scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An annotated chromosome-level genome of a troglophilic spider (*Pimoa clavata*)

Yiming Zhang[1,2,3], Xiaoqing Zhang[3,4], Tongyao Jiang[3], Yejie Lin[1,2,3] & Shuqiang Li[1,3] ✉

Climatic upheavals throughout Earth's history have driven species to subterranean refugia, where stable microclimates buffer environmental extremes. The spider family Pimoidae, relict lineages sensitive to thermal fluctuations, exemplifies this climate-driven habitat transition. Here, we present the first chromosome-level genome of *Pimoa clavata*, a troglophilic spider endemic to Beijing's mountainous caves. Combining PacBio HiFi and Hi-C sequencing, we assembled a 1.94 Gb genome with 29 scaffolds in chromosomes level, which were further subjected to comprehensive annotation. The well-annotated genome of this species provides foundational insights into the genomic basis of subterranean adaptation, offering critical resources for evolutionary genomics and conservation of climate-vulnerable cave species.

Background & Summary

Earth's geological history has been punctuated by dramatic climatic oscillations, from glacial-interglacial cycles to hyperthermal events, each acting as a crucible for evolutionary innovation. Over deep time, species survival during abrupt environmental shifts has hinged on two primary strategies: adaptive genetic remodeling or habitat migration to refuge zones[1]. Among terrestrial ecosystems, subterranean environments—particularly caves—have served as critical biodiversity arks during climatic upheavals. These systems buffer extreme temperature fluctuations and maintain stable humidity levels, creating microclimatic havens that decouple from surface environmental stressors[2]. The persistence of climate change can drive some species to abandon their thermally unstable surface habitats and instead seek underground shelters. This habitat transition often initiates speciation cascades: geographic isolation in fragmented cave networks promotes morphological reduction (e.g., eye and pigment loss), metabolic reprogramming for nutrient scarcity, and sensory specialization for perpetual darkness[3,4]. Over evolutionary timescales, these adaptations crystallize into distinct ecological guilds—troglophiles (facultative cave-dwellers) and troglobites (obligate subterranean specialists)—whose genomes encode extraordinary solutions to extreme environmental challenges[5,6].

The spider family Pimoidae (Araneae: Araneomorphae) represents a historically phylogenetically enigmatic lineage that epitomizes this climate-driven transition. With 86 described species across two genera (World Spider Catalog v25.5), pimoids are relict taxa exhibiting pronounced thermal sensitivity and relictual distribution patterns, persisting primarily in cool, stable microhabitats of mountainous caves and forest litter[7]. Preliminary phylogeographic analyses suggest that Pimoidae diverged during the mid-Miocene climatic transition (42–33 millions of years ago, MYA), when global cooling and orogenic uplift fragmented their ancestral habitats[8]. This evolutionary trajectory positions them as living archives of paleoclimatic adaptation—a status further underscored by their extreme vulnerability to contemporary warming trends. Some species in this group, as troglophile species, occupy an ecological nexus between surface and subterranean biomes, making them ideal models for studying genomic signatures of habitat transition under climatic pressure. However, the absence of high-quality genomic resources has hindered investigations into the mechanistic links between their relict status, thermal sensitivity, and cave adaptation.

Here, we present the first chromosome-level genome assembly of *Pimoa clavata*, a troglophilic spider endemic to natural caves in Beijing's mountainous regions. This species embodies a unique confluence of

[1]College of Life Sciences, Anhui Normal University, Wuhu, 241000, China. [2]Hebei Key Laboratory of Animal Diversity, College of Life Sciences, Langfang Normal University, Langfang, 065000, China. [3]Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Chinese Academy of Sciences, Beijing, 100101, China. [4]College of Life Science, Shenyang Normal University, Shenyang, 110034, China. ✉e-mail: lisq@ioz.ac.cn

