



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the Japanese sawyer beetle *Monochamus alternatus*

Yong-Fu Gao^{1,2}, Fang-Yuan Yang¹ , Wei Song², Li-Jun Cao², Jin-Cui Chen², Xiu-Jing Shen², Liang-Jian Qu³, Shi-Xiang Zong¹ & Shu-Jun Wei¹

The Japanese sawyer beetle *Monochamus alternatus* (Coleoptera: Cerambycidae) is a pest in pine forests and acts as a vector for the pine wood nematode *Bursaphelenchus xylophilus*, which causes the pine wilt disease. We assembled a high-quality genome of *M. alternatus* at the chromosomal level using Illumina, Nanopore, and Hi-C sequencing technologies. The assembled genome is 767.12 Mb, with a scaffold N50 of 82.0 Mb. All contigs were assembled into ten pseudo-chromosomes. The genome contains 63.95% repeat sequences. We identify 16,284 protein-coding genes in the genome, of which 11,244 were functionally annotated. The high-quality genome of *M. alternatus* provides an invaluable resource for the biological, ecological, and genetic study of this beetle and opens new avenues for understanding the transmission of pine wood nematode by insect vectors.

Background & Summary

The pine wilt disease is currently considered one of the most serious threats to pine forests worldwide^{1–3}. This disease is caused by the pinewood nematode *Bursaphelenchus xylophilus* (Steiner and Buhner) (Nematoda: Aphelenchoididae), an invasive species originally from North America². The natural spread of pinewood nematode usually requires insect vectors⁴. The longhorn beetles from the *Monochamus* (Coleoptera: Cerambycidae) are the primary vectors of the pinewood nematode^{5–7}. The Japanese sawyer beetle *Monochamus alternatus* (Hope) (Coleoptera: Cerambycidae: Lamiinae) is an effective vector of the pinewood nematode⁸. The *M. alternatus* can cause damage directly to various species of pine trees from the genera *Pinus*, *Cedrus*, *Abies*, *Picea*, and *Larix*⁴. This beetle is widely distributed in Japan, Korea, Laos, Vietnam, and the surrounding countries^{9,10}. Single *M. alternatus* can harbor, on average, 15,000 and up to 280,000 pinewood nematodes in its tracheal system^{11,12}. *Monochamus saltuarius* is another species that was reported as the vector beetle of pinewood nematode in Japan, Europe, and China. It was first reported to transmit the pinewood nematode to native *Pinus* species in Liaoning Province, China¹³. It is crucial to understand the ecology and genetics of *M. alternatus* and how it transmits pinewood nematodes^{14,15}. The genome of *M. saltuarius* has been sequenced and assembled¹⁶. However, the genome of *M. alternatus* has yet to be determined. Bridging this knowledge gap will greatly aid our control efforts against *M. alternatus* and pine wilt disease¹⁷.

In this study, we assembled the chromosome-level genome of *M. alternatus* using a combination of Nanopore, Illumina short-read sequencing, and chromosome conformation capture (Hi-C) technologies to provide genomic resources for future investigations on the ecology, genetics, and evolution of the *M. alternatus* and the interaction between the pinewood nematode and its insect vector.

Methods

Sample preparation. Samples of *M. alternatus* were from a laboratory strain reared at the Key Laboratory of Forest Protection of National Forestry and Grassland Administration, Ecology and Nature Conservation Institute, Chinese Academy of Forestry, Beijing, China. This strain was reared for about 30 generations in the laboratory. A single female adult was used to construct libraries of Illumina short read, Oxford Nanopore Technology (ONT) long read sequencing, and Hi-C. The samples were starved for 24 hours, and the guts of the adults were removed

¹Beijing Key Laboratory for Forest Pest Control, Beijing Forestry University, Beijing, 100083, China. ²Institute of Plant Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 1000097, China. ³Key Laboratory of Forest Protection of National Forestry and Grassland Administration, Ecology and Nature Conservation Institute, Chinese Academy of Forestry, Beijing, 100091, China. ✉e-mail: zongshixiang@bjfu.edu.cn; shujun268@163.com

Sequencing strategy	Platform	Usage	Insertion size	Clean data (Gb)	Coverage (X)
Short-read	Illumina	Genome survey	300 bp	42.50	55
Long-read	Nanopore	Genome assembly	10–20 kb	142.7	186
Hi-C	Illumina	Hi-C assembly	300 bp	81.7	106
RNA-seq	Illumina	Anno-evidence	300 bp	18.9	27

Table 1. Library sequencing data and methods used in this study to assemble the *Monochamus alternatus* genome.

to minimize contamination from gut microbes. In addition, we collected three larvae, pupae, and adults of *M. alternatus* for transcriptome sequencing. All samples were frozen in liquid nitrogen and stored at -80°C until further usage.

