scientific data

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of eriophyoid mite *Setoptus koraiensis*

Zi-Kai Shao, Lei Chen, Jing-Tao Sun & Xiao-Feng Xue ✉

Eriophyoidea represents a highly diverse superfamily of herbivorous mites in the Acariformes, including over 5,000 named species that are distributed worldwide. However, the lack of chromosome-level genome prevents our understanding of the evolution in this group. Here, we report the first chromosome-level genome assembly of *Setoptus koraiensis* using Illumina, PacBio, and Hi-C sequencing technologies. The assembled genome has a size of 47 Mb with an N50 of 24.53 Mb, anchored into two chromosomes. The chromosome-level genome assembly had a BUSCO completeness of 89%. We identified 5,954 protein-coding genes, with 4,770 genes that could be functionally annotated. This genome provides resources to further understand the genetic and evolution of eriophyoid mites.

Background & Summary

Eriophyoid mites (Acariformes, Eriophyoidea) are among the largest superfamilies in the Arachnida, comprising over 5,000 name species[1,2] and exhibiting a worldwide distribution[3]. These tiny (~200 um in length, among the smallest arthropods), vermiform to fusiform mites have only two pairs of legs, and are strictly phytophagous, reflecting high hostplant specificity[4,5]; some of them can cause massive economic losses in agriculture and forestry[6].

Despite the need to understand the ecology and evolution among eriophyoid mites, there are no chromosome-level assembled genomes for eriophyoid mites yet. A near chromosome genome assembly has been published for tomato russet mite *Aculops lycopersici*[7], but the lack of high-quality chromosome-level genome resources has limited further comparative genomic analyses among eriophyoid mites.

In this study, we assembled a chromosome-level genome for the *Setoptus koraiensis* (Eriophyoidea, Phytoptidae) using PacBio long-reads sequencing, Illumina short-reads sequencing, and high-throughput chromatin conformation capture (Hi-C) sequencing. Our assembly resulted in a genome size of 47 Mb across two chromosomes, with scaffold N50 lengths of 24.53 Mb (Table 1). This genome is the first chromosome-level genome among eriophyoid mites, providing significant new data resources for understanding the Eriophyoidea.

Methods

Sample collection.

At least 100,000 wild *S. koraiensis* individuals, including eggs, juveniles and adults, were collected from *Pinus koraiensis* Siebold & Zucc. (Pinaceae), in Lishui, Nanjing city, Jiangsu province, China (31.3921°N, 118.5417°E). Samples were identified by morphological characteristics with molecular evidence (mitochondrial COI). Vouchers were deposited in the Arthropod/Mite Collection of the Department of Entomology, Nanjing Agricultural University, Jiangsu Province, China.

Genome sequencing.

Genomic DNA was extracted from more than 100,000 individuals using MagAttract HMW DNA Kit. The Pacbio 30 kb SMRTbell library was prepared with more than 5 μg gDNA using the SMRTbell^{TM} Prep Kit 2.0 (Pacific Biosciences). The mode of Continuous Long Read (CLR) was run on the Sequel II platform. Illumina whole-genome sequencing was prepared using a 350 bp-insert fragment library (150 bp paired-end) by Truseq DNA PCR-free Kit, which was further sequenced on an Illumina NovaSeq 6000 platform. High-throughput chromosome conformation capture (Hi-C) included cross-linking, HindIII restriction enzyme digestion, end repair, DNA cyclization, purification and capture. The Hi-C library with 300–700 bp insert size library was sequenced on the NovaSeq 6000 platform. Finally, we generated 24.25 Gb (~496X) PacBio long reads, 9.5 Gb (~194X) Illumina short reads, and 9 Gb Hi-C (~184X) reads for our genome assembly.

Department of Entomology, Nanjing Agricultural University, Nanjing, Jiangsu, 210095, China. ✉e-mail: xfxue@njau.edu.cn

Characteristics	Setoptus koraiensis
Genome Size (Mb)	47
Number of contigs	266
Number of chromosomes	2
Scaffold N50 length (Mb)	24.53
BUSCO completeness (%)	89
Repetitive elements Size (Mb)	6.42 (13.82%)

Table 1. Statistics of *Setoptus koraiensis* genome assembly. **State**: We would be happy to be published without further edits.

