scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Schizodactylus jimo* He, 2021 from China (Ensifera: Schizodactylidae)

Binbin Yao[1,4], Tao Jiang[1,4], Lulu Yang[1], Jiawei Wang[1], Delong Guan[2,3]✉ & Sheng-Quan Xu[1]✉

The Schizodactylidae family, an ancient lineage within Orthoptera, is ideal for studying adaptive evolution due to its sand-dwelling habits and unique tarsi, but genomic data scarcity has hindered research. To fill this knowledge gap, this study presents the first chromosome-level genome assembly and annotation for *Schizodactylus jimo*, a species endemic to China known for its unique adaptations to sandy environments. Using a combination of PacBio HiFi long-read sequencing and Hi-C scaffolding technologies, we constructed a high-quality reference genome. The final assembly spans 1.192 Gb with an impressive scaffold N50 of 198.77 Mb. A total of 94.65% of the sequence was successfully anchored into 9 pseudochromosomes. The genome's integrity was confirmed by a 98.0% BUSCO completeness score. Annotation identified that 40.51% of the genome consists of repetitive elements and predicted a total of 14,215 protein-coding genes. Of these genes, 12,747 (accounting for 89.67%) have had their functions annotated in different databases.

Background & Summary

The order Orthoptera represents an ancient and highly diverse lineage within the class Insecta, comprising over 29,000 described species. Members of this order fulfill critical ecological roles as herbivores, predators, and decomposers in terrestrial ecosystems[1,2]. In recent years, the rapid advancement of genome sequencing technologies has propelled genomics into the forefront of research aimed at elucidating the genetic underpinnings of rapid speciation, unique adaptations, and phylogenetic relationships within Orthoptera[3–5]. However, a notable characteristic of orthopteran genomes is their considerable size, which has presented challenges in assembly. Consequently, high-quality, chromosome-level reference genomes remain scarce, substantially limiting in-depth investigations into the macroevolution, chromosomal structural variations, and the evolution of key functional gene families across this taxon[6].

As an ancient and phylogenetically pivotal lineage of Orthoptera, the family Schizodactylidae is critical for understanding the evolution of Ensifera[7]. Its systematic position is a key node for clarifying the evolutionary relationships between major superfamilies like Grylloidea and Tettigonioidea[8,9]. Phylogenetic evidence indicates that Schizodactylidae diverged from its sister taxon, Grylloidea, no later than the Triassic period[7,10]. The extant representatives of this family exhibit a stark disjunct distribution, comprising two subfamilies: the Asian Schizodactylinae and the southern African Comicinae. These insects are well-known for their unique morphological specializations, including tarsi adapted for sandy substrates and their fossorial, predatory lifestyles[11,12]. Despite their evolutionary significance, this family remains a genomic desert; consequently, the genetic mechanisms underlying their distinct morphology and ecological adaptations are largely unknown.

Schizodactylus jimo, the sole species of dune cricket currently documented in China, is narrowly distributed in Yunnan Province[13]. It is a classic nocturnal, carnivorous insect that shelters in self-constructed burrows during the day and emerges at night to prey on small arthropods. Its most striking feature is the presence of expandable, paddle-like lobes on its terminal tarsomeres, which are considered a key adaptation for efficient locomotion

[1]College of Life Sciences, Shaanxi Normal University, Xi'an, China. [2]Guangxi Key Laboratory of Sericulture Ecology and Applied Intelligent Technology, Hechi University, Hechi, China. [3]Guangxi Collaborative Innovation Center of Modern Sericulture and Silk, Hechi University, Hechi, China. [4]These authors contributed equally: Binbin Yao, Tao Jiang. ✉e-mail: 2023660006@hcnu.edu.cn; xushengquan@snnu.edu.cn

Library type	Data size (Gb)	Average reads length (bp)
HIFI(PacBio)	23.47	16,554
Hi-C	157.76	150
Nova-seq(Illumina)	82.83	150
RNA-seq	11.95	150

Table 1. Statistics of the DNA sequence data used for genome assembly.

and digging in loose sand. This unique mode of movement and its associated morphological structures make *S. jimo* an ideal model for studying the evolution of adaptations to specialized environments. However, the lack of high-quality genomic resources has made it difficult to understand the evolutionary driving forces of these complex traits at the molecular level.