Genomic DNA and RNA sequencing. For short-read sequencing, genomic DNA was extracted using the QIAGEN[®] Genomic DNA extraction kit (Qiagen, Hilden, Germany) according to the standard operating procedure provided by the manufacturer. The pair-end library with an insert size of about 300 bp was prepared using VAHTSTM Universal DNA Library Prep Kit for Illumina[®] V3 (Vazyme, ND607, Nanning, China) and sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA). We obtained 42.5 Gb Illumina short reads (Table 1).

For long-read sequencing, high molecular weight genomic DNA was isolated using the QIAGEN[®] Genomic DNA extraction kit (Qiagen, Hilden, Germany) according to the standard operating procedure provided by the manufacturer. A total of 3–4 μg DNA was used as input material for the ONT library preparation. Long DNA fragments were selected using the PippinHT system (Sage Science, USA). The A-ligation reaction was conducted with the NEBNext Ultra II End Repair/dA-tailing Kit (Ipswich, MA, USA). The adapter in the SQK-LSK109 (Oxford Nanopore Technologies, UK) was used for further ligation reaction. About 700 ng DNA library was constructed and performed on a Nanopore PromethION sequencer instrument (Oxford Nanopore Technologies, UK) at the GrandOmics Biosciences Co., Ltd. (Wuhan, China), and 142.7 Gb long reads were generated (Table 1).

For Hi-C sequencing, the library was prepared according to the standard protocol described by Belton with minor modifications¹⁸. An adult of *M. alternatus* was cut into pieces and mixed with 2% formaldehyde solution for cross-linking. Glycine (2.5 M) was added to stop this reaction, and the sample was homogenized to separate the nuclei. The purified nuclei were dissolved in SDS and incubated at 65°C for 10 min. After quenching the SDS with Triton X-100, the sample was digested with Dpn II and marked by incubating with biotin-14-dCTP. Biotin from nonligated DNA ends was removed by T4 DNA polymerase. Then, the Hi-C library was prepared by Truseq Nano DNA HT Kit (Illumina, USA) and sequenced on the Illumina HiSeq platform with paired-end 150-bp reads (Illumina, San Diego, CA, USA) at Annoroad Gene Technology Co., Ltd. (Beijing, China). A total of 81.7 Gb ($106 \times$ coverage) of clean data was generated (Table 1).

For transcriptome sequencing, total RNA was extracted from a single *M. alternatus* (larva, pupa, and adult, respectively) using the RNAPrep Pure Tissue Kit (Tiangen, China). Library was constructed using a TruSeq RNA sample preparation kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) with the paired-end mode at GrandOmics Biosciences Co., Ltd. (Wuhan, China). A total of 18.9 Gb transcriptome data was obtained (Table 1).

Estimation of genomic characteristics. The Illumina raw reads were checked and filtered using Trimmomatic version 0.39-2¹⁹ to discard reads with adaptors, unknown nucleotides (Ns), or $>20\%$ low-quality bases. Genome size, heterozygosity, and duplication were estimated by using Jellyfish version 2.2.10²⁰ and GenomeScope version 2.0²¹ based on the 17-mer depth distribution. The estimated genome size was 667 Mb, with a heterozygosity rate of 1.31% and a duplication rate of 1.55% (Fig. 1A).

Genome assembly. A draft genome at contig level was assembled using NextDenovo version 1.2.5 (<https://github.com/Nextomics/NextDenovo>) with default parameters (genome-size = 667, read-cutoff = 3k) based on Nanopore long reads. Purge_dups was used to remove alternative haplotype and redundant fragments in the contig assembly. We performed Hi-C analysis to further anchor the assembly into chromosome-scale linkage groups. The Hi-C clean reads were cleaned using Fastp²² and mapped to the contigs using BWA. YaHS version 1.2a.1²³ and JuicerTools version 1.19.02²⁴ were used to assemble and manual correction. As a result, 98.21% of the contigs were anchored to 10 pseudo-chromosomes, which were presented in the heatmap of the chromatin contact matrix (Fig. 1B). At last, two rounds of polishing with ONT reads and Illumina reads were performed using NextPolish version 1.4.0²⁵. The output chromosome-level genome has a size of 767.12 Mb, N50 of 82.0 Mb, maximum length of 149.24 Mb, and GC content of 32.35% (Table 2).

