Scientific Data

https://doi.org/10.1038/s41597-026-07026-4

Article in Press

Chromosome-level genome assembly and annotation of the termite *Reticulitermes chinensis* Snyder

Received: 30 June 2025

Accepted: 4 March 2026

Cite this article as: Yue, Z., Xin, P., Wang, J. *et al.* Chromosome-level genome assembly and annotation of the termite *Reticulitermes chinensis* Snyder. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07026-4

Zhiyong Yue, Peidong Xin, Jinpei Wang, Qi Jiang, Baozhen Zhou, Chenguang Feng, Jihu Sun & Jia Wu

Title: Chromosome-level genome assembly and annotation of the termite *Reticulitermes chinensis* Snyder

Zhiyong Yue[1], Peidong Xin[2,*], Jinpei Wang[1], Qi Jiang[1], Baozhen Zhou[3], Chenguang Feng[2], Jihu Sun[1], Jia Wu[1,*]

[1] Key Laboratory of Natural Anti-aging Product Mining and Biosynthesis of Shaanxi Higher Education Institutes, Applied Research Institute of Life Sciences, Xi'an International University, Xi'an, 710077, China

[2] Shaanxi Key Laboratory of Qinling Ecological Intelligent Monitoring and Protection, School of Ecology and Environment, Northwestern Polytechnical University, Xi'an, 710072, China

[3] College of Medicine, Xi'an International University, Xi'an, 710077, China

* Corresponding authors: Peidong Xin, yxpeidong@163.com; Jia Wu, wujia@xaiu.edu.cn

Abstract

Termites belong to the infraorder Isoptera within the order Blattodea. *Reticulitermes chinensis* Snyder, a species in the Rhinotermitidae family, remains a major pest causing severe damage to wooden structures in buildings. Here, the chromosome-level genome of *R. chinensis* was assembled using PacBio and Hi-C technologies. The final genome comprises 21 pseudochromosomes, with a total size of 1.02 Gb. The scaffold N50 is 47.29 Mb, the GC content is 29.46%, and 94% of sequences were anchored to pseudochromosomes. The BUSCO completeness score is 97.6%. The genome contains 23,733 predicted protein-coding genes, and 48.74% of it consists of repetitive sequences. This high-quality genome is conducive to understanding unique termite traits such as social behavior, efficient lignocellulose digestion, and evolutionary adaptations.

Background & Summary

Termites, highly social insects, play a critical ecological role by decomposing plant material and recycling nutrients. However, their wood-chewing behavior also causes significant economic damage to timber structures and crops, resulting in global losses of billions of dollars annually[1]. Previously classified under the order Isoptera, termites have been reclassified, based on phylogenetic evidence, as the infraorder Isoptera within the order Blattodea[2]. Genomic studies of termites are essential for understanding their unique biological traits, including social behavior, efficient lignocellulose digestion, and evolutionary adaptations, with implications for evolutionary biology, bioenergy, and pest management.

Reticulitermes chinensis Snyder, a member of the Rhinotermitidae family within the genus *Reticulitermes*, is a major pest in southern China, notorious for damaging wooden structures in buildings[3]. To date, only two species in Reticulitermes, *R. speratus* (Japan) and *R. lucifugus* (Europe), have had their raw genomes published[4,5]. The lack of genomic resources for *R. chinensis* underscores the need for a high-quality reference genome to support further research. Sequencing and analyzing the *R. chinensis* genome can elucidate genetic adaptations in the evolutionary history of insects. As highly social insects, termites offer critical insights into the origins and evolution of eusocial behavior through their genomic data[4]. Furthermore, genomic research will reveal the molecular basis of *R. chinensis*' efficient lignocellulose degradation, driven by symbiotic gut microorganisms and endogenous enzymes. This could yield novel enzymes for bioenergy applications, promoting the sustainable use of lignocellulosic biomass. Additionally, identifying genes regulating growth, development, and reproduction lays the foundation for precise, environmentally friendly pest control strategies, offering a sustainable alternative to conventional chemical methods[6].

Methods

Sample collection and genomic DNA preparation

Samples of *R. chinensis* were collected from a single colony in Xiaohe Town (33.14° N, 109.13° E), Xunyang City, Shaanxi Province, China. Using a single colony minimized genomic heterozygosity owing to the species' parthenogenetic reproduction. To ensure DNA purity, termites maintained without food for 48 h, rinsed with distilled water, flash-frozen in liquid nitrogen, and ground into fine powder to reduce the impact of chitin. High-quality genomic DNA was extracted using a sodium dodecyl sulfate (SDS)-based protocol[7]. Briefly, powdered samples were incubated in SDS solution, followed by sequential treatments with sodium chloride (NaCl) and chloroform/isoamyl alcohol (24:1). DNA was then precipitated with isopropanol, washed with 75% ethanol, air-dried, and dissolved in Tris-EDTA (TE) buffer. DNA quality was assessed for concentration, purity, and integrity using multiple approaches. Concentration was measured with a NanoDrop 2000 spectrophotometer

(Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit 4.0 fluorometer (Thermo Fisher Scientific), yielding values of 97.46 ng/μL and 99.62 ng/μL, respectively (ratio: 1.02). Purity was confirmed by absorbance ratios of 1.84 (260/280 nm) and 2.07 (260/230 nm), indicating high-quality DNA. Integrity was evaluated using a CHEF pulsed-field gel electrophoresis system (Bio-Rad, Hercules, CA, USA), revealing a primary band exceeding 60 kb with minimal smearing as compared to a 4.9–98 kb DNA ladder (Bio-Rad, Catalog No. 1703624), confirming suitability for sequencing.

