scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale genome and lncRNA annotation of the Orange Baboon Tarantula (*Pterinochilus murinus*)

Yiming Zhang[1,2], Fengyuan Li[3], Pengyu Jin[4], Tongyao Jiang[3], Haifeng Chen[1,2], Jiangni Li[1,2] ✉ & Xunyou Yan[1] ✉

We present a high-quality chromosome-scale genome assembly for the Orange Baboon Tarantula (*Pterinochilus murinus*). The 4.7 Gb genome is organized into 22 chromosomes and demonstrates 96.32% completeness based on BUSCO analysis (-m genome) using the arachnida_odb10 dataset. Genome annotation supported by extensive RNA-seq evidence (~189 Gb) identified 40,403 protein-coding genes and 14,286 long non-coding RNA loci. These comprehensive genomic resources—including a chromosome-scale assembly and lncRNA annotations—support a wide range of comparative and functional genomic analyses in spiders and related arthropods.

Background & Summary

Tarantulas (family Theraphosidae) are a large and distinctive branch of spiders, including the world's largest and exceptionally long-lived spider species. For example, the Goliath birdeater (*Theraphosa blondi*) attains a body mass of ~175 g and lengths up to 13 cm, and members of Theraphosidae can live for >30 years in captivity[1–3]. These ground-dwelling predators occur in a wide range of habitats – spiders in general "occupy habitats from the most arid deserts to the extreme Arctic"[4,5]– but many studied Tarantulas (e.g. Brazilian or rainforest species) inhabit humid environments.

The Orange Baboon Tarantula (*Pterinochilus murinus*, OBT) is a highly recognizable theraphosid native to the dry savanna regions of East and southern Africa (Democratic Republic of Congo, Burundi, Kenya, Tanzania, Angola, Zambia, Mozambique, Zimbabwe and South Africa)[4]. It is famed among pet-keepers as the "Orange Bitey Thing" for its aggressive temperament, but is also notable for its bright orange coloration. In addition, the OBT exhibits five widely recognized color variants—Brown Colour Form (BCF; Tete, Mozambique), Dark Colour Form (DCF; Botswana, Zimbabwe, Kenya, Kigoma, Mikumi), Orange Colour Form (OCF; Usambara Mountains Region; formerly UMW), Red Colour Form (RCF; Usambara Mountains Region), and Typical Colour Form (TCF; Kenya, Mozambique). These morphs, frequently described in hobbyist literature and online reference sources (e.g., Wikipedia, Keith's Arachnids), are tentatively associated with specific geographic locales or soil types. Morphological distinctions among these forms are primarily limited to differences in body pigmentation and setal patterns, particularly in the coloration of the carapace, abdomen, and leg rings, as well as the brightness and contrast of the starburst (radiating) carapace pattern and the fishbone abdominal design[6]. To date, however, no peer-reviewed study has systematically documented these morphs or investigated their ecological and genetic correlates. Investigation of the OBT not only mitigates the bias in current Tarantula genomic research toward humid-environment taxa (e.g., *Acanthoscurria geniculate*[5] and *Chaetopelma lymberakisi*[7]), but also yields critical insights into the genomic underpinnings of adaptation to diverse arid habitats.

Despite the diversity of Theraphosidae (hundreds of species)[4,8], genomic resources for Mygalomorphae spiders remain scarce. To date, only a handful of spider genomes have been published, most of which belong to

[1]Hebei Key Laboratory of Animal Diversity, College of Life Sciences, Langfang Normal University, Langfang, 065000, China. [2]Langfang Key Laboratory of Bioinformatics, 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 Sciences, Anhui Normal University, Wuhu, 241000, China. ✉e-mail: lijiangni@lfnu.edu.cn; yanxunyou@lfnu.edu.cn