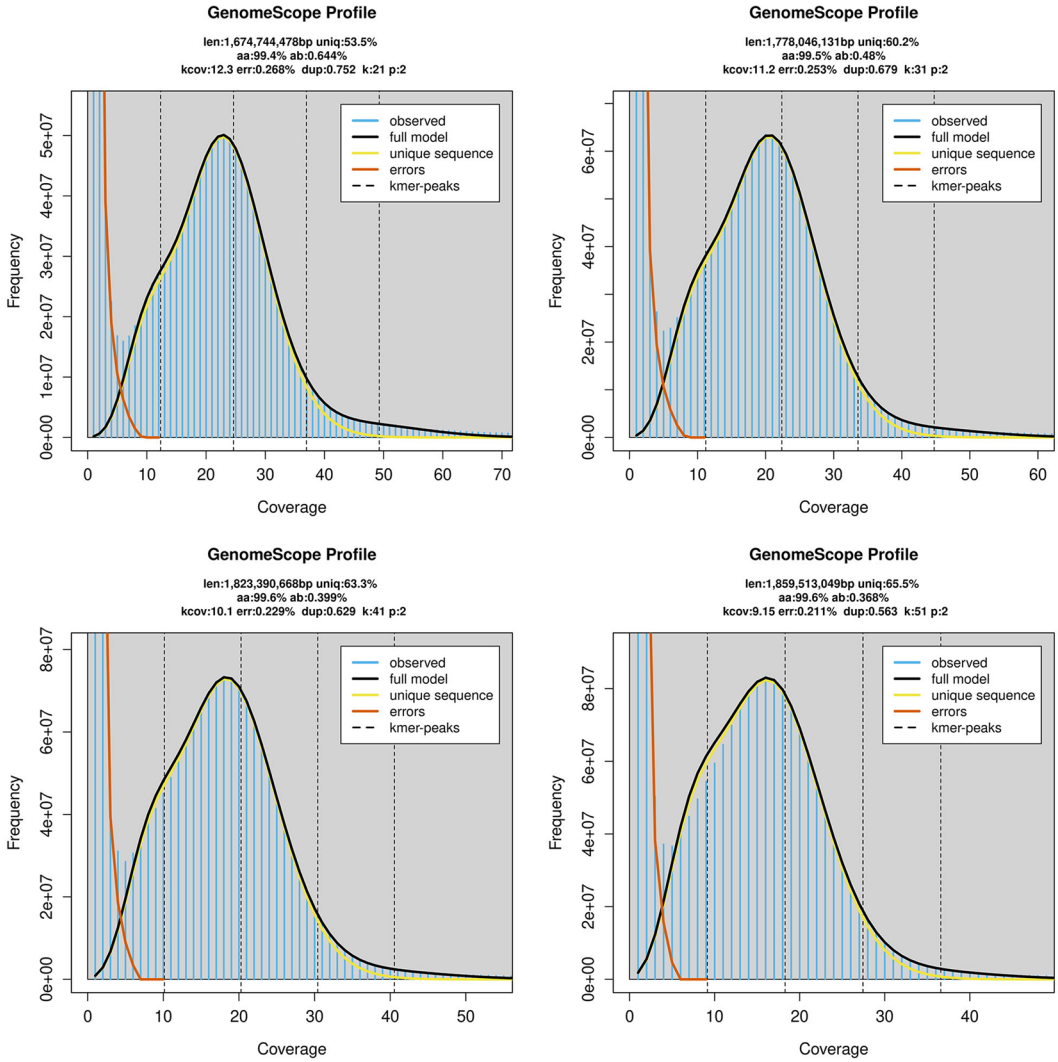

Fig. 1 Genome size estimation of *Pimoa clavata* using k-mer analysis with different k values. K-mer frequency distributions of Illumina short reads were generated using GenomeScope v2.0 with varying k values (k = 21, 31, 41, 51).