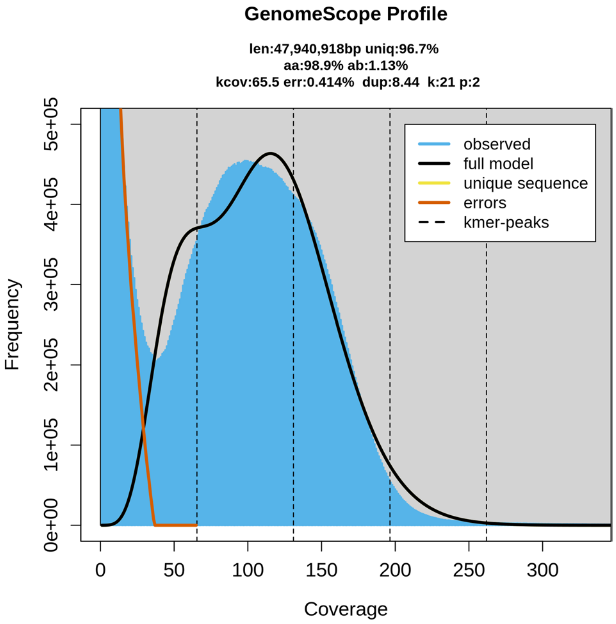

Fig. 1 GenomeScope genome size estimates for *Setoptus koraiensis*.

Genome survey. Duplicate and low-quality Illumina raw reads (base quality < Q20, length < 15 bp, polymer A/G/C/ > 10 bp) were trimmed and removed using BBtools package v38.82[8]. The 21-mer depth distribution was counted using script 'khist.sh' of BBtools. Genome Scope v2.0[9] was used to estimate the genome size and heterozygosity of *S. koraiensis* with the maximum kmer coverage at 1,000×. Based on the distribution of kmer coverage and frequency, the estimated genome size of *S. koraiensis* was 45.72 Mb, with a heterozygosity rate of around 1.13% and a repeat content proportion of approximately 3.3% (Fig. 1).

Genome assembly. The CLR reads were set as input to Flye v2.6[10] to assemble continuous long reads. One round of built-in long reads polishing was performed by Flye v2.6. Then, two rounds of short reads were used to polish and fill in gaps of the primary assembly with NextPolish v1.4.1[11]. Haplotigs and duplication caused by haplotype divergence were eliminated by Purge_dups v1.2.5[12] using the alignment program Minimap2 v2.28[13]. Hi-C reads were aligned to the purged genome using BWA v0.7.18[14] and Juicer v1.6[15] to anchor, order and orient contigs into chromosomal assembly following 3D-DNA[16] pipeline. Then, we manually reviewed and corrected assembled errors using Juicebox v2.17[17]. Contaminations were checked and deleted against the UniVec and NCBI nucleotide databases using BLAST + v2.11.0[18] and MMseqs2 v16[19]. The completeness of genome assembly was evaluated by BUSCO version 5.2.2[20] using the eukaryota_odb10 dataset (creation date 2020-09-10). The reads from the whole genome sequencing were aligned back to the genome assembly to access the mapping rate. After de novo assembly, polishing and contaminant removal, the *S. koraiensis* genome has a genome size of 49.9 Mb with 565 scaffolds, an N50 length of 24.53 Mb, with 94.2% of assembled genomes anchored to two chromosomes (Fig. 2) resulting in a final genome size of 47 Mb (Table 1).

Genome annotation. The repetitive elements were identified using RepeatModeler v2.0.5[21], which discovered the complete long terminal repeats (LTR) with the '-LTRstruct' pipeline. RepeatMasker v4.1.6[22] was searched against the custom repeat library of Dfam 3.8[23] and Repbase v20181026[24] with options '-no_is -norna -xsmall -q' to soft mask repeats of the genome assembly.