Araneomorphae (e.g., orb-weavers and cobweb spiders)[9–14]. In contrast, genomic resources for Mygalomorphae and Mesothelae—representing the entire suborder Opisthothelae—are extremely limited. The few existing genome assemblies for Mygalomorphae (e.g., trapdoor spiders) and Mesothelae (e.g., segmented trapdoor spiders) either have poor contiguity or lack publicly available, complete annotations[10,15,16]. In particular, *A. geniculata* (a rainforest Tarantula) was the first mygalomorph genome sequenced, but it is highly fragmented and lacks a reference-quality annotation[5]. Until now, no Tarantula genome has been available with both chromosome-scale assembly and comprehensive annotation. Here we address this gap by presenting the first chromosome-level, annotated genome for a Theraphosidae spider. We sequenced and assembled the *P. murinus* genome from an adult female specimen (sourced from a Chinese pet trade) using 152 Gb of PacBio HiFi long-read data (~34 × coverage of the estimated ~4.5 Gb genome) and 272 Gb of Hi-C chromatin interaction data. This yielded a highly contiguous assembly into 22 chromosome-scale scaffolds (Fig. 1a; Figure S1, see Supplementary Information), with nearly the entire genome (99.60%) represented in these large scaffolds. The total assembly size is 4.7 Gb, which falls within the size range of previously reported and *k-mer* based estimates for theraphosid genomes[17], and the presence of 22 chromosomes is consistent with earlier karyotypic analyses of this species[18]. Repeat annotation revealed that the *P. murinus* genome contains a high proportion of repetitive elements, accounting for 75.8% of the assembly. The most abundant category was DNA transposons (34.9%), a feature consistent with the repeat landscape reported for most previously characterized spider genomes. Long terminal repeat (LTR) retrotransposons represented the second most prevalent class (12.7%), while unclassified repeats accounted for 13.1% (Fig. 1b,d; Table 1).

Automated genome annotation was conducted by integrating multiple evidence types—including RNA-seq transcriptome data (~189 Gb), homology-based protein reference annotation, and *ab initio* gene prediction (see Methods). The combined annotation initially yielded 114,219 putative protein-coding genes. To minimize potential false positives, we retained only the transcriptome-supported gene models, resulting in 40,403 relatively high-confidence protein-coding genes and 14,286 long non-coding RNA (lncRNA) genes.

Assessment of the transcript-supported gene models using BUSCO (-m prot) revealed 95.0% completeness for the full transcriptome and 93.9% for the longest isoform set (Table 2), indicating high overall completeness. Among the protein-coding genes, 32,942 received functional annotations, with 30,540 genes matched to functions in the EggNOG-mapper database and 30,766 showing significant similarity to entries in the NCBI non-redundant (nr) protein database. Additionally, 24,361 genes were assigned clear Gene Ontology (GO) or KEGG functional annotations, including 22,151 genes with GO terms and 21,505 with KEGG pathway annotations.

The new *P. murinus* genome fills a critical gap in arachnid genomics. It is the first high-quality, chromosome-scale and annotated genome of a Theraphosidae/Mygalomorphae spider, and will be a valuable reference for studies of Tarantula biology. This dataset enables comparative genomic analyses between Mygalomorph and Araneomorph spiders, for example to investigate the genetic basis of their differing life histories and ecological niches. It also provides a resource for exploring genes related to the OBT's striking coloration, venom, and adaptations to dry environments. In summary, our assembly meets the standards of a reference-grade genome and provides the research community with a richly annotated dataset for this iconic Tarantula species.

Methods

Sample collection and sequencing. All sequencing data were generated from a single adult female *P. murinus* (RCF) obtained from a commercial supplier in the mainland Chinese exotic pet market (https://m.tb.cn/h.hr9InvjZ1Sb7uWH). The specimen was euthanized following standard ethical handling protocols for invertebrates. The fourth pair of legs was removed and processed separately for high-fidelity (HiFi) long-read sequencing and Hi-C chromatin conformation capture sequencing. The remaining body tissue was homogenized and used for RNA extraction to generate transcriptomic data.

High-molecular-weight genomic DNA was extracted from leg tissue by Novogene Co., Ltd. (Beijing, China) using an SDS-based protocol, and samples with a primary band >30 kb were selected for library construction. Sequencing libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 by Novogene Co., Ltd. The SMRTbell libraries were sequenced on a PacBio Sequel II platform with SMRT Cell 8 M Tray following the manufacturer's protocol. Raw sequencing output consisted of polymerase reads, which are circular consensus sequences containing adapters at both ends. Adapter sequences were removed, and subreads were generated by splitting polymerase reads at the adapter sites. Subreads shorter than 50 bp were filtered out. Circular consensus sequences (HiFi reads) were then generated from the filtered subreads using the PacBio *ccs* software (v6.4.0; Pacific Biosciences) with the parameters–min-passes 3 and–min-rq 0.99, ensuring all retained reads had a quality value (QV) ≥ 20. These high-quality HiFi reads were used as the primary input for downstream genome assembly[19].