evolutionary singularities: as a climate-vulnerable relict, a cave-adapted troglophile, and a lineage occupying the basal branches of the Asian clade system of the genus *Pimoa*. Based on k-mer analysis of Illumina short-read data, the genome size of *Pimoa clavata* was estimated to be between 1.67 Gb and 1.85 Gb (Fig. 1). Using an integrated sequencing approach—PacBio HiFi long reads (33.0 Gb) and Hi-C chromatin conformation data (210.4 Gb)—we assembled a 1.94 Gb genome, with 95.8% (1.87 Gb) anchored into 29 chromosomal scaffolds (Fig. 2a,c), slightly larger than the predicted size. Assembly metrics confirm exceptional completeness and continuity, positioning this genome among the highest-quality spider genomes currently available. Repetitive elements dominate the genome (66.45%), with DNA transposons (14.87%) representing the most abundant transposable component, consistent with patterns observed in other spider lineages. However, LTR retrotransposons (14.56%) are disproportionately enriched compared to other arachnid genomes[9–13]—a distinctive feature potentially linked to transposon-mediated regulatory evolution in isolated cave populations (Fig. 2d, Table 1).

Gene annotation pipelines identified 16,829 protein-coding genes, with 95.2% precisely mapped to chromosomes. Functional characterization revealed robust completeness and broad annotation coverage: 97.1% (16,341) of genes received functional assignments via Nr (16,295; 96.8%) or eggNOG (14,968; 88.9%) databases.

Methods

Sample collection and sequencing. Live *P. clavata* were field-captured from Beijing suburban mountainous regions (GPS: 39°40′10.92″N, 115°49′34.68″E; altitude: 592 m) under strict ethical protocols. Given the species' obligate subterranean ecology and extreme sensitivity to environmental perturbation—individuals succumb to mortality within 4–6 hours post-cave egress—collected specimens were immediately stabilized in cryogenic transport modules (pre-cooled to 0 °C) and flash-frozen in liquid nitrogen within 2 hours of capture. To mitigate sampling bias and preserve population integrity, pooled sequencing libraries were constructed from a