For gene structure annotation, we performed a pipeline integrating *ab initio* and homolog-based methods. Braker v2.1.5[25] was used to obtain *ab initio* gene predictions employing GeneMark-ES/ET/EP v4.33[26] and Augustus v3.4.0[27] based on reference proteins from the OrthoDB v11 database[28]. GeMoMa v1.9[29] was used for

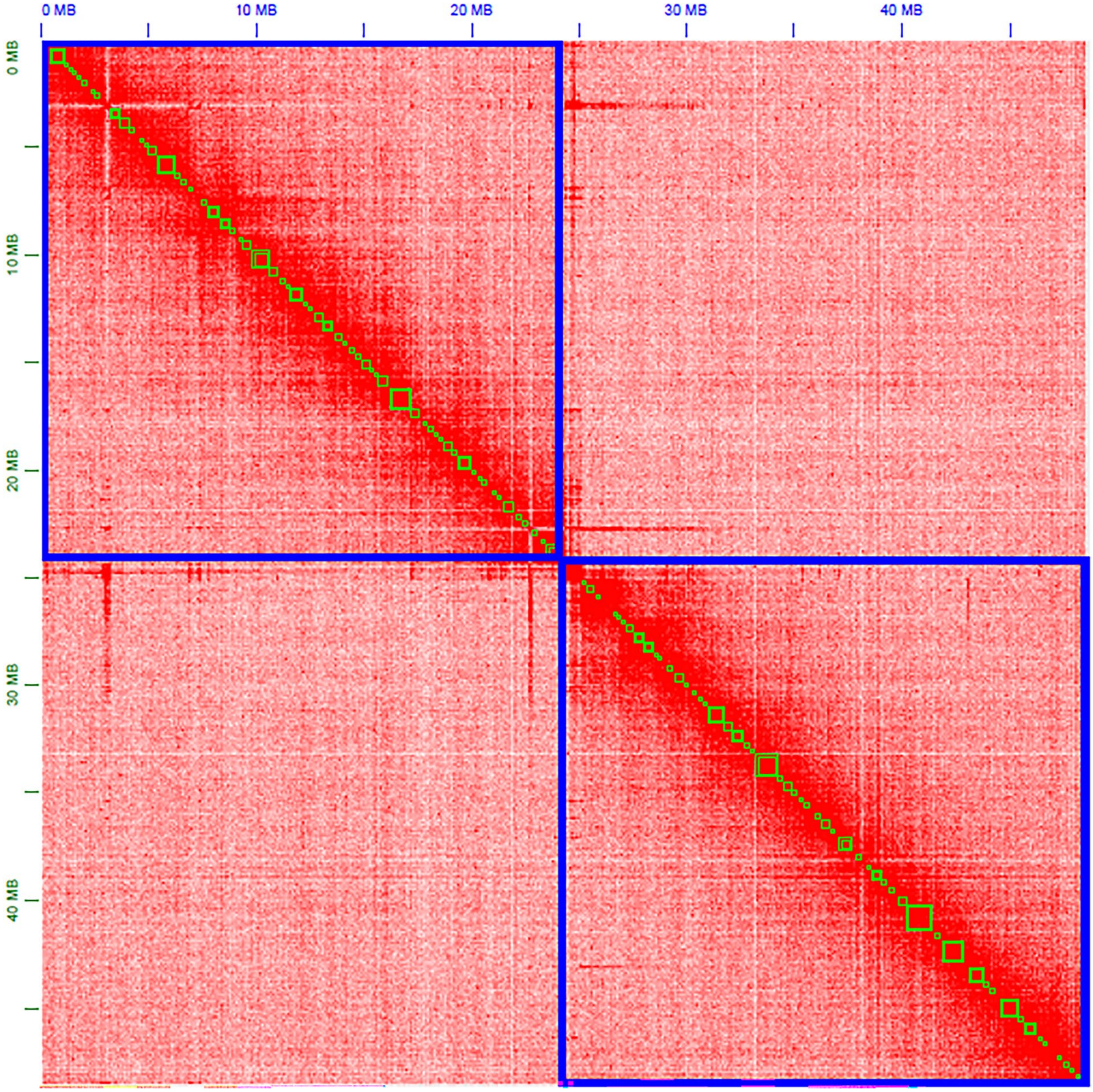

Fig. 2 Genome-wide chromosomal heatmap of *Setoptus koraiensis*, the blue boxes show super scaffolds.