Library construction and sequencing

For short-read sequencing, a library was constructed using the MGIEasy FS DNA Library Prep Set (MGI Technology, Shenzhen, China, Catalog No. 1000006987) and sequenced on the MGISEQ-2000 platform in paired-end mode (150 bp reads). After filtering adapters and low-quality reads using fastp v0.20[8] with the parameters "--length_required 20 -q 15 -n 0", 50.56 Gb of clean data was obtained, equivalent to ~51× genome coverage (Table S1, short-read sequencing statistics). For long-read sequencing, a PacBio high-fidelity (HiFi) library was prepared using the SMRTbell Prep Kit 3.0 (PacBio, Menlo Park, CA, USA). Genomic DNA was fragmented to an average size of ~15 kb, ligated with SMRTbell adapters, and sequenced on the PacBio Revio system. HiFi reads were generated using PacBio's SMRT Link pipeline, yielding 40.24 Gb of high-quality HiFi reads, corresponding to ~40× coverage (Table S1, HiFi sequencing data statistics). For Hi-C sequencing, chromatin was cross-linked with formaldehyde, digested with DpnII, biotin-labeled, and ligated. After de-crosslinking, DNA was fragmented into 300–500 bp segments, captured with streptavidin magnetic beads, and used to construct a library with the VAHTS Universal DNA Library Prep Kit for Illumina V3 (Vazyme, Nanjing, China, Catalog No. ND607-01) and VAHTS DNA Adapters (set3–set6) (Vazyme, Nanjing, China, Catalog No. N805/N806/N807/N808). Sequencing on the Illumina NovaSeq 6000 platform with 150 bp paired-end reads yielded 55.31 Gb of clean Hi-C reads (~55× coverage) after filtering using fastp v0.20 (--length_required 20 -q 15 -n 0). (Table S1, Hi-C sequencing statistics).

For transcriptome sequencing, pooled samples were used due to the small body size of individuals. Total RNA was extracted with the TRIzol kit (TIANGEN, Cat. No. DP424, China), and sequencing libraries were constructed for each replicate. Libraries were sequenced on the Illumina NovaSeq 6000 platform (150 bp paired-end) with two biological replicates. Raw reads were quality-filtered using fastp v0.20 (--length_required 20 -q 15 -n 0), yielding 13.63 Gb and 8.66 Gb of clean data for the two replicates, respectively (Table S1, transcriptome sequencing statistics). De novo assembly of the filtered reads was performed using SPAdes v3.1.1, and coding sequences were predicted with TransDecoder v5.5.0 (https://github.com/TransDecoder/).

Genome assembly

Genome size, heterozygosity, and repeat content were estimated using k-mer analysis of MGI short reads. Reads were trimmed with fastp v0.20 (--length_required 20 -q 15 -n 0)[8], and k-mer frequency distribution was calculated using Jellyfish v2.3.0[9]. Using a k-mer size of 21 (ploidy = 2), GenomeScope v2.0[10] estimated a genome size of 1.01 Gb with 0.5% heterozygosity (Fig. 1a, k-mer distribution curve), guiding the assembly process. Contig assembly was performed using Hifiasm v0.19.8[11] with default parameters, leveraging PacBio HiFi long reads. Duplicates were removed with Purge_dups v1.0.1[12]. The resulting contig assembly had a total length of 1.02 Gb and an N50 = 3.34 Mb (Table 1, HiFi assembly statistics).

Hi-C data were used for scaffolding. Reads were aligned to contigs using BWA v0.7.12[13], and scaffolds were generated with YaHS v1.1a-r3[14]. Manual curation was performed using JuiceBox v1.11.08 [15] to identify and correct potential misassemblies based on chromatin contact maps. Consistent with karyotype studies of multiple closely related species[16,17], our final chromosome-level assembly comprised 21 pseudochromosomes, with a total size of 1.02 Gb, a scaffold N50 = 47.29 Mb, a GC content of 29.46%, and 94% of assembled sequences anchored to the 21 pseudochromosomes (Fig. 1b, Hi-C contact map; Table 1, Hi-C assembly statistics). Genome completeness was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO v5.5.0) with the “insecta_odb10” database[18]. BUSCO analysis indicated a completeness of 97.6%, including 96.3% complete single-copy, 1.3% complete duplicated, 0.4% fragmented, and 2.0% missing orthologs (Table 1, BUSCO assessment of the Hi-C assembly). Whole-genome alignment and synteny analysis were performed between *R. chinensis* and *R. speratus* (GCA_021186555.1)[19] using LAST v1282[20] (lastdb: -u NEAR; lastal: -i2G -m10; last-split -m10). From this analysis, ~788.93 Mb of syntenic blocks were identified, corresponding to 77.10% of the *R. chinensis* genome and 89.59% of the *R. speratus* genome. Chromosomal collinearity was visualized using Circos v0.69-921 (Fig. 2a, synteny Circos plot).