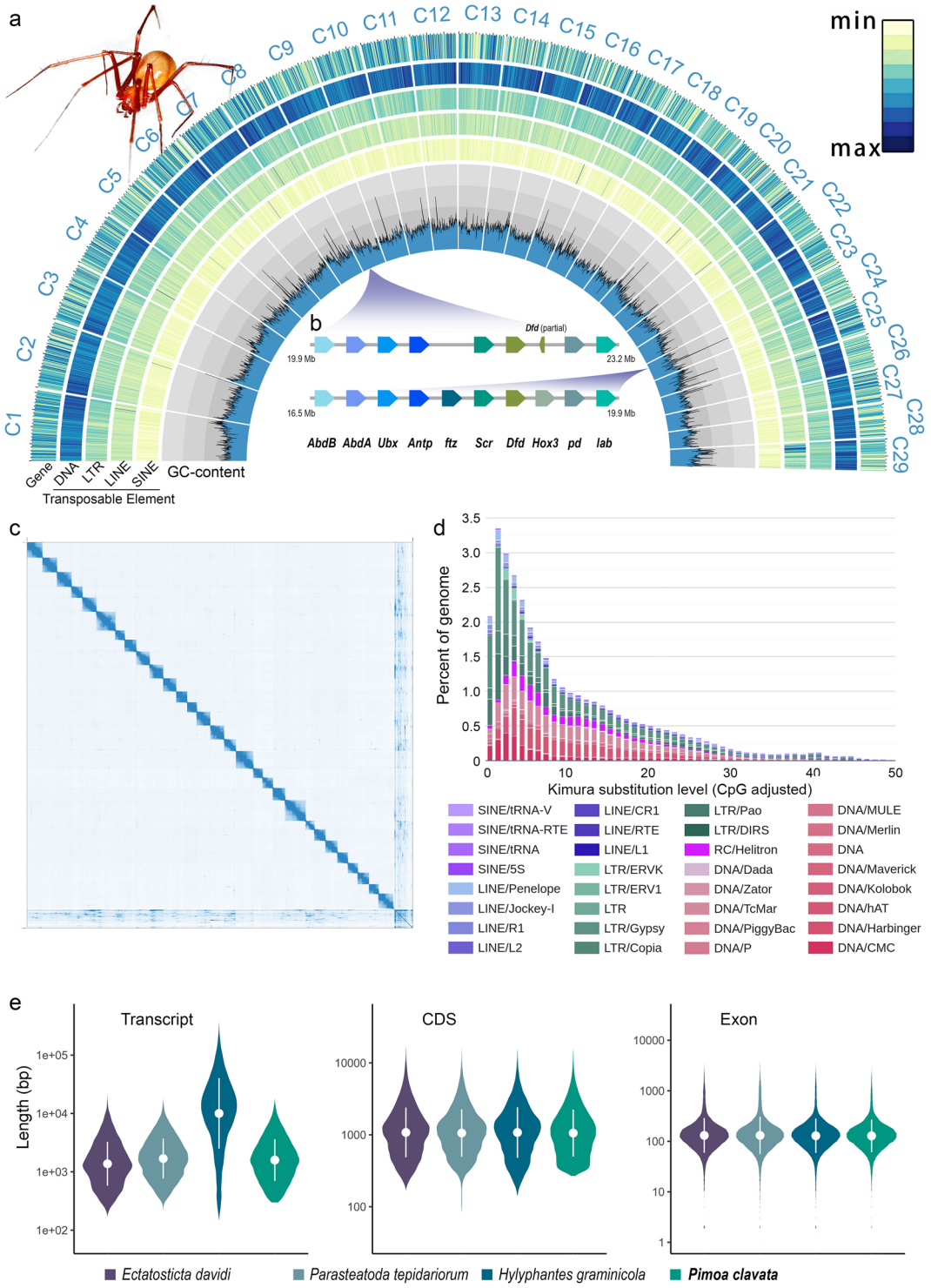

Fig. 2 Characterization of *Pimoa clavata* genome. (**a**) Circos plot of the genomic characteristics. A female *P. clavata* is shown in the upper left corner (Photographer: Yejie Lin). From the outer ring to the inner ring are the distributions of gene density, Transposable Elements (DNA, LTR, LINE, SINE), and GC content. (**b**) Hox genes in *P. clavata* genome and their positions on chromosomes. (**c**) Hi-C interaction matrix demonstrating chromosome-scale assembly continuity. (**d**) Repeat element landscape showing evolutionary dynamics of interspersed repeats. (**e**) Comparative genomic feature analysis of transcripts, coding sequence (CDS) and exon length distributions across *P. clavata* and three reference species.

non-sex-discriminated cohort (n = 25 adults). High-molecular-weight genomic DNA was extracted from homogenized whole-body tissue using the QIAGEN Blood & Cell Culture DNA Kit.

file name: Pimo.genome.fa					
sequences: 639					
total length: 1948072705 bp (1946714705 bp excl N/X-runs)					
GC level: 31.73%					
bases masked: 1294405075 bp (66.45%)					
			number of elements*	length occupied (bp)	percentage of sequence
Retroelements			727354	344805057	17.70%
	SINEs:		48222	5912525	0.30%
		Penelope	41973	19610960	1.01%
	LINEs:		128839	55348449	2.84%
		CRE/SLACS	0	0	0.00%
		L2/CR1/Rex	10018	3530785	0.18%
		R1/LOA/Jockey	45206	20056402	1.03%
		R2/R4/NeSL	0	0	0.00%
		RTE/Bov-B	18917	5862569	0.30%
		L1/CIN4	3870	1320844	0.07%
	LTR elements:	550293	283544083	14.56%	
		BEL/Pao	111745	61811316	3.17%
		Ty1/Copia	56194	31277211	1.61%
		Gypsy/DIRS1	233802	133110933	6.83%
		Retroviral	24673	13547760	0.70%
DNA transposons			886279	289637036	14.87%
	hobo-Activator	310361	98245820	5.04%	
	Tc1-IS630-Pogo	279044	87680886	4.50%	
	En-Spm		0	0	0.00%
	MuDR-IS905	0	0	0.00%	
	PiggyBac		10315	3091148	0.16%
	Tourist/Harbinger	2144	296460	0.02%	
	Other(Mirage, P-element, Transib)	9009	5479667	0.28%	
Rolling-circles			140456	48220169	2.48%
Unclassified:			2113332	566445090	29.08%
Total interspersed repeats:				1200887183	61.64%
Small RNA:			38902	6242396	0.32%
Satellites:			1618	1246723	0.06%
Simple repeats:			400866	39771423	2.04%
Low complexity:			39685	1962795	0.10%

Table 1. Repeat annotation in the *Pimoa clavata* genome. *most repeats fragmented by insertions or deletions have been counted as one element. RepeatMasker version 4.1.2-p1, default mode. run with rmblastn version 2.11.0+.