Hi-C libraries were constructed based on a previously described protocol with slight adjustments[14]. In brief, fresh tissue samples were treated with 1% formaldehyde to cross-link DNA and associated proteins. The cross-linked chromatin was digested using the MboI restriction enzyme, and the resulting DNA ends were labeled with biotin-14-dCTP to enable the selective enrichment of ligation junctions. DNA fragments generated from proximity ligation were purified using the QIAamp DNA Mini Kit (Qiagen). The recovered DNA was then randomly fragmented to an average size of ~350 bp, after which standard Illumina library preparation steps were applied. Sequencing was carried out on an Illumina HiSeq platform with 150 bp paired-end reads. A total of 272 Gb of Hi-C sequencing data (~60 × genome coverage) was generated. The Hi-C reads were processed using the Juicer v1.6.2[20] pipeline to generate Hi-C contact information.

Transcriptome sequencing procedures were conducted by commercial service provider (Novogene Co., Ltd.) using standard protocols.

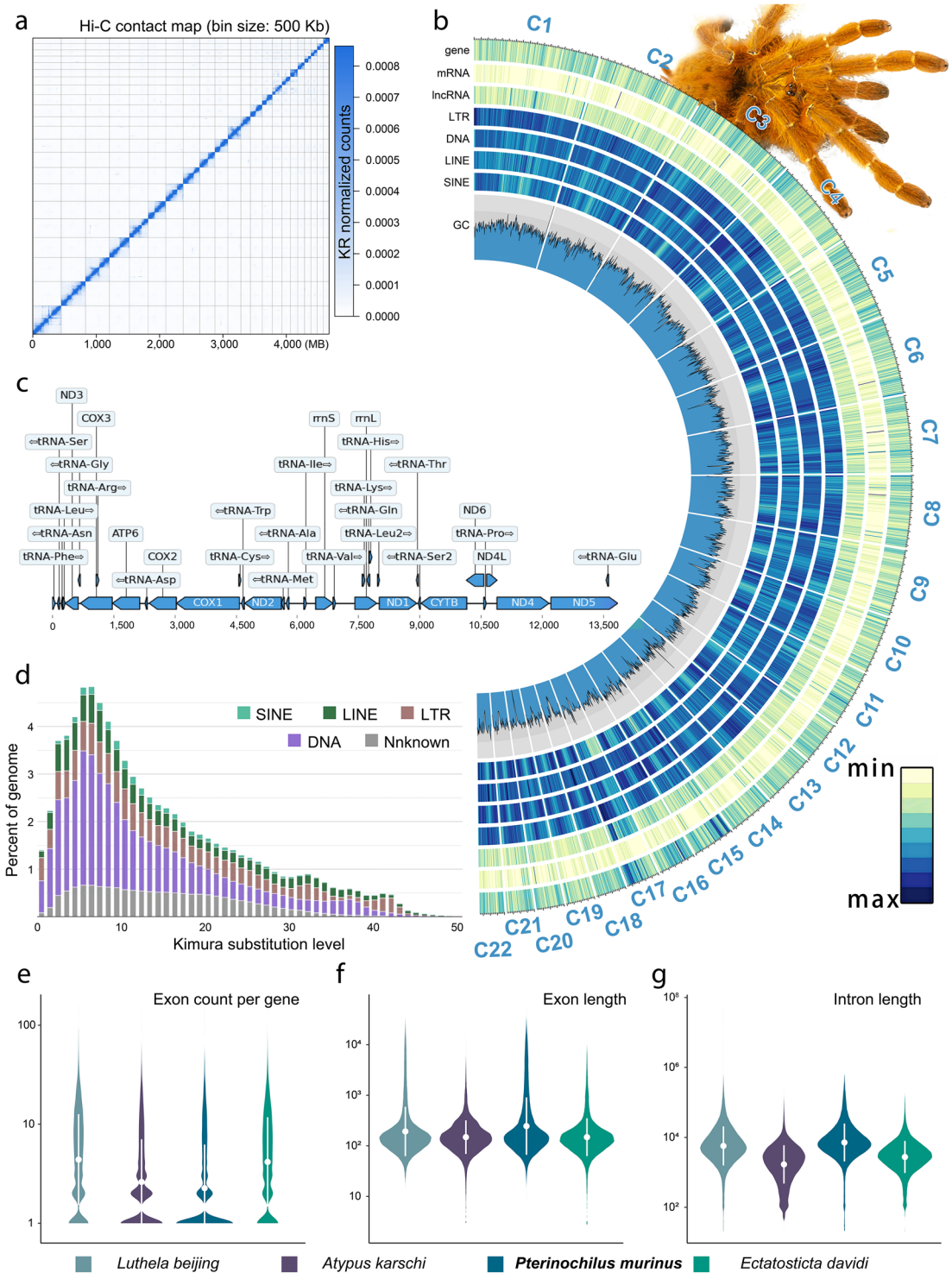

Fig. 1 Characterization of *Pterinochilus murinus* genome. (**a**) Hi-C interaction matrix demonstrating chromosome-scale assembly continuity. (**b**) Circos plot of the genomic characteristics. A female *P. murinus* is shown in the upper right corner (Photographer: Tongyao Jiang). From the outer ring to the inner ring are the distributions of gene, mRNA and lncRNA density, Transposable Elements (LTR, DNA, LINE, SINE), and GC content. (**c**) Mitochondrial assembly and annotation information. (**d**) Repeat element landscape showing evolutionary dynamics of interspersed repeats. (**e,f**) Comparative genomic feature analysis of Exon count per gene, Exon length and Intron length distributions across *P. murinus* and three reference species.