Genome annotation. The protein-coding genes in the *M. alternatus* genome were predicted under three lines of evidence, including RNA-based, *ab initio*, and homology-based methods. For the RNA-based method, short transcriptome reads were mapped to the genome using Hisat2²⁶. Then, the aligned BAM files were used to assemble the transcripts using Stringtie version 2.1.4²⁷. The genes were predicted using PASA version 2.0.2 with default settings²⁸. The *ab initio* prediction was performed using Augustus version 3.4.0²⁹ and SNAP version 2006-07-28³⁰. The gene models in Augustus and SNAP were trained based on transcripts longer than 300 bp generated by PASA. In the homology-based prediction, we gathered evidence of homologous genes from

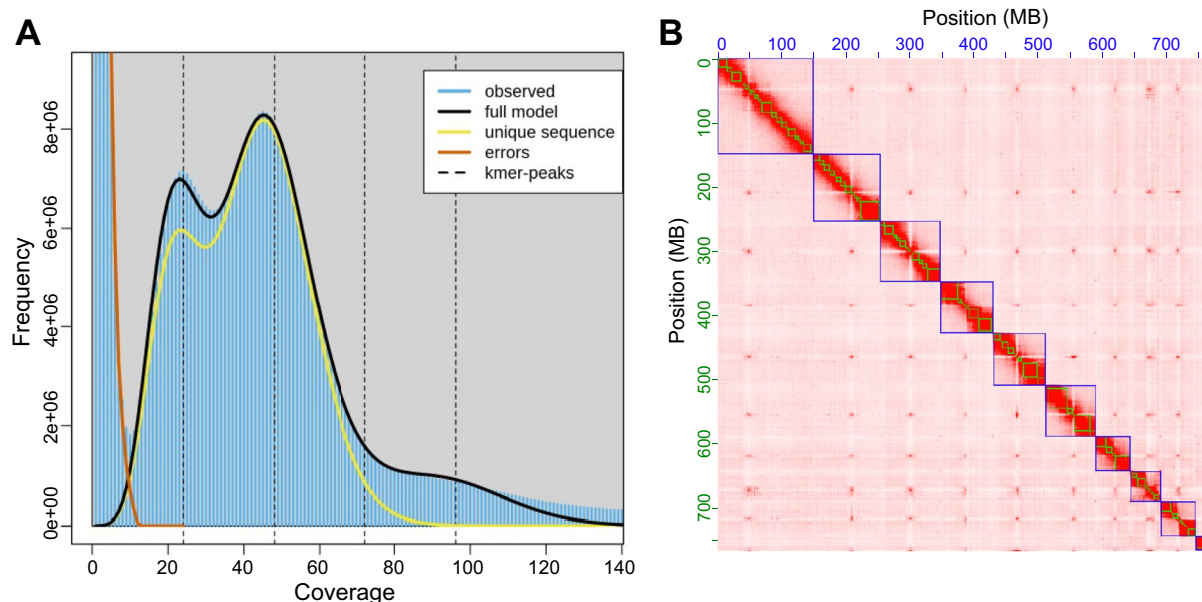


Fig. 1 Feature estimation and assembly of *Monochamus alternatus* genome. (A) Estimation of *M. alternatus* genomic features. The 17-mer distributions showed double peaks: the first peak with a coverage of 100 indicates genome duplication, and the highest peak with a coverage of 200 represents a genome-size peak. *M. alternatus* genome size was calculated to be 667 Mb with heterozygosity rate of 1.31% and duplication rate of 1.55%. (B) **Genome-wide contact matrix of *Monochamus alternatus* generated using Hi-C data.** Each black square represents a pseudo-chromosome. The color bar indicates the interaction intensity of Hi-C contacts.

Survey statistics	Sequencing platform	Illumina
	Estimated genome size (bp)	667,998,064
	Heterozygosity rate (%)	1.31
	Duplication rate (%)	1.55
Assembly statistics	Sequencing platform	Illumina, Nanopore, Hi-C
	Total length (bp)	767,125,294
	Longest scaffold length (bp)	149,245,057
	Scaffold N50 (bp)	82,003,078
	GC content (%)	32.35
Annotation statistics	Anchored to chromosome (%)	98.21%
	Bases masked	63.95%
	Number of annotated protein-coding gene	16,284
	Number of functionally annotated gene	11,244

Table 2. Statistics for the chromosomal-level genome of the *Monochamus alternatus*.