To address this knowledge gap and provide a crucial reference point for comparative genomics and evolutionary biology in Orthoptera, this study presents the first chromosome-level genome assembly and annotation for *S. jimo*, generated using PacBio long-read sequencing and Hi-C scaffolding technologies. The successful construction of this high-quality genome will not only greatly facilitate the understanding of the genetic basis of the unique biological characteristics of Schizodactylidae but will also provide new insights and data to support research on adaptive evolution and biodiversity conservation. Furthermore, it offers an invaluable genetic resource for resolving persistent phylogenetic questions and for reconstructing the history of chromosome evolution within Orthoptera.

Methods & Results

Sample collection, DNA and RNA extraction. *S. jimo* specimens were procured on May 1st, 2023 from the right bank of the Nujiang River, located in Baoshan, Yunnan Province, China. For genomic analysis, muscle tissue was excised from two male specimens, and high-molecular-weight DNA was isolated using the cetyltrimethylammonium bromide (CTAB) method. The quality of the resultant DNA was rigorously assessed; its integrity was visualized on an agarose gel, purity was determined spectrophotometrically (NanoDrop), and the concentration was precisely measured using a Qubit fluorometric assay. In parallel, to aid in gene model prediction and genome annotation, transcriptomic data were generated. Total RNA was extracted from dissected head and muscle tissues via a magnetic bead protocol. The integrity of the extracted RNA was verified by gel electrophoresis prior to library preparation. All nucleic acid samples that passed these stringent quality control criteria were submitted for next-generation sequencing.

Genome sequencing. For Illumina short-read sequencing, A short-read sequencing library was prepared using the TruSeq Nano DNA HT Sample Preparation Kit (Illumina Inc., San Diego, CA, USA). Following construction, the library was initially quantified using a Qubit 2.0 Fluorometer. The insert size distribution was then assessed with an Agilent 2100 Bioanalyzer. Once the insert size was confirmed to be within the expected range, the effective concentration of the library was accurately quantified by qPCR to ensure quality. The qualified library was then sequenced on an Illumina HiSeq X-Ten platform.

For long-read sequencing, a library constructed from long DNA fragments was sequenced on a PacBio Sequel instrument using SMRT (Single Molecule, Real-Time) technology (Pacific Biosciences of California, Inc., Menlo Park, CA, USA). The raw reads were subsequently processed using the SMRT Link software suite to generate high-fidelity (HiFi) reads.

Hi-C libraries were constructed following a standard protocol. Briefly, chromatin was cross-linked, digested with the DpnII restriction enzyme, and the resulting fragments were labeled with biotin. After proximity ligation and purification of ligated junctions, the final library was sequenced on an Illumina NovaSeq. 6000 platform to generate 150 bp paired-end (PE150) reads.

To support high-quality gene annotation, total RNA was extracted from a pool of diverse tissues (including head, thorax, abdomen, and legs) of another individual for transcriptome sequencing. Transcriptome libraries were subsequently constructed following the standard protocol of the New England BioLabs (NEB) kit[14]. The insert size of the libraries was validated using an Agilent 2100 Bioanalyzer. Subsequently, the effective library concentration was precisely determined by quantitative real-time PCR (qRT-PCR). Libraries that passed quality control were pooled according to their effective concentrations and target data output, followed by sequencing on an Illumina NovaSeq. 6000 platform.

Finally, 82.83 Gb of Illumina short-read data, 23.47 Gb of HiFi data, 157.76 Gb of Hi-C data, as well as 11.95 Gb of transcriptome data were obtained. (Table 1).