Genome annotation

Repetitive elements were identified using *de novo* and homology-based approaches. For de novo prediction, Tandem Repeats Finder (TRF v4.0.9)[22] was applied to detect simple tandem repeats with the parameters “2 5 7 80 10 50 2000 -d -h -ngs.” Transposable elements were annotated de novo using RepeatModeler v2.0.7[23] and LTR_FINDER v1.07[24]. For homology-based annotation, we employed RepeatMasker v4.0.6[25] and RepeatProteinMask v1.0.8 with the parameters “-engine ncbi -noLowSimple -pvalue 1e-04”, against the Repbase database [26]. Finally, the annotated repetitive regions were soft-masked by converting them to lowercase letters in the genome assembly using BEDtools v2.29.2[27], generating a masked reference sequence for downstream analysis. Repetitive sequences comprised 498.74 Mb (48.74%) of the genome, including 62.07 Mb of DNA elements, 94.62 Mb of long interspersed nuclear elements (LINEs), 89.73 Mb of short interspersed

nuclear elements (SINEs), 150.88 Mb of tandem repeats (TRF), 6.71 Mb of long terminal repeats (LTRs) and 194.03 Mb of unclassified elements (unknown) (Fig. 2b, circos plot of repetitive elements; Table 2, repeat annotation of *R. chinensis*).

Protein-coding genes were predicted using *de novo*, homology-based protein prediction and transcriptome-based prediction methods. Augustus v2.5.5[28] was employed for de novo prediction. Subsequently, homology-based prediction leveraged protein sequences from *Coptotermes formosanus* (GCA_013340265.1)[29], *Periplaneta americana* (GCF_040183065.1)[30], *Diploptera punctata* (GCA_030220185.1)[31], *Zootermopsis nevadensis* (GCF_000696155.1)[32], and *Cryptotermes secundus* (GCF_002891405.2)[33] obtained from National Center for Biotechnology Information (NCBI). The protein sets from these five species were aligned to the *R. chinensis* genome using BLAT v. 35[34], and gene models were predicted using GeneWise v2.4.1[35] with default parameters. Finally, transcripts with complete open reading frames (ORFs) were mapped to the *R. chinensis* genome using BLAT v. 35[34] and further processed using GeneWise v2.4.1[35] to refine gene structures. Predictions were consolidated using EVidenceModeler v1.1.1[36], resulting in 30,609 protein-coding genes. BUSCO v5.5.0[18] assessment (insecta_odb10 database) indicated 98.9% completeness, including 97.4% complete single-copy, 1.5% complete duplicated, 0.2% fragmented, and 0.9% missing orthologs (Table 3, BUSCO assessment of protein-coding genes in *R. chinensis*).

Finally, repetitive element and protein-coding gene annotations were conducted for the related species *C. formosanus*, *R. speratus*, and *Z. nevadensis* using the same pipeline.

Functional annotation was performed against Gene Ontology (GO) annotations (http://geneontology.org/; 9,498 genes annotated)[37,38] , Swiss-Prot (www.uniprot.org; 17,687 genes annotated)[39,40], TrEMBL (www.uniprot.org; 25,762 genes annotated)[39,40], non-redundant proteins (NR: https://ftp.ncbi.nlm.nih.gov/blast/db; 25,681 genes annotated)[41], Kyoto Encyclopedia of Genes and Genomes (KEGG: https://www.kegg.jp/; 15,802 genes annotated)[42,43], and InterPro (https://www.ebi.ac.uk/interpro/; 17,235 genes annotated) databases[44], using InterProScan v 5.75-106.0[45], with an e-value threshold of 1e-5. In total, 26,328 protein-coding genes were functionally annotated, accounting for 86% of all predicted protein-coding genes (Table 3, statistics of functional annotation of protein-coding genes), and the number of genes annotated on each pseudo-chromosome is shown in the Fig.3.

Data Records

All raw sequencing data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) database under BioProject accession number PRJNA1335780[46]. The genome assembly has been deposited in GenBank under accession number JBRIKC000000000.1[47]. Specifically, the raw data includes genome survey sequencing data

(SRR35653617)[48], PacBio HiFi sequencing data (SRR35653619)[49], Hi-C sequencing data (SRR35653618)[50], and RNA-seq sequencing data (SRR35653616 and SRR35653615)[51,52].

Technical Validation

Genome assembly quality was assessed using multiple methods. BUSCO analysis against the insecta_odb10 database indicated 97.4% complete single-copy orthologs, confirming high completeness. Mapping rates of short reads and long reads to the assembly were 99.56% and 99.99% respectively, indicating robust assembly accuracy. Whole-genome alignment and synteny analysis identified conserved syntenic blocks covering 77.10% of the *R. chinensis* genome and 89.59% of the *R. speratus* genome, and the high collinearity between our assembly and the previously published *R. speratus* genome further supports its high quality. For gene annotation, we re-annotated the genomes of *Coptotermes formosanus* (GCA_013340265.1)[29] and *Reticulitermes speratus* (GCA_021186555.1)[19] using repeat and protein-coding gene annotation methods as described for *R. chinensis,* while also leveraging published annotation data from *Zootermopsis nevadensis* (GCF_000696155.1)[32] as a reference. Compared with these species in terms of repeat content (Table 2, repeat annotation statistics across multiple species), gene number (*C. formosanus*: 25,958; *R. speratus*: 23,601; *Z. nevadensis*: 30,187) and BUSCO completeness (insecta_odb10 database; *C. formosanus*: C:95.3%[S:93.3%,D:2.0%],F:2.7%,M:2.0%,n:1367; *R. speratus*: C:80.7%[S:79.7%,D:1.0%],F:2.6%,M:16.7%,n:1367; *Z. nevadensis*: C:98.9%[S:54.9%,D:44.0%],F:0.4%,M:0.7%,n:1367), the gene annotation of *R. chinensis* in this study is consistent and within a reasonable range. These validations ensure the reliability of the genome assembly and annotations for downstream analyses.