Coleoptera species, including *Anoplophora glabripennis*³¹, *Tribolium castaneum*³², *Dendroctonus ponderosae*³³ and *Diabrotica virgifera*³⁴. Redundant genes in the pooled gene set were removed using CD-HIT³⁵. Maker version 3.01.04³⁶ pipeline was used to perform the homology-based prediction. At last, the evidence from these methods was combined using EvidenceModeler (EVM) version 1.1.1³⁷ to obtain a non-redundant consensus official gene set (OGS).

The predicted genes were functionally annotated using EggNog-Mapper version 2.1.9³⁸. Five methods were used to search against several public databases, including Gene Ontology (GO), Clusters of Orthologous Groups of Proteins (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), CAZY, and Pfam. In summary, we identified 16,284 protein-coding genes (Table 2), of which 11,244 were functionally annotated (Table 3).

Repeats prediction. Homology-based and *de novo* prediction methods were used to detect transposable elements (TEs). Briefly, repeats sequences were detected using RepeatMasker version 4.1.2 (-no_is -norna -xsmall -q)³⁹, against the Repbase, Dfam database, and species-specific repeat library identified by RepeatModeler version 2.0.3. Finally, 63.95% of the genome was identified to be repeat DNA. Overall, 576,182 transposable elements (TEs), including 178,967 retroelements (189 short interspersed nuclear elements (SINEs), 144,289 long interspersed nuclear elements (LINEs), and 34,489 long terminal repeats (LTR)) and 397,215 DNA transposons were

Database name	Annotated number	Percent (%)
GO	8311	73.91
KEGG	4751	42.25
COG	10635	94.58
CAZY	295	2.62
PFAM	10519	93.55
Total	11244	98.44

Table 3. Number of functionally annotated protein-coding gene in different databases.

Item	Number of element	Length occupied (bp)	Percentage of sequence (%)
Retroelements	178,967	76,112,632	9.72
SINEs	189	14,122	0
LINEs	144,289	56,571,154	7.23
LTR elements	34,489	19,527,356	2.49
DNA transposons	397,215	133,929,988	17.11
Rolling-circles	7,406	2,474,246	0.32
Unclassified	1,235,289	287,908,114	36.78
Total interspersed repeats	NA	497,950,734	63.61
Small RNA	0	0	0
Satellites	523	126,782	0.02
Simple repeats	678	105,396	0.01
Low complexity	0	0	0

Table 4. Repeats elements statistics in genome of *Monochamus alternatus* using RepeatMasker. SINEs: short interspersed nuclear elements; LINEs: long interspersed nuclear elements; LTR: long terminal repeat.

Class	Type	Count
rRNA count	8s_rRNA	98
	28s_rRNA	5
	18s_rRNA	4
Cove stats	Candidate tRNAs read	9436
	Cove-confirmed tRNAs	498
	Bases scanned by Cove	987,197
	Seq scanned by Cove	0.1%
tRNA count	tRNAs decoding Standard 20 AA	452
	Selenocysteine tRNAs (TCA)	0
	Possible suppressor tRNAs (CTA, TTA)	2
	tRNAs with undetermined/unknown isotypes	3
	Predicted pseudogenes	41
Total tRNAs		498
tRNAs with intron		43

Table 5. Statistics of non-coding RNAs in genomes of *Monochamus alternatus*.

identified. Five hundred twenty-three satellites and 678 simple repeats were identified as tandem repeats (TRs), accounting for 0.03% of the *M. alternatus* genome (Table 4).

Non-coding RNA annotation. For non-coding RNA annotation, the transfer RNA (tRNA) was annotated by tRNAscan-SE version 1.3.1 based on the structural characteristics of tRNA⁴⁰, whereas the ribosome RNA (rRNA) was predicted by RNAmmer version 1.2^{41,42}. We obtained 498 tRNA and 107 rRNA genes, including 98 8s_rRNA, five 28s_rRNA, and four 18s_rRNA genes in the *M. alternatus* genome (Table 5).