Estimation of genome size. The Illumina short-reads were employed to get a preliminary understanding of the genome size. We used the Trimmomatic v0.38 to quality control[15]. The high-quality reads were used for genome size estimation by the K-mer method. The Jellyfish tool from Trinity v2.15.1 software computed the k-mer number and distribution ($k = 21$)[16]. GenomeScope v2.0[17] estimated the genome size and performed depth frequency distribution analysis. The genome size was determined to be 1.178 Gb (Fig. 1a).

Genome assembly. First, HiFi reads and Hi-C reads were assembled with Hifiasm v0.19.4-r575[18], with default parameters. Subsequently, redundant sequences were identified and eliminated using Purge_Haplotigs

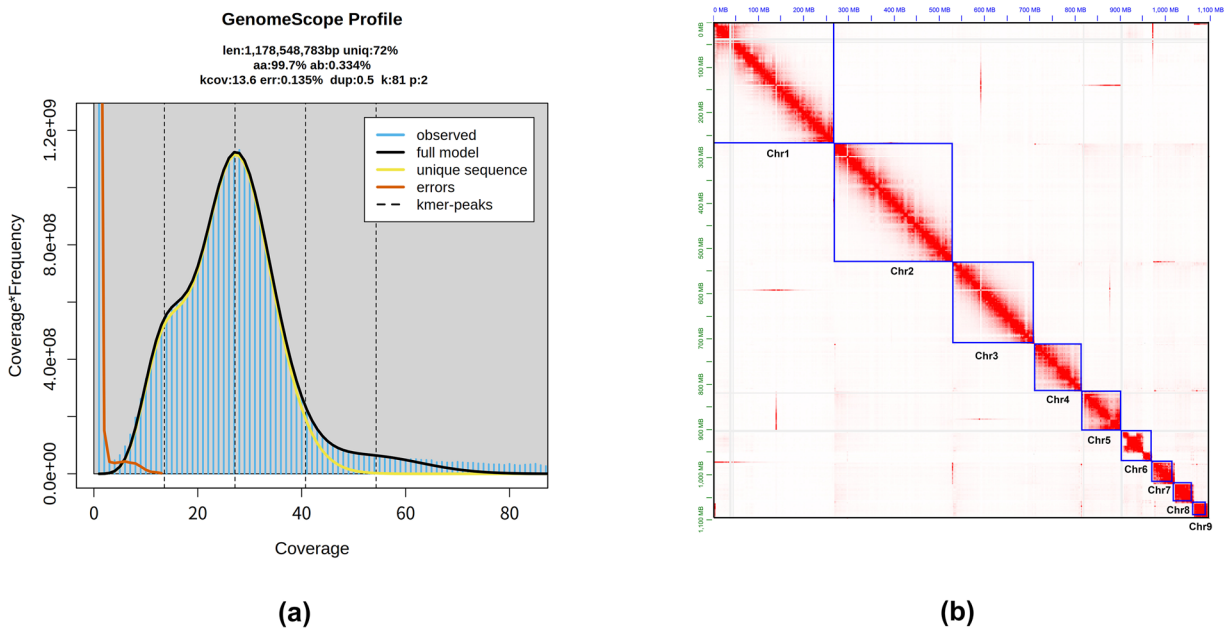

Fig. 1 (**a**): GenomeScope k-mer plot of *S. jimo*. (**b**): Hi-C contact heatmap of the *S. jimo* genome.

Genome assembly	Contigs	Scaffolds
Number of sequences	373	58
Total length (bp)	1,192,309,203	1,192,340,703
N50 (bp)	42,669,167	198,773,512
GC content	40.18%	40.18%
BUSCO (insecta_odb10)	/	98.00%
Ambiguous bases (N's)	/	31,500

Table 2. Assembly statistics of *S. jimo*.