homology prediction with the parameters "GeMoMa.c = 0.4 GeMoMa.p = 10", and the protein sequences of six species (*Aculops lycopersici* (GCA_015350385.1), *Tetranychus urticae* (GCA_039701765.1), *Tetranychus piercei* (GCA_036759885.1), *Panonychus citri* (GCA_014898815.1), *Pyemotes zhonghuajia* (GCA_025170145.1), *Blomia tropicalis* (GCA_029204025.1)) were provided to assist gene prediction. The results obtained from BRAKER and GeMoMa were combined and provided to MAKER v3.01.03[30]. The functional annotation of predicted protein sequences was searched against UniProt, InterProScan and eggNOG databases. Diamond v2.1.10[31] was used to assign the gene function of the best hits in the UniProt database under the 'very sensitive' mode. Gene Ontology (GO) and pathway (KEGG) were annotated using InterProScan v5.72[32] and eggnog-mapper v2.1.12[33] against Pfam[34], SMART[35], Superfamily[36], CDD[37], and EggNOG 5.0.2 database[38].

Data Records

The raw reads and genome assembly have been deposited in the NCBI databases under BioProject PRJNA1196018. The PacBio, Illumina, and Hi-C data are available under identification numbers SRR32458739-SRR32458741[39]. The final chromosome assembly has been deposited at GenBank under the accession number GCA_048013815.1[40]. The mitochondrial COI sequence has been deposited at GenBank under the accession number PV163833[41]. The genome assembly and annotation files are available in Figshare[42].

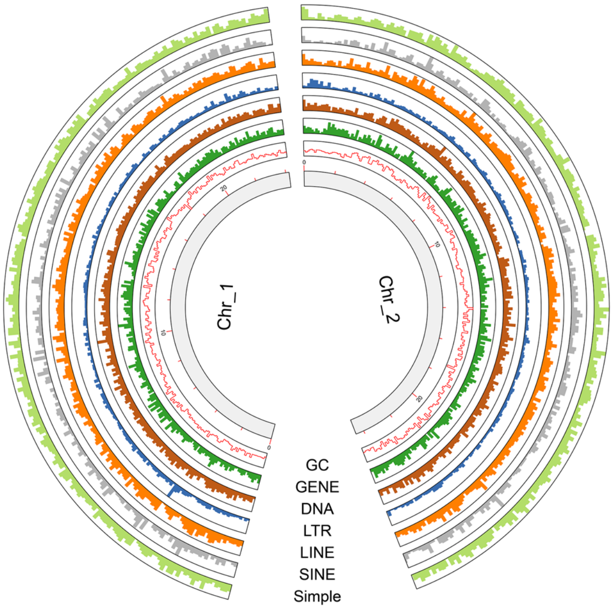

Fig. 3 Circular karyotype representation of the chromosomes of *Setoptus koraiensis*. Tracks from inside to outside are GC content (GC), density of protein-coding genes (GENE), DNA transposons (DNA), LTR/LINE/SINE retrotransposons (LTR, LINE, SINE), and simple repeats (Simple).

Technical Validation

We mapped the Illumina sequencing data to the final assembly with BWA v0.7.18, and the mapping rate was 92.9%. We assessed the completeness of the genome assembly using BUSCO v5.4.2 with the 'eukaryota_odb10' database, and a total of 89% (83.9% single-copied genes, 5.1% duplicated genes, 5.5% fragmented, and 5.5% missing genes) completed BUSCOs were identified, which is higher than that of *A. lycopersici* (86.3%). We masked 13.82% (6.42 Mb) repetitive regions of the *S. koraiensis* genome. Among them, 0.2% of repeat sequences were short interspersed elements (SINEs), 1.29% were long interspersed elements (LINEs), 0.92% were long terminal repeats (LTRs), 1.61% were DNA transposons, and 5.14% were unclassified (Fig. 3). We identified 5,954 protein-coding genes, with 4,770 genes that could be functionally annotated. The BUSCO completeness for protein sequence is 77.3% (71.4% single-copied genes, 5.9% duplicated genes, 3.9% fragmented, and 18.8% missing genes) with the 'eukaryota_odb10' database. All evidence strongly supported the completeness and accuracy of *S. koraiensis* genome assembly.