Data Availability

The dataset has been deposited to National Genomics Data Center as a BioProject under accession number PRJNA1335780. The final Genome assembly data of the termite *Reticulitermes chinensis Snyder* is available in the GenBank under the accession number JBRIKC000000000.1. The raw reads generated from three platform specific sequencing runs have been deposited in the NCBI Sequence Read Archive (SRA, accession numbers: SRR35653615 - SRR35653619).

Code Availability

No custom code was developed for this study. All analyses used publicly available bioinformatics tools, with software versions and parameters detailed in the Methods section to ensure reproducibility.

References

1. Govorushko, S. Economic and ecological importance of termites: A global review. Entomological Science 22, 21-35 (2019).
2. Lo, N. et al. Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. Current Biology 10, 801-804 (2000).
3. Khan, Z. et al. A comprehensive review on the documented characteristics of four Reticulitermes termites (Rhinotermitidae, Blattodea) of China. Brazilian Journal of Biology 84, e256354 (2022).
4. Shigenobu, S. et al. Genomic and transcriptomic analyses of the subterranean termite Reticulitermes speratus: Gene duplication facilitates social evolution. Proceedings of the National Academy of Sciences 119, e2110361119 (2022).
5. Martelossi, J. et al. Wood feeding and social living: Draft genome of the subterranean termite Reticulitermes lucifugus (Blattodea; Termitoidae). Insect Molecular Biology 32, 118-131 (2023).
6. Konishi, T., Tasaki, E., Takata, M. & Matsuura, K. King-and queen-specific degradation of uric acid contributes to reproduction in termites. Proceedings of the Royal Society B 290, 20221942 (2023).
7. Peñafiel, N., Flores, D.M., Rivero De Aguilar, J., Guayasamin, J.M. & Bonaccorso, E. A cost-effective protocol for total DNA isolation from animal tissue. Neotropical Biodiversity 5, 69-74 (2019).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764-770 (2011).
10. Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature communications 11, 1432 (2020).
11. 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 (2021).
12. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36, 2896-2898 (2020).
13. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics 25, 1754-1760 (2009).
14. Zhou, C., McCarthy, S.A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics 39, btac808 (2023).
15. Durand, N.C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell systems 3, 99-101 (2016).
16. Martins, V.G. Karyotype evolution in the Termitidae (Isoptera). (1999).
17. Jankásek, M. & Varadínová, Z.K. Blattodea karyotype database. European Journal of Entomology 118, 192-199 (2021).
18. 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. Molecular biology and evolution 38, 4647-4654 (2021).
19. Sawada, Y. et al. Unsupervised AI reveals insect species-specific genome signatures. PeerJ 12, e17025 (2024).
20. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P. & Frith, M.C. Adaptive seeds tame genomic sequence comparison. Genome research 21, 487-493 (2011).
21. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome research 19, 1639-1645 (2009).
22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 27, 573-580 (1999).
23. Flynn, J.M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences 117, 9451-9457 (2020).
24. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research 35, W265-W268 (2007).
25. Bedell, J.A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics 16, 1040-1041 (2000).
26. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile Dna 6, 11 (2015).
27. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).
28. Hoff, K.J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. Current protocols in bioinformatics 65, e57 (2019).
29. Hellemans, S. et al. Genomic data provide insights into the classification of extant termites. Nature Communications 15, 6724 (2024).
30. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763-767 (2014).
31. Fouks, B. et al. Live-bearing cockroach genome reveals convergent evolutionary mechanisms linked to viviparity in insects and beyond. Iscience 26(2023).
32. Terrapon, N. et al. Molecular traces of alternative social organization in a termite genome. Nature communications 5, 3636 (2014).
33. Fraser, R. et al. Evidence for a novel X chromosome in termites. Genome Biology and Evolution 16, evae265 (2024).
34. Kent, W.J. BLAT—the BLAST-like alignment tool. Genome research 12, 656-664 (2002).
35. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. Genome research 14, 988-995 (2004).
36. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and

the Program to Assemble Spliced Alignments. Genome biology 9, R7 (2008).
37. Ashburner, M. et al. Gene ontology: tool for the unification of biology. Nature genetics 25, 25-29 (2000).
38. Aleksander, S.A. et al. The gene ontology knowledgebase in 2023. Genetics 224, iyad031 (2023).
39. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic acids research 25, 31-36 (1997).
40. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research 28, 45-48 (2000).
41. Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic acids research 51, D29 (2022).
42. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 27-30 (2000).
43. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. Nucleic acids research 49, D545-D551 (2021).
44. Blum, M. et al. The InterPro protein families and domains database: 20 years on. Nucleic acids research 49, D344-D354 (2021).
45. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236-1240 (2014).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP629289 (2025).
47. Xin, P. Reticulitermes chinensis isolate PX-2025, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JBRIKC000000000 (2025).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35653617 (2025).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35653619 (2025).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35653618 (2025).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35653616 (2025)
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35653615 (2025).