Name	Length (bp)
Chromosome 1	269,344,292
Chromosome 2	270,192,479
Chromosome 3	198,773,512
Chromosome 4	115,349,299
Chromosome 5	90,170,527
Chromosome 6	60,955,701
Chromosome 7	51,884,644
Chromosome 8	34,170,425
Chromosome 9	37,802,904
Unplaced	63,696,920

Table 3. Summary of chromosome length of *S. jimo* genome.

v1.1.2[19]. The total size of the draft genome assembly was 1.192 Gb with an N50 length of 42.67 Mb. The overall GC content of the *S. jimo* genome was estimated to 40.18% (Table 2).

We used SAMtools v1.6[20] to index the draft genome and used BWA v0.7.17[21] to align the Hi-C data with the draft genome. Then, the contigs were anchored into chromosomes by Hi-C sequencing reads through the Juicer v2.0[22] and YAHS v1.2a.1[23] software workflows. To further improve the chromosome-scale assembly, it was subjected to manual review and refinement using Juicebox v2.17.00[22] Assembly Tools. Finally, we used the software RACON v1.4.13-cb13104[24] for error correction. The integrity of the genome assembly was also assessed using BUSCO v5.8.2[25] based on the insecta_odb10 datasets.

Ultimately, Hi-C reads were employed to link these contigs to 58 scaffolds (Table 2), resulting in a total of 9 pseudochromosomes (Fig. 1b, Table 3), with a mounting rate of up to 94.65%, and the chromosome lengths ranged from 34.17 Mb to 270.19 Mb (Table 3). The chromosomal-level assembly constituted a total length of

Name	Length(bp)	Percentage
DNA transposon elements	24,120,857	2.02%
LINEs	327,768,136	27.49%
LTRs	608,996	0.05%
Rolling Circle	220,453	0.02%
SINEs	1,623,538	0.14%
Penelope	733,725	0.06%
Other (Simple Repeat, Microsatellite, RNA)	47,699,161	4.00%
Unclassified	80,205,577	6.73%
Total	482,980,443	40.51%

Table 4. Statistics of repeat annotation result.