For long-read sequencing, SMRTbell libraries—double-stranded DNA templates capped with hairpin loops at both ends—were constructed following the standard PacBio protocol using a 15 kb preparation solution (PacBio). The high-fidelity (HiFi) libraries were subsequently sequenced with the PacBio Sequel II system in Circular Consensus Sequencing mode at Novogene Technology, yielding a total of 33.0 Gb of HiFi data, comprising 2,054,123 reads[14].

Chromosome conformation capture (Hi-C) experiments were performed to elucidate the three-dimensional organization of the genome[15]. Hi-C libraries were prepared following a published protocol with minor modifications[16]. Briefly, samples were fixed in 1% formaldehyde to induce cross-linking, after which the cross-linked chromatin was digested with the MboI restriction enzyme. Biotin-14-dCTP was then incorporated to label the DNA ends, facilitating the removal of non-ligated fragments. The ligated DNA was subsequently extracted using the QIAamp DNA Mini Kit (Qiagen). Following extraction, the purified DNA was sheared into ~350 bp fragments and processed using a standard Illumina library preparation protocol[17]. Hi-C sequencing was conducted on the Illumina Hiseq platform with paired-ends (PE) 150 bp. We then filtered the raw reads using Juicer v1.6.2[18] to remove low-quality reads and adapters, yielding 210.4 Gb of clean data for *P. clavata*.

***De novo* genome assembly.** Short-insert libraries of *P. clavata* were sequenced with the Illumina NovaSeq platform using 150 bp PE reads. To remove low-quality reads and adapter sequences, raw reads were trimmed by Trimmomatic v0.39[19]. A total of 46.3 Gb of clean data were obtained for subsequent genome survey analysis. K-mer frequency analysis was performed using Jellyfish v2.3.0[20] with multiple k-mer sizes (k = 21, 31, 41, and 51). The distribution of k-mers was analyzed using GenomeScope v2.0[21,22] to estimate genome characteristics.

	hifiasm	wtdbg2	final assembly (from hifiasm)	gene set (protein)
BUSCO				
Complete	2886	2863	2873	2829
Complete and single-copy	2713	2704	2704	2646
Complete and duplicated	173	159	169	183
Fragmented	11	16	20	25
Missing	37	55	41	80
n = 2934				
Assembly (quast)				
# contigs (>= 0 bp)	1607	9788	639	
# contigs (>= 1000 bp)	1607	9788	639	
# contigs (>= 5000 bp)	1607	9621	628	
# contigs (>= 10000 bp)	1607	8105	622	
# contigs (>= 25000 bp)	1530	5237	585	
# contigs (>= 50000 bp)	1153	4177	305	
Total length (>= 0 bp)	2176822041	1876077947	1948072705	
Total length (>= 1000 bp)	2176822041	1876077947	1948072705	
Total length (>= 5000 bp)	2176822041	1875375566	1948050653	
Total length (>= 10000 bp)	2176822041	1864182935	1948006344	
Total length (>= 25000 bp)	2175209366	1818037405	1947340237	
Total length (>= 50000 bp)	2161777910	1781163686	1939400095	
# contigs	1607	9788	639	
Largest contig	23962897	3890033	104109586	
Total length	2176822041	1876077947	1948072705	
GC (%)	32.26	31.59	31.73	
N50	4126853	713765	66374000	
N75	2204322	336541	53370419	
L50	147	773	13	
L75	323	1724	21	
# N's per 100 kbp	0	0	69.71	

Table 2. Evaluation results of genome assembly.

Across different k-mer values, the estimated genome size of *P. clavata* ranged from approximately 1.67 Gb to 1.85 Gb (Fig. 1).