Author Contributions

Jia Wu, Zhiyong Yue, Chenguang Feng, and Jihu Sun contributed to the research design. Jinpei Wang, Qi Jiang, Baozhen Zhou, and Jia Wu collected the samples. Peidong Xin and Zhiyong Yue analyzed the data. Peidong Xin and Chenguang Feng contributed to data quality control. Zhiyong Yue, Peidong Xin, and Jia Wu wrote the draft manuscript and revised the manuscript. All co-authors contributed to this manuscript and approved it.

Competing Interests

The authors declare that there are no competing interests that could influence the objectivity of this study.

Acknowledgements

This work was supported by the General Program of Shaanxi Natural Science Foundation [2025JC-YBMS-247], the Fundamental Research Funds for the Central Universities, Northwestern Polytechnical University [D5000220464], the Scientific and Technological Plan Project of Xi'an Science and Technology Bureau [24GXFW0080], and the Initiation Funds for High level Talents Program of Xi’an International University [XAIU202402].

Figure legends

Fig.1 Genome assembly of *Reticulitermes chinensis*.
(a) K-mer analysis (21-mer) of Illumina short reads estimates a genome size of 1,005,249,533 bp and heterozygosity of 0.5%.
(b) Hi-C linkage density heatmap of *R. chinensis*. The x- and y-axes represent genomic positions, with red dots indicating high-density paired reads, suggesting chromosomal proximity.

Fig. 2 Overview of the *Reticulitermes chinensis* genome.
(a) Synteny alignment between *R. chinensis* (blue circle, chromosomes) and *R. speratus* (grey circle, scaffolds).
(b) Density of protein-coding genes, repeat sequences, and GC content across chromosomes 1–21, shown in 10 Mb windows.

Fig. 3 Distribution of annotated genes across the *Reticulitermes chinensis* genome.
The x-axis represents the 21 pseudo-chromosomes and unplaced scaffolds ("others"). The y-axis indicates the gene count.

Tables

Table 1. Statistics of the genome assembly.

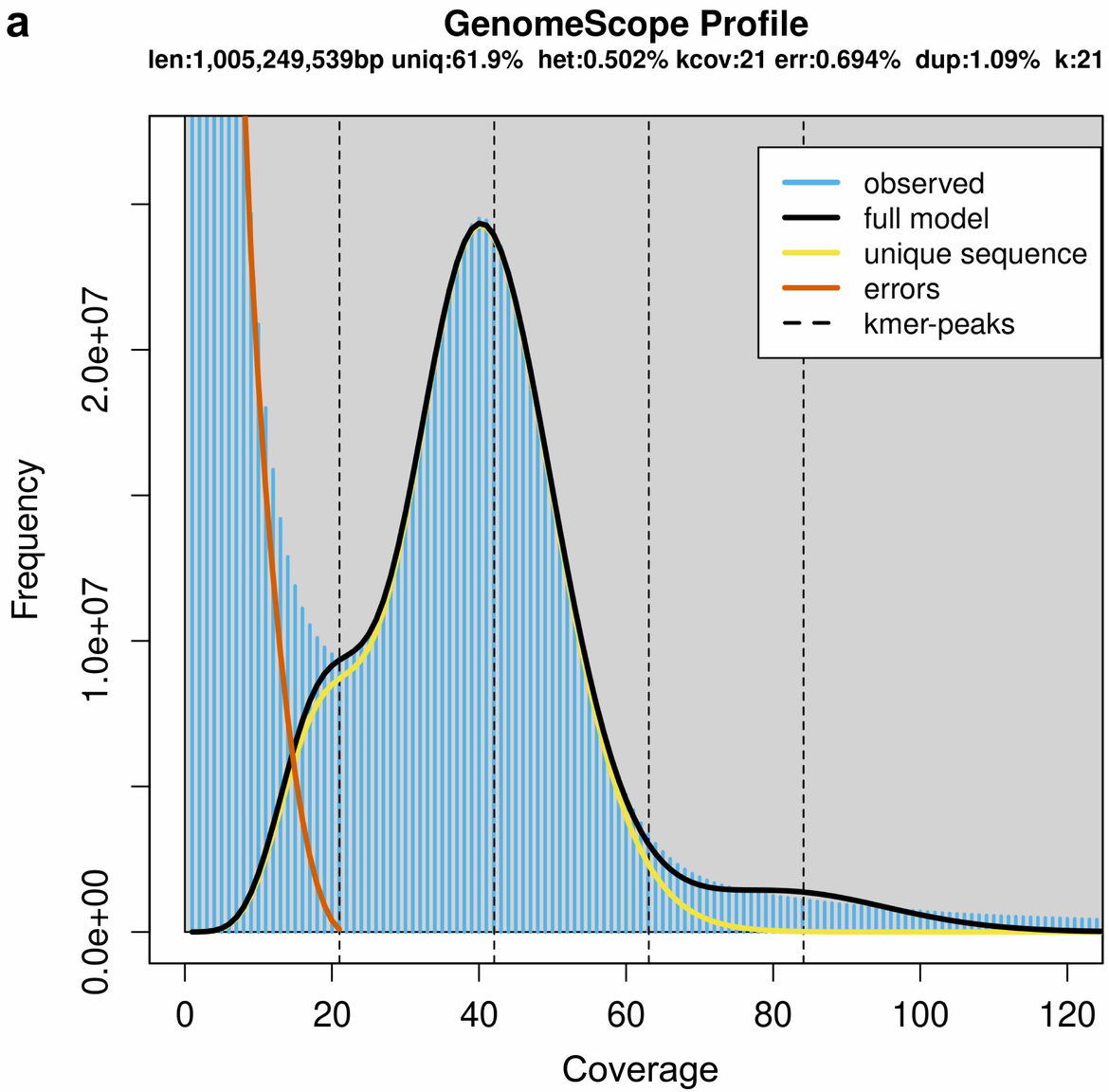
Term	HiFi Number	Hi-C Number
Total sequence size (bp)	1,023,208,222	1,023,293,324
Total sequence number	853	367
Average length (bp)	1,199,540	2,788,265
N90 (bp)	777,275	32,204,279
N90 number	318	20
N80 (bp)	1,424,706	35,933,611
N80 number	223	17
N70 (bp)	1,913,000	39,638,661
N70 number	162	15
N60 (bp)	2,435,996	44,642,800
N60 number	115	12
N50 (bp)	3,344,424	47,294,433
N50 number	78	10
N number	0	85,102
N rate %	0	8.32E-05
GC content %	29.42	29.46
Complete BUSCOs (%)	98	98.4
Complete single-copy BUSCOs (%)	95.8	96.8
Complete duplicated BUSCOs (%)	2.2	1.6