Code availability

No custom scripts or code were used in this study.

Received: 27 December 2024; Accepted: 12 March 2025;
Published online: 18 March 2025

References

1. Zhang, Z.-Q. Eriophyoidea and allies: where do they belong? *Syst. Appl. Acarol.* **22**, 1091–1095 (2017).
2. Zhang, Z.-Q. Phylum Arthropoda von Siebold, 1848. in *Animal biodibersity: An Outline of Higher-Level Classification and Survey of Taxonomic Richness* (ed. Zhang, Z.-Q.) 99–103 (Magnolia Press, 2011)
3. Li, N., Sun, J.-T., Yin, Y., Hong, X.-Y. & Xue, X.-F. Global patterns and drivers of herbivorous eriophyoid mite species diversity. *J. Biogeogr.* **50**, 330–340 (2022).
4. Skoracka, A., Smith, L., Oldfield, G., Cristofaro, M. & Amrine, J. W. Host-plant specificity and specialization in eriophyoid mites and their importance for the use of eriophyoid mites as biocontrol agents of weeds. *Exp. Appl. Acarol.* **51**, 93–113 (2010).
5. Yin, Y. *et al.* DNA barcoding uncovers cryptic diversity in minute herbivorous mites (Acari, Eriophyoidea). *Mol. Ecol. Resour.* **22**, 1986–1998 (2022).
6. de Lillo, E., Pozzebon, A., Valenzano, D. & Duso, C. An intimate relationship between eriophyoid mites and their host plants–a review. *Front. Plant. Sci.* **9**, 1786 (2018).
7. Greenhalgh, R. *et al.* Genome streamlining in a minute herbivore that manipulates its host plant. *Elife* **9** (2020).
8. Bushnell, B. BBtools. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 October 2024) (2014).
9. 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).
10. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
11. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
12. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
13. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

15. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell. Syst.* **3**, 95–98 (2016).
16. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
17. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell. Syst.* **3**, 99–101 (2016).
18. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
19. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
20. Manni, M., Berkeley, M. R., Seppey, M., Simao, 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).
21. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
22. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: http://www.repeatmasker.org (accessed on 1 October 2024) (2013–2015).
23. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic. Acids. Res.* **44**, D81–89 (2016).
24. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
25. Bruna, 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. Genom. Bioinform.* **3**, lqaa108 (2021).
26. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR. Genom. Bioinform.* **2**, lqaa026 (2020).
27. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic. Acids. Res.* **32**, W309–312 (2004).
28. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic. Acids. Res.* **51**, D445–D451 (2023).
29. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic. Acids. Res.* **44**, e89 (2016).
30. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
31. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
32. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic. Acids. Res.* **45**, D190–D199 (2017).
33. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
34. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic. Acids. Res.* **47**, D427–D432 (2019).
35. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic. Acids. Res.* **46**, D493–D496 (2018).
36. Wilson, D. *et al*. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic. Acids. Res.* **37**, D380–386 (2009).
37. Marchler-Bauer, A. *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic. Acids. Res.* **45**, D200–D203 (2017).
38. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic. Acids. Res.* **47**, D309–D314 (2019).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP565774 (2024).
40. Shao, Z.-K. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_048013815.1 (2025).
41. Shao, Z.-K. *GenBank* https://identifiers.org/ncbi/insdc:PV163833 (2025).
42. Shao, Z.-K., Chen, L., Sun, J.-T. & Xue, X.-F. A chromosome-level genome assembly of eriophyoid mite *Setoptus koraiensis*. *figshare* https://doi.org/10.6084/m9.figshare.28087958 (2025).

Acknowledgements

This research was funded by the National Natural Science Foundation of China (32170466). This work was also supported by the high-performance computing platform of Bioinformatics Center, Nanjing Agricultural University.

Author contributions

Z.-K.S. and X.-F.X. conceived and designed the study. Z.-K.S. analyzed the data. X.-F.X., L.C. and J.-T.S. had substantial contributions to the interpretation of the data, writing, and review of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.-F.X.

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.