PacBio reads were assembled *de novo* using two independent assemblers: hifiasm v0.19.7[23] and wtdbg2 v2.5[24]. The resulting contig sets were then polished using Racon v1.4.17 (https://github.com/isovic/racon) and NextPolish v1.4.0[25], and the optimal assembly was selected for subsequent analyses (Table 2). Hi-C sequencing data were mapped to the scaffold-level assembly of *P. clavata* using Juicer, and the 3D-DNA v180922 pipeline[26] was employed to construct chromosomal scaffolds and correct misassemblies. Final refinements were performed using Juicebox Assembly Tools v1.11.08[27].

Genome annotation. In this study, repetitive sequences were annotated using the RepeatModeler v2.0.2[28] pipeline. Repetitive elements were predicted and subsequently masked by screening against a reference repeat database with RepeatMasker v4.1.2-p1[29]. Overall, repeats spanned 1.3 Gb, accounting for 66.45% of the genome (Table 1). The most prevalent repeat types included DNA elements (14.87%), LTR elements (14.56%), LINEs (2.84%), rolling-circle transposons (2.48%), and simple repeats (2.04%). Protein-coding gene annotation was performed based on transcriptome data using the TransDecoder pipeline (v5.5.0; https://github.com/TransDecoder). Given conservation constraints for this vulnerable species, transcriptomic data were strategically acquired from the SRA database, including one *P. clavata* dataset and phylogenetically related species' resources (SRR6998915, SRR6425931, SRR3144083)[30]. Initially, RNA-seq reads were aligned to the genome assembly using STAR v2.7.1[31]. The aligned reads were then assembled into transcripts with StringTie v2.2.1[32], using the genome assembly as a reference. Finally, coding sequences (CDS) were identified from the assembled transcripts via TransDecoder, resulting in the final gene annotations. Function annotation is based on NCBI-Nr, and EggNOG v5.0 databases[33] (Fig. 2a,b,e).

Data Records

The raw sequencing data—including HiFi long reads, Hi-C, and resequencing short reads—have been submitted to the NCBI Sequence Read Archive (SRA, SRR34027933–SRR34027937 of SRP592911 in BioProject PRJNA1277451)[34]. The genome assembly of *P. clavata* is available at NCBI Genome under accession number GCA_051201355.1[35] and ScienceDB (sciencedb.21439)[36]. In addition to the raw reads and assembled genome, the structural and functional annotation results, including predicted genes, protein sequences, and associated functional terms, are publicly available only via ScienceDB[36].