Data Records

The genome project was deposited at NCBI under BioProject number PRJNA819115. Illumina sequencing data for genome survey were deposited in the Sequence Read Archive at NCBI under accession number SRR26115523⁴³. Hi-C sequencing data were deposited in the Sequence Read Archive at NCBI under accession number SRR26146338⁴⁴. Nanopore sequencing raw data were deposited in the Sequence Read Archive at NCBI under accession number SRR26157698⁴⁵. RNA-seq data were deposited in the Sequence Read Archive at NCBI under accession numbers SRR26116071- SRR26116073⁴⁶⁻⁴⁸. The final chromosome assembly, genome structure

Description	Contig-level		Chromosome-level		Official gene set (OGS)	
	Number	Percent (%)	Number	Percent (%)	Number	Percent (%)
Complete gene (C)	1282	93.8	1363	99.7	1322	96.7
Complete and single-copy gene (S)	1274	93.2	1353	99	1315	96.2
Complete and duplicated gene (D)	8	0.6	10	0.7	7	0.5
Fragmented gene (F)	45	3.3	0	0	21	1.5
Missing gene (M)	40	2.9	4	0.3	24	1.8
Total gene searched	1367	/	1367	/	1367	/

Table 6. BUSCO evaluation of genome assemblies.

annotation, amino acid sequences and functional annotation results of protein-coding genes were deposited to Figshare repository under a DOI number of <https://doi.org/10.6084/m9.figshare.c.6849162.v1>⁴⁹. The final chromosome assembly was deposited in GenBank under accession number JAYMDT000000000⁵⁰.

Technical Validation

The Hi-C heatmap exhibits the accuracy of genome assembly, with relatively independent Hi-C signals observed between the ten pseudo-chromosomes (Fig. 1B). We assessed the accuracy of the final genome assembly by mapping Illumina short reads to the *M. alternatus* genome with BWA-MEM2 version 0.7.1721⁵¹. The mapping rate for Illumina reads was 98.71%. The findings indicate that the quality of our assembled genome is high.

To assess the completeness of genome assembly and OGS, we run Benchmarking Universal Single-Copy Orthologues (BUSCO version 5.2.2) using the insecta_odb10 database, which contains 1367 conserved genes⁵². For Contig-level, in the first round, the BUSCO analysis showed that 93.8% (single-copied gene: 93.2%, duplicated gene: 0.6%) of 1367 single-copy genes were identified as complete, 3.3% of genes were fragmented, and 2.9% of genes were missing in the assembled genome. For the chromosome-level assembly, BUSCO analysis showed that 99.7% (single-copied gene: 99.0%, duplicated gene: 0.7%) of 1367 genes were identified as complete, 0% of genes were fragmented, while 0.3% of genes were missing in the assembled genome. For OGS, BUSCO analysis showed 96.7% completeness, with only 0.5% of genes duplicated, 1.5% fragmented, and 1.8% missing (Table 6).

Code availability

There were no custom scripts or code utilized in this study.

Received: 25 September 2023; Accepted: 5 February 2024;