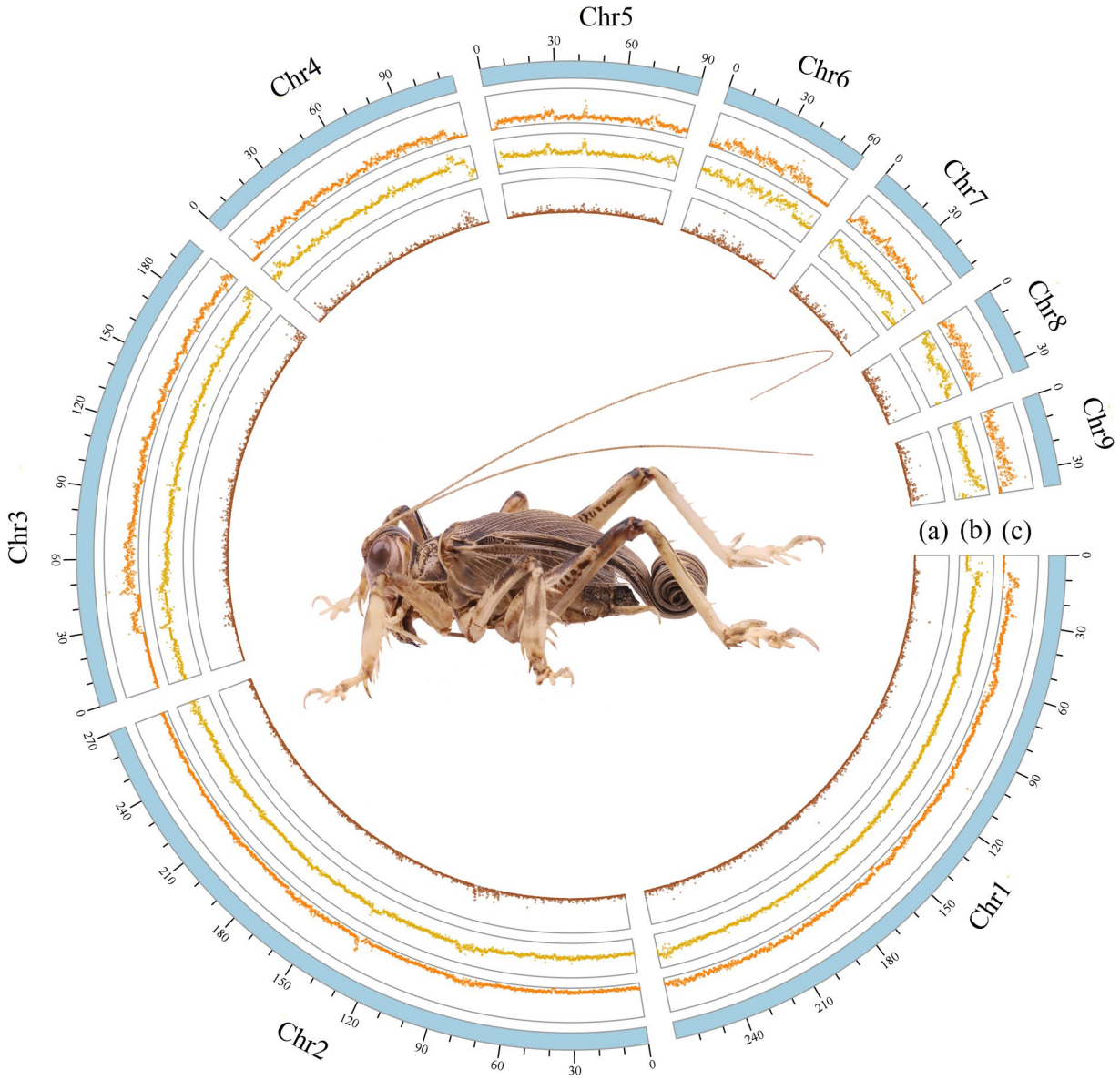

Fig. 2 Schematic representation of the genomic characteristics of *S. jimo*. (**a**): Gene density; (**b**): GC skew; (**c**): Repetitive elements.

1.178 Gb of *S. jimo* genome, with the longer scaffold N50 (198.77 Mb), the number of ambiguous bases (N's) is 31,500 and a high completeness 98.0% of BUSCO evaluation (Table 2).

Gene annotation	Value
Coding gene number	14,215
BUSCO (metazoa_odb10)	94.10%
Mean Gene length	13,780.20
Mean CDS length	1,589.2

Table 5. Statistics of genome annotation.

	Number	Percent(%)
Total	14,215	/
Nr	12,373	87.04%
InterPro	11,722	82.46%
Pfam	10,561	74.29%
GO	8,121	57.13%
Annotated	12,747	89.67%
Unannotated	1,468	10.33%

Table 6. Statistics of gene functional annotation.

Repeat annotation. We used Earl Grey v4.2.4 Pipeline[26] to detect repetitive elements (REs). A total of 482.98 Mb REs were identified (constituting 40.51% of *S. jimo* genome), including 2.02% DNA transposon elements, 27.49% long interspersed nuclear elements (LINEs), 0.05% long terminal repeats (LTRs), 0.06% Penelope, 0.02% Rolling Circle, 0.14% short interspersed nuclear elements (SINEs), 4.00% Tandem Repeats (TRs), and 6.73% unclassified (Table 4, Fig. 2).

Gene structure and functional annotation. We used EVidenceModeler (EVM) v2.0.0[27] to integrate 3 strategies, including *de novo* prediction, homologous genes, and transcriptomes to annotate protein-coding genes. For transcriptome-based annotation, RNA-seq data were firstly mapped to our assembly using HISAT2[28], and the transcriptome information in BAM alignments was produced, With our reference assembly, transcriptome data were further annotate using StringTie v2.1.438[29] with the default parameters. For homolog-based annotation, protein sequences of the Anabrus simplex[30], *Locusta migratoria*[3] and *Apteronemobius asahinai*[31] genomes were aligned to *S. jimo* genome sequences using Genewiss v2.4.1[32]. The *de novo* prediction was performed using Augustus v3.3.434[33], which was automatically trained using the mapped transcriptome data and protein homology information.