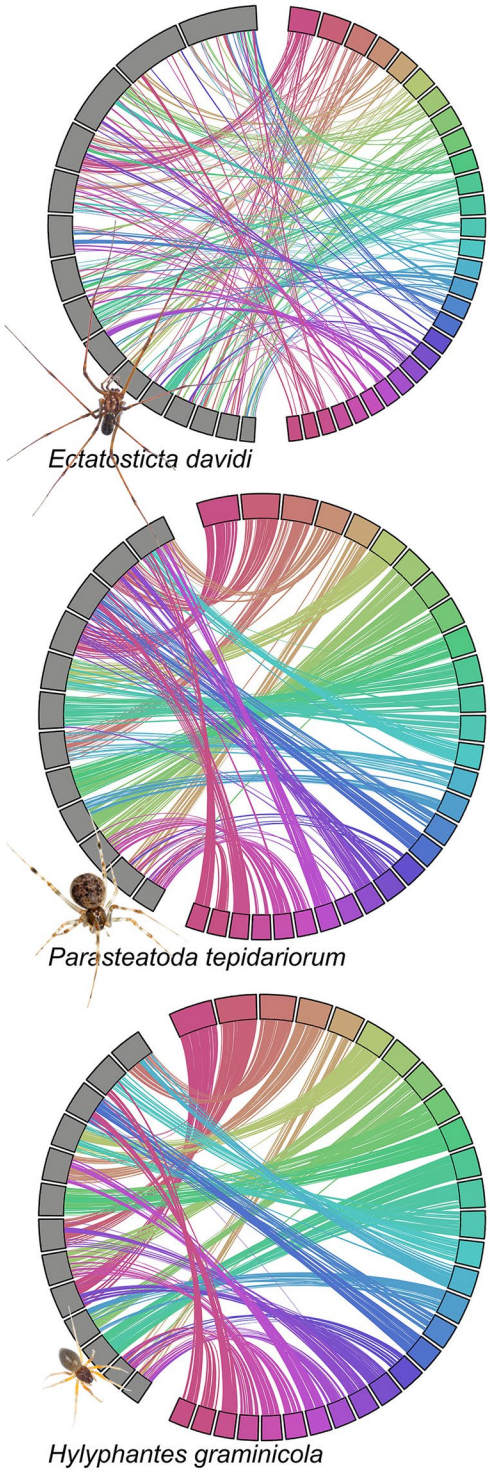

Fig. 3 Chromosomal collinearity between *Pimoa clavata* (colored) and reference species (gray). Representative images of reference species are shown in the lower left corner of each circular plot (Photographer: Tongyao Jiang).

Technical Validation

The completeness and accuracy of the *P. clavata* genome assembly and gene annotations were rigorously evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.2[37] with the arachnida_odb10 database (2,934 conserved arthropod orthologs). Dual-mode analysis was performed: (1) Genome mode assessed assembly continuity through sequence mapping and metaeuk-based gene prediction; while (2) Protein mode evaluated gene set completeness by aligning predicted proteins to orthologous sequences. Results demonstrated high genomic completeness, with 98.0% of BUSCO orthologs successfully recovered. Detailed gene-wise analysis revealed 2,829 complete protein-coding genes (96.4%), including 183 duplicated hits (6.2%). Notably,

only 25 genes (0.9%) were classified as fragmented, and 80 orthologs (2.7%) were missing entirely, indicating robust assembly and annotation quality (Table 2). Subsequently, we extracted gene structure-related metrics from the *P. clavata* genome and conducted a comparative analysis with other spider species (*Ectatosticta davidi*, *Parasteatoda tepidariorum*, and *Hylyphantes graminicola*). The results showed that the annotated genes in *P. clavata* exhibit structural characteristics comparable to those observed in other spider lineages (Fig. 2e).

To further validate the chromosomal-level assembly quality of *P. clavata*, we conducted whole-genome synteny analysis with three phylogenetically diverse spider species (*E. davidi*, *P. tepidariorum*, and *H. graminicola*) using the MCScanX pipeline implemented in the JCVI toolkit[38], employing default parameters. Macro-syntenic relationships were visualized using Circos v0.69-9[39]. Comparative analysis revealed conserved collinear patterns across all examined species, confirming the reliability and robustness of our genome assembly (Fig. 3).

Code availability

All bioinformatics analyses were executed in strict compliance with developer-provided documentation and standard operational protocols. The study exclusively employed officially distributed software modules without implementation of custom algorithms or script modifications, ensuring complete methodological transparency and replicability.

Received: 26 March 2025; Accepted: 25 July 2025;
Published online: 11 August 2025