Published online: 13 February 2024

References

- Togashi, K. & Jikumaru, S. Evolutionary change in a pine wilt system following the invasion of Japan by the pinewood nematode, *Bursaphelenchus xylophilus*. *Ecol. Res.* **22**, 862–868 (2007).
- Mamiya, Y. Pathology of the pine wilt disease caused by *Bursaphelenchus xylophilus*. *Annu. Rev. Phytopathol.* **21**, 201–220 (1983).
- Hao, Z., Huang, J., Li, X., Sun, H. & Fang, G. A multi-point aggregation trend of the outbreak of pine wilt disease in China over the past 20 years. *For. Ecol. Manag.* **505**, 119890 (2022).
- Akbulut, S. & Stamps, W. Insect vectors of the pinewood nematode: a review of the biology and ecology of *Monochamus* species. *For. Pathol.* **42**, 89–99 (2012).
- Linit, M. J. Nematode-vector relationships in the pine wilt disease system. *J. Nematology* **20**, 227–235 (1988).
- Zhao, L. L., Mota, M., Vieira, P., Butcher, R. A. & Sun, J. H. Interspecific communication between pinewood nematode, its insect vector, and associated microbes. *Trends Parasitol.* **30**, 299–308 (2014).
- Aikawa, T. *Transmission biology of Bursaphelenchus xylophilus in relation to its insect vector*. Tokyo, Japan edn, 123–138 (2008).
- Kobayashi, F., Yamane, A. & Ikeda, T. The Japanese pine sawyer beetle as the vector of pine wilt disease. *Annu. Rev. Entomol.* **29**, 115–135 (1984).
- Fan, J., Sun, J. & Shi, J. Attraction of the Japanese pine sawyer, *Monochamus alternatus*, to volatiles from stressed host in China. *Ann. For. Sci.* **64**, 67–71 (2007).
- Kwon, T.-S. *et al.* Distribution patterns of *Monochamus alternatus* and *M. saltuarius* (Coleoptera: Cerambycidae) in Korea. *J. Korean For. Soc.* **95**, 543–550 (2006).
- Tang, X. *et al.* Hypoxia-induced tracheal elasticity in vector beetle facilitates the loading of pinewood nematode. *eLife* **12**, e84621 (2023).
- Bossen, J., Kuhle, J.-P. & Roeder, T. The tracheal immune system of insects—a blueprint for understanding epithelial immunity. *Insect Biochem. Mol. Biol.* **157**, 103960 (2023).
- The First Record of *Monochamus saltuarius* (Coleoptera: Cerambycidae) as Vector of *Bursaphelenchus xylophilus* and Its New Potential Hosts in China *Insects*. **11**(9), 636, <https://doi.org/10.3390/insects11090636> (2020).
- Kim, B.-N. *et al.* A short review of the pinewood nematode, *Bursaphelenchus xylophilus*. *Toxicol. Environ. Health Sci.* **12**, 297–304 (2020).
- Maehara, N., Kanzaki, N., Aikawa, T. & Nakamura, K. Potential vector switching in the evolution of *Bursaphelenchus xylophilus* group nematodes (Nematoda: Aphelenchoididae). *Ecol. Evol.* **10**, 14320–14329 (2020).
- Fu, N. N. *et al.* Chromosome-level genome assembly of *Monochamus saltuarius* reveals its adaptation and interaction mechanism with pine wood nematode. *Int J Biol Macromol* **222**, 325–336 (2022).
- Zhao, L. *et al.* Chemical signals synchronize the life cycles of a plant-parasitic nematode and its vector beetle. *Curr. Biol.* **23**, 2038–2043 (2013).
- Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
21. 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).
22. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
23. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
24. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, <https://doi.org/10.1016/j.cels.2016.07.002> (2016).
25. Hu, J., Fan, J. P., Sun, Z. Y. & Liu, S. L. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
26. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
27. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
28. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22 (2008).
29. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
30. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, <https://doi.org/10.1186/1471-2105-1185-1159> (2004).
31. 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).
32. Tribolium Genome Sequencing, C. *et al.* The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**, 949–955 (2008).
33. Keeling, C. I. *et al.* Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biol.* **14**, R27 (2013).
34. Lata, D., Coates, B. S., Walden, K. K. O., Robertson, H. M. & Miller, N. J. Genome size evolution in the beetle genus *Diabrotica*. *G3-Genes Genom. Genet.* **12**, jkac052 (2022).
35. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
36. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
37. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
38. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
39. Tarailo-Graovac, M. & Chen, N. S. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **5**, 4.10.11–14.10.14 (2009).
40. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686–689 (2005).
41. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
42. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
43. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26115523> (2023).
44. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26146338> (2023).
45. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26157698> (2023).
46. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26116071> (2023).
47. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26116072> (2023).
48. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRR26116073> (2023).
49. Wei, S. J. & Gao, Y. F. Chromosome-level genome assembly of the Japanese sawyer beetle *Monochamus alternatus* (Coleoptera: Cerambycidae), an insect vector of the pine wood nematode. *Figshare. Collection*. <https://doi.org/10.6084/m9.figshare.c.6849162.v1> (2023).
50. Gao, Y. F., Wei, S. J. & Zong, S. X. *Monochamus alternatus* Hope Genome sequencing and assembly. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_035320865.1 (2024).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Manni, M., Berkeley, M. R., Seppely, 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).

Acknowledgements

This research was supported by the National Key R&D Program of China (2021YFD1400900), the Program of Beijing Academy of Agriculture and Forestry Sciences (JKZX202208), and Fundamental Research Funds of Chinese Academy of Forestry (CAFYBB2021ZG001).

Author contributions

S.J.W. and S.X.Z. designed the study. L.J.Q., J.C.C., and X.J.S. contribute to the samples; Y.F.G., L.J.C. W.S. and F.Y.Y. contribute to the genome assembly and annotation. Y.F.G. and S.J.W. wrote the draft manuscript. S.J.W. contributed substantially to the revisions. The final manuscript has been read and approved by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.-X.Z. or S.-J.W.

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) 2024