Overall, a total of 14,215 coding genes were annotated within the assembled *S. jimo* genome. The average gene length was 13,780.2 bp, with an average CDS length of per gene 1,473.5 bp. Based on BUSCO analysis, 94.1% of the BUSCO database (insecta_odb10) genes were identified, further underlining the accuracy and completeness of gene predictions (Table 5).

Protein motifs and domains were annotated using InterProScan v5.61[34] in combination with publicly available databases, including InterPro[35] and Pfam[36]. Gene Ontology (GO) terms were assigned based on the corresponding InterPro entries. Functional annotation was further conducted by aligning the protein sequences against the Nr database[37] using the DIAMOND v2.1.6[38] with an E-value threshold of $<1e{-}5$. After comparing the protein sequences predicted from gene models with the aforementioned protein databases, we successfully assigned putative functions to 12,747 protein-coding genes, accounting for 89.67% of the total gene set(Table 6).

Data Records

The assembled genome, Illumina sequencing data, genomic PacBio sequencing data, transcriptomic Illumina sequencing data, and Hi-C sequencing data are being deposited in the GenBank database under project number PRJNA1293729. The assembled genome has been deposited in Genbank under the accession number GCA_052149075.1[39]. The Illumina sequencing data of the genome, PacBio sequencing data of the genome, Illumina sequencing data of the transcriptome, and Hi-C sequencing data have been deposited in the SRA database, with the accession number SRP620269[40]. Genome annotation information of repeated sequences, gene structure and functional prediction is available in the Figshare database[41].

Technical Validation

We evaluated the quality of the *S. jimo* genome assembly and annotation using BUSCO v5.8.2. The results showed that 98.0% of the BUSCO core genes were identified. Furthermore, we employed the GAEP pipeline[42] to align Illumina short reads and long reads to the assembled genome, achieving mapping rates of 99.76% and 99.87%, respectively. In addition, the Quality Value (QV) of the genome, estimated by Merqury[43], was 42.87. Collectively, these results indicate that the *S. jimo* genome assembly exhibits high completeness, accuracy, and overall quality.

Data availability

The genome assembly results of this study have been deposited in two public databases, namely NCBI[39] (https://identifiers.org/ncbi/insdc.gca:GCA_052149075.1) and Figshare[41] (https://doi.org/10.6084/m9.figshare.29626556.v4). All raw data used in this study are available and have been submitted to GenBank (The BioProject number: PRJNA1293729). Files related to repetitive sequences and gene annotations can be accessed on Figshare[41] (https://doi.org/10.6084/m9.figshare.29626556.v4).

Code availability

No specific script was used in this work. The codes and pipelines used in data processing were all executed according to the manual and protocols of the corresponding bioinformatics software.

Received: 8 August 2025; Accepted: 29 October 2025;
Published online: 10 December 2025