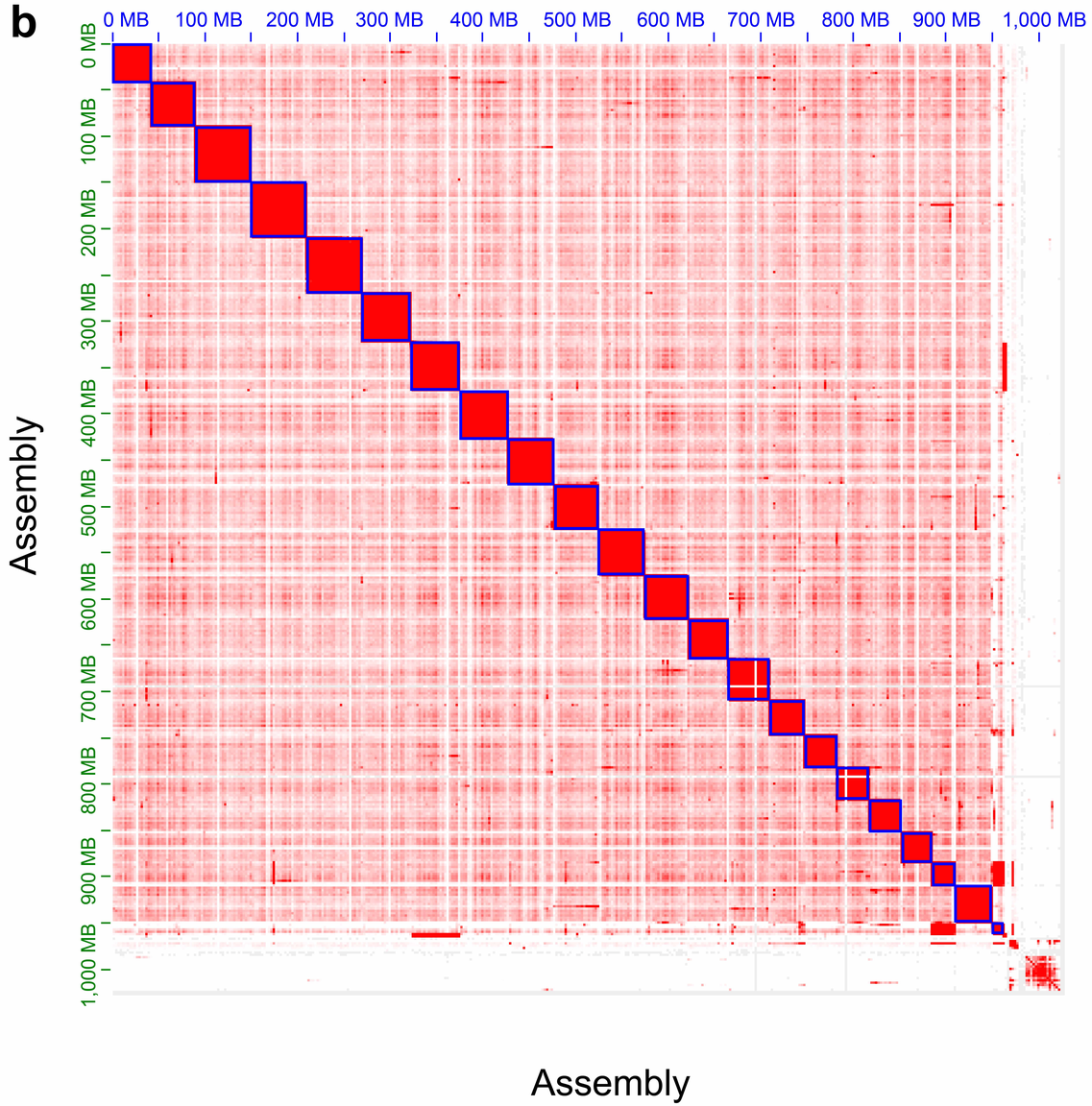
Table 2. Summary of the repetitive sequences in *R. chinensis, C. formosanus, R. speratus and Z. nevadensis* genome

