
14. Li, S., Wu, X., Fan, G., Du, K. & Deng, L. Exploring cantharidin and its analogues as anticancer agents: a review. *Curr. Med. Chem.* **30**, 2006–2019 (2023).
15. Jin, D., Huang, N. N. & Wei, J. X. Hepatotoxic mechanism of cantharidin: insights and strategies for therapeutic intervention. *Front. Pharmacol.* **14**, 1201404 (2023).
16. Zhu, M. *et al*. Cantharidin treatment inhibits hepatocellular carcinoma development by regulating the JAK2/STAT3 and PI3K/Akt pathways in an EphB4-dependent manner. *Pharmacol. Res.* **158**, 104868 (2020).
17. Frenzel, M. & Dettner, K. Quantification of cantharidin in canthariphilous ceratopogonidae (Diptera), anthomyiidae (Diptera) and cantharidin-producing oedemeridae (Coleoptera). *J. Chem. Ecol.* **20**, 1795–1812 (1994).
18. Carrel, J. E. & Eisner, T. Cantharidin: potent feeding deterrent to insects. *Science.* **183**, 755–757 (1974).
19. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods.* **58**, 268–276 (2012).
20. Bushnell, B. B. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 October 2022). (2014).
21. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. **33**, 2202–2204 (2017).
22. 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).
23. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898 (2020).
24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).

25. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*. **3**, 95–98 (2016).
26. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**, 92–95 (2017).
27. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol*. **35**, 1026–1028 (2017).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410 (1990).
29. De Almeida, M. C., Zacaro, A. A. & Cella, D. M. Cytogenetic analysis of *Epicauta atomaria* (Meloidae) and *Palembus dermestoides* (Tenebrionidae) with Xyp sex determination system using standard staining, C-bands, NOR and synaptonemal complex microspreading techniques. *Hereditas*. **133**, 147–157 (2000).
30. Zacaro, A. A., De Almeida, M. C. & Cella, D. M. Recombination nodules in coleopteran species: *Palembus dermestoides* (Tenebrionidae) and *Epicauta atomaria* (Meloidae). *Cytogenet. Genome Res*. **103**, 185–191 (2003).
31. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
32. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*. **117**, 9451–9457 (2020).
33. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. **44**, D81–89 (2016).
34. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. **6**, 11 (2015).
35. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: http://www.repeatmasker.org (accessed on 1 October 2022). (2013–2015).
36. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
37. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. **33**, D121–124 (2005).
38. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol. Biol*. **1962**, 1–14 (2019).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. **12**, 491 (2011).
40. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics*. **3**, lqaa108 (2021).
41. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*. **19**, 189 (2018).
42. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).
43. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*. **32**, W309–312 (2004).
44. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR genomics and bioinformatics*. **2**, lqaa026 (2020).
45. 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).
46. McKenna, D. D. *et al*. Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol*. **17**, 227 (2016).
47. Keeling, C. I. *et al*. Chromosome-level genome assembly reveals genomic architecture of northern range expansion in the mountain pine beetle, *Dendroctonus ponderosae* Hopkins (Coleoptera: Curculionidae). *Molecular ecology resources*. **22**, 1149–1167 (2022).
48. Jin, J., Zhao, Y., Zhang, G., Pan, Z. & Zhang, F. The first chromosome-level genome assembly of *Entomobrya proxima* Folsom, 1924 (Collembola: Entomobryidae). *Scientific data*. **10**, 541 (2023).
49. Hoskins, R. A. *et al*. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res*. **25**, 445–458 (2015).
50. Kim, S. W. *et al*. Whole-genome sequences of 37 breeding line Bombyx mori strains and their phenotypes established since 1960s. *Scientific data*. **9**, 189 (2022).
51. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*. **12**, 59–60 (2015).
52. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol*. **34**, 2115–2122 (2017).
53. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res*. **45**, D190–d199 (2017).
54. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res*. **47**, D427–d432 (2019).
55. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. **46**, D493–d496 (2018).
56. Wilson, D. *et al*. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*. **37**, D380–386 (2009).
57. Lewis, T. E. *et al*. Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Res*. **46**, D435–d439 (2018).
58. Marchler-Bauer, A. *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*. **45**, D200–d203 (2017).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29001182 (2024).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29001180 (2024).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29001183 (2024).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29001181 (2024).
63. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_039776855.1 (2024).
64. Shen, C. Genome annotation (repeats and protein-coding genes). *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.25687182 (2024).
65. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol*. **35**, 543–548 (2018).
66. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience*. **10** (2021).

Acknowledgements

This research was supported by the National Major Special Project of Breeding for Genetically Modified Organisms in China (No. 2016ZX08012-005) and the Innovation Project of Chinese Academy of Agricultural Sciences (No. 2060302-049-091).

Author contributions

G.W. and X.Y. supervised the project. C.S., G.Y. and M.T. contributed to the research design. C.S., G.Y., M.T., X.L., L.Z. and W.L. collected the samples for genome sequencing and conducted bioinformatics analysis. C.S., G.Y., M.T., X.L., L.Z., W.L., L.J., P.D., H.Z., Q.Z., G.W. and X.Y. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to G.W. or X.Y.

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.