References

1. Alexandre, V. L. Locusts and remote sensing: a review. *Journal of Applied Remote Sensing* **7**, 075099 (2013).
2. Cigliano, M. M., Braun, H., Eades, D. C. & Otte, D. Orthoptera Species File Online. on 2025-07-17 at http://orthoptera.speciesfile.org/.
3. Wang, X. *et al*. The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications* **5**, 2957 (2014).
4. Li, X., Mank, J. E. & Ban, L. The grasshopper genome reveals long-term gene content conservation of the X Chromosome and temporal variation in X Chromosome evolution. *Genome Research* **34**, 997–1007 (2024).
5. Zhang, X. *et al*. Temporal genomics in Hawaiian crickets reveals compensatory intragenomic coadaptation during adaptive evolution. *Nature Communications* **15**, 5001 (2024).
6. Hawlitschek, O. *et al*. New estimates of genome size in Orthoptera and their evolutionary implications. *PLoS One* **18**, e0275551 (2023).
7. Song, H. *et al*. Phylogenomic analysis sheds light on the evolutionary pathways towards acoustic communication in Orthoptera. *Nature Communications* **11**, 4939 (2020).
8. Jost, M. C. & Shaw, K. L. Phylogeny of Ensifera (Hexapoda: Orthoptera) using three ribosomal loci, with implications for the evolution of acoustic communication. *Molecular Phylogenetics and Evolution* **38**, 510–530 (2006).
9. Laure, D. G. Phylogeny and the Evolution of Acoustic Communication in Extant Ensifera (Insecta, Orthoptera). *Zoologica Scripta* **32**, 525–561 (2003).
10. Song, H. *et al*. 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. *Cladistics* **31**, 621–651 (2015).
11. Leubner, F., Bradler, S. & Wipfler, B. The thoracic morphology of the wingless dune cricket Comicus calcaris (Orthoptera: Schizodactylidae): Novel apomorphic characters for the group and adaptations to sand desert environments. *Arthropod Struct Dev* **46**, 449–461 (2017).
12. Heads, S.W. & Leuzinger, L. On the placement of the Cretaceous orthopteran Brauckmannia groeningae from Brazil, with notes on the relationships of Schizodactylidae (Orthoptera, Ensifera). *Zookeys*, 17–30 (2011).
13. He, Z. Q. & Liu, Y. J. A new species of Dune Cricket from China (Orthoptera: Ensifera: Schizodactylidae). *Zootaxa* **4999**, 356–362 (2021).
14. Parkhomchuk, D. *et al*. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research* **37**, e123 (2009).
15. Singla, D. & Yadav, I. S. GAAP: A GUI-based Genome Assembly and Annotation Package. *Current Genomics* **23**, 77–82 (2022).
16. Hesse, U. K-Mer-Based Genome Size Estimation in Theory and Practice. *Methods in Molecular Biology* **2672**, 79–113 (2023).
17. 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**, 1–10 (2020).
18. 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).
19. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
20. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
21. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics* (2013).
22. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
23. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
24. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* **27**, 737–746 (2017).
25. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology* **14**, 227–245 (2019).
26. Baril, T., Galbraith, J., Hayward, A. & Arkhipova, I. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Molecular Biology and Evolution* **41**, msae068 (2024).
27. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
28. 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).
29. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).
30. Childers, A. K. *et al*. in Insects, **Vol. 12** (2021).
31. Satoh, A., Takasu, M., Yano, K. & Terai, Y. De novo assembly and annotation of the mangrove cricket genome. *BMC Research Note* **14**, 387 (2021).
32. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* **14**, 988–995 (2004).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
34. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
35. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**, D211–215 (2009).
36. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–d419 (2021).
37. Yu, K. & Zhang, T. Construction of customized sub-databases from NCBI-nr database for rapid annotation of huge metagenomic datasets using a combined BLAST and MEGAN approach. *PLoS One* **8**, e59831 (2013).
38. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).

39. Yao, B. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052149075.1 (2025).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP620269 (2025).
41. Yao, B. Chromosome-level genome assembly and annotation of Schizodactylus jimo. *figshare* https://doi.org/10.6084/m9.figshare.29626556.v4 (2025).
42. Zhang, Y., Lu, H. W. & Ruan, J. GAEP: a comprehensive genome assembly evaluating pipeline. *Journal of Genetics and Genomics* **50**, 747–754 (2023).
43. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

Acknowledgements

This research was supported by the Fundamental Research Funds for the Central Universities (GK202105003, and TD2020041Y). This work was partly supported by the National Natural Science Foundation of China (no. 31872273).

Author contributions

B.Y., T.J. and S.X. conceived this study; B.Y., T.J. and J.W. collected the samples and performed the experiments; B.Y., T.J., D.G. and L.Y. performed the research and analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.G. or S.-Q.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.