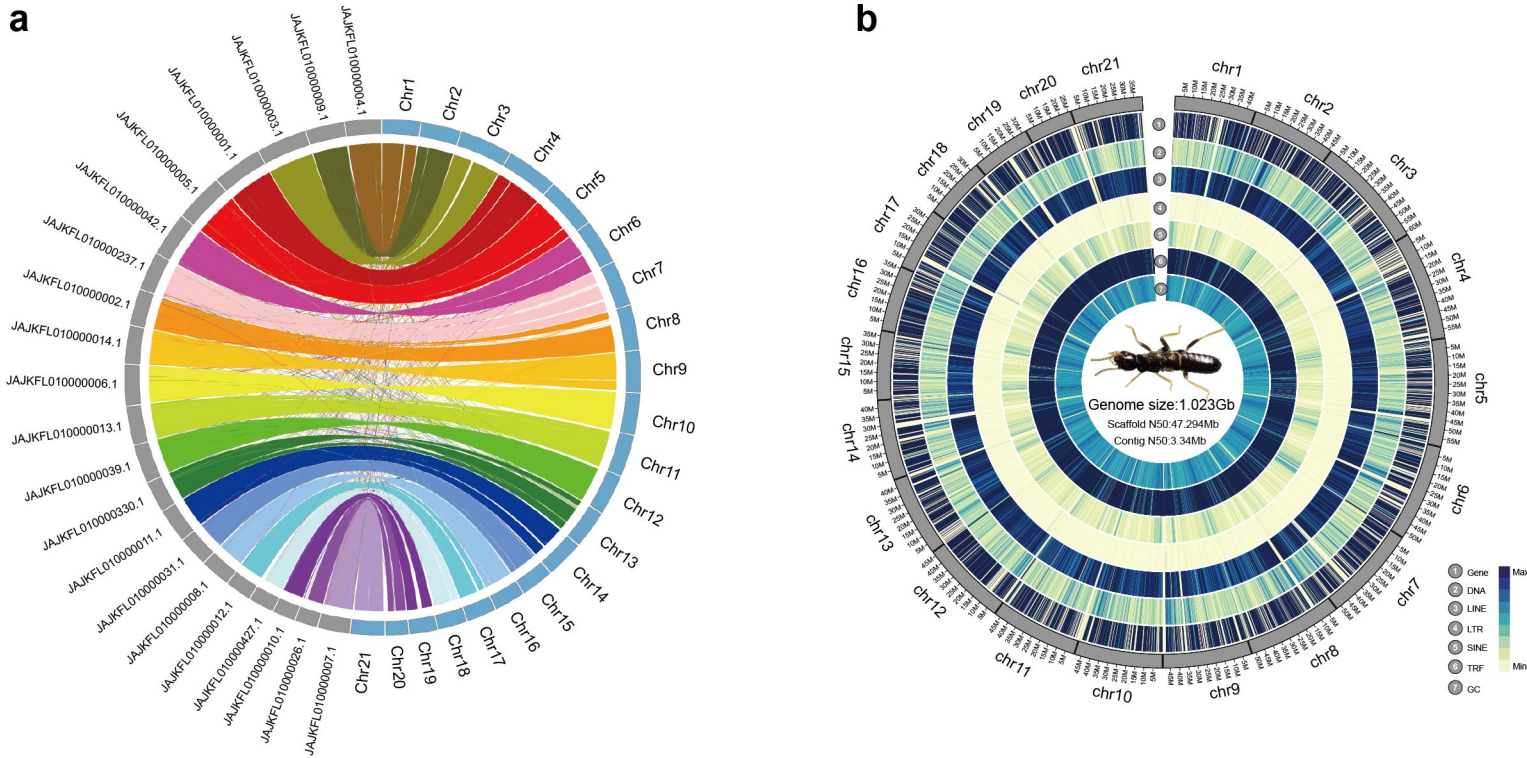
Type	*R. chinensis* Length (bp)	*C. formosanus* Length (bp)	*R. speratus* Length (bp)	*Z. nevadensis* Length (bp)
SINE	89,728,487	61,698,303	79,989,998	13,245,651
LINE	94,623,280	69,655,985	85,210,840	31,701,711
LTR	6,712,222	4,991,102	5,354,218	1,108,421
DNA	62,071,457	49,104,008	59,199,613	19,146,192
TRF	150,883,204	31,443,398	91,715,274	8,534,730
Satellite	1,006,151	763,779	693,263	17,755
Simple	13,594,505	16,176,466	12,928,625	3,966,556
Unknown	194,033,941	164,612,275	136,725,563	60,015,156
total	498,740,045 (48.74%)	356,715,075 (41.89%)	417,875,139 (47.45%)	130,654,087 (28.14%)