References

1. Hoffmann, A. & Sgrò, C. Climate change and evolutionary adaptation. *Nature* **470**, 479–485, https://doi.org/10.1038/nature09670 (2011).
2. Bryson, R. W., Prendini, L., Savary, W. E. & Pearman, P. B. Caves as microrefugia: Pleistocene phylogeography of the troglophilic North American scorpion *Pseudouroctonus reddelli*. *BMC Ecology and Evolution* **14**, 9, https://doi.org/10.1186/1471-2148-14-9 (2014).
3. Jeffery, W. R. Cavefish as a Model System in Evolutionary Developmental Biology. *Developmental Biology* **231**, 1–12, https://doi.org/10.1006/dbio.2000.0121 (2001).
4. Romero, A. *Cave Biology: Life in Darkness*. https://doi.org/10.1017/CBO9780511596841 (Cambridge Univ. Press, 2009).
5. Ran, J. & Yang, W. A Review of Progress in Chinese Troglofauna Research. *Journal of Resources and Ecology* **6**, 237–246, https://doi.org/10.5814/j.issn.1674-764x.2015.04.007 (2015).
6. Mao, T. *et al*. Evolution in *Sinocyclocheilus* cavefish is marked by rate shifts, reversals, and origin of novel traits. *BMC Ecology and Evolution* **21**, 45, https://doi.org/10.1186/s12862-021-01776-y (2021).
7. Mammola, S., Hormiga, G., Arnedo, M. A. & Isaia, M. Unexpected diversity in the relictual European spiders of the genus *Pimoa* (Araneae: Pimoidae). *Invertebrate Systematics* **30**, 566–587, https://doi.org/10.1071/IS16017 (2016).
8. Hormiga, G., Kulkarni, S., DA Silva Moreira, T. & Dimitrov, D. Molecular phylogeny of pimoid spiders and the limits of Linyphiidae, with a reassessment of male palpal homologies (Araneae, Pimoidae). *Zootaxa* **1**, 71–101. Published 2021 Aug 24. https://doi.org/10.11646/zootaxa.5026.1.3 (2021).
9. Zhang, Y. *et al*. A trade-off in evolution: the adaptive landscape of spiders without venom glands. *GigaScience* **13**, e48, https://doi.org/10.1093/gigascience/giae048 (2024).
10. Hu, W. *et al*. A molecular atlas reveals the tri-sectional spinning mechanism of spider dragline silk. *Nature Communications* **14**, 837, https://doi.org/10.1038/s41467-023-36545-6 (2023).
11. Fan, Z. *et al*. A chromosome-level genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. *GigaScience* **10**, b16, https://doi.org/10.1093/gigascience/giab016 (2021).
12. Liu, S., Aagaard, A., Bechsgaard, J. & Bilde, T. DNA Methylation Patterns in the Social Spider, *Stegodyphus dumicola*. *Genes* **10**, 137, https://doi.org/10.3390/genes10020137 (2019).
13. Yu, N. *et al*. Chromosome-level genome of spider *Pardosa pseudoannulata* and cuticle protein genes in environmental stresses. *Scientific Data* **11**, 121, https://doi.org/10.1038/s41597-024-02966-1 (2024).
14. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
15. Lu, L. *et al*. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Molecular Cell* **79**, 521–534, https://doi.org/10.1016/j.molcel.2020.06.007 (2020).
16. Rao, S. S. P. *et al*. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680, https://doi.org/10.1016/j.cell.2014.11.021 (2014).
17. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **2010**, t5448, https://doi.org/10.1101/pdb.prot5448 (2010).
18. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
21. Vurture, G. W. *et al*. M. C. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
22. 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, https://doi.org/10.1038/s41467-020-14998-3 (2020).
23. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
24. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* **17**, 155–158, https://doi.org/10.1038/s41592-019-0669-3 (2020).
25. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
26. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
27. Dudchenko, O. *et al*. The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at https://www.biorxiv.org/content/10.1101/254797v1 (2018).

28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in Bioinformatics* **4**, 4–10, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
30. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP139794 (2018).
31. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
32. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
33. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
34. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/SRP592911 (2025).
35. *NCBI Genome*, https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_051201355.1 (2025).
36. Zhang, Y. & Li, S. Genome sequencing and assembly of *Pimoa clavata*. V4. *Science Data Bank* https://doi.org/10.57760/sciencedb.21439 (2025).
37. Seppey, M., Manni, M. & Zdobnov, E. M. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 227–245 (Springer New York, 2019).
38. Tang, H. *et al*. Synteny and Collinearity in Plant Genomes. *Science* **320**, 486–488, https://doi.org/10.1126/science.1153917 (2008).
39. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645, https://doi.org/10.1101/gr.092759.109 (2009).

Acknowledgements

This research was supported by the Science & Technology Fundamental Resources Investigation Program of China (2023FY100200) and the National Natural Science Foundation of China (NSFC 32430015).

Author contributions

Y.Z. and S.L. designed the project. X.Z., T.J. and Y.L. were involved in sample collection and preparation. Y.Z. contributed to the bioinformatics analysis and drafted the manuscript. All authors read, revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L.

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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2025