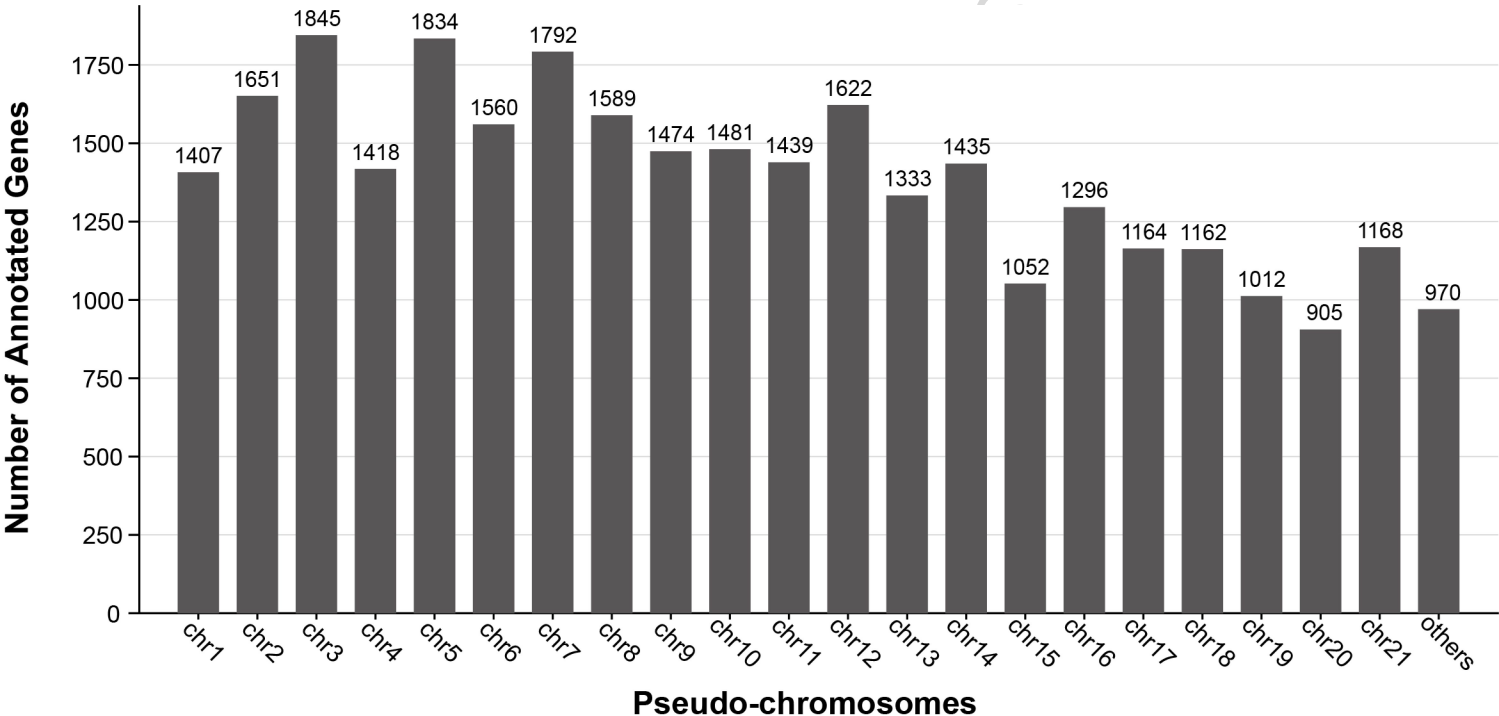
Table 3. Annotation statistics of protein-coding genes in the genome of *R. chinensis*

Libraries	Number
Total	30,609
Complete BUSCOs (%)	98.9
Complete single-copy BUSCOs (%)	97.4
Complete duplicated BUSCOs (%)	1.5
Functionally Annotated genes	26,328 (86.01%)
Missing genes	4,281 (13.99%)
GO	9,498 (31.03%)
Swiss-Prot	17,687 (57.78%)
TrEMBL	25,762 (84.16%)
NR	25,681 (83.90%)
KEGG	15,802 (51.63%)
InterPro	17,235 (56.31%)






a
JAJKFL010000004.1
JAJKFL010000009.1
JAJKFL010000003.1
JAJKFL010000001.1
JAJKFL010000005.1
JAJKFL010000042.1
JAJKFL010000237.1
JAJKFL010000002.1
JAJKFL010000014.1
JAJKFL010000006.1
JAJKFL010000013.1
JAJKFL010000039.1
JAJKFL010000330.1
JAJKFL010000011.1
JAJKFL010000031.1
JAJKFL010000008.1
JAJKFL010000012.1
JAJKFL010000427.1
JAJKFL010000010.1
JAJKFL010000026.1
JAJKFL010000007.1
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
b
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
Genome size:1.023Gb
Scaffold N50:47.294Mb
Contig N50:3.34Mb
1 Gene
2 DNA
3 LINE
4 LTR
5 SINE
6 TRF
7 GC
Max
Min


Number of Annotated Genes
0
250
500
750
1000
1250
1500
1750
1407
1651
1845
1418
1834
1560
1792
1589
1474
1481
1439
1622
1333
1435
1052
1296
1164
1162
1012
905
1168
970
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
others
Pseudo-chromosomes