***De novo* genome assembly.** The genome size of *P. murinus* has been previously estimated by Feulgen image analysis densitometry (FIAD)[21] to range between 4.6 Gb and 5.3 Gb[17], and the chromosome number has been determined by karyotyping to be 2n = 22[18]. HiFi *k-mer*–based genome profiling was performed with Jellyfish[22] and

file name: C8.genome.fa					
sequences: 1151					
total length: 4686383565 bp (4684823565 bp excl N/X-runs)					
GC level: 39.97%					
bases masked: 3552120872 bp (75.80%)					
			number of elements*	length occupied (bp)	percentage of sequence
Retroelements			3484903	1276592612	27.24%
	SINEs:		675338	136551597	2.91%
		Penelope	456334	143630687	3.06%
	LINEs:		1371597	545325948	11.64%
		CRE/SLACS	0	0	0.00%
		L2/CR1/Rex	292195	156009168	3.33%
		R1/LOA/Jockey	184999	83194899	1.78%
		R2/R4/NeSL	1014	242743	0.01%
		RTE/Bov-B	183075	57904764	1.24%
		L1/CIN4	11127	900738	0.02%
	LTR elements:	1437968	594715067	12.69%	
		BEL/Pao	35163	14345767	0.31%
		Ty1/Copia	239378	126558755	2.70%
		Gypsy/DIRS1	268100	198796546	4.24%
		Retroviral	5077	1935156	0.04%
DNA transposons			3876008	1635631836	34.90%
	hobo-Activator	320813	108286543	2.31%	
	Tc1-IS630-Pogo	3291397	1418049861	30.26%	
	En-Spm		0	0	0.00%
	MuDR-IS905	0	0	0.00%	
	PiggyBac		699	359127	0.01%
	Tourist/Harbinger	0	0	0.00%	
	Other(Mirage, P-element, Transib)	1684	1141555	0.02%	
Rolling-circles			20356	3413448	0.07%
Unclassified:			3712125	612039370	13.06%
Total interspersed repeats:				3524263818	75.20%
Small RNA:			615880	127425930	2.72%
Satellites:			1102	551564	0.01%
Simple repeats:			478828	19072689	0.41%
Low complexity:			64392	3359107	0.07%

Table 1. Repeat annotation in the *Pterinochilus murinus* 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+.

GenomeScope2[23] ($k = 21$ and 31), yielding genome size estimates of approximately 3.9–4.3 Gb and low heterozygosity (0.72–0.76%, Figure S2, see Supplementary Information). Given the heterozygosity level, we performed *de novo* contig assembly of the HiFi reads using both hifiasm v0.19.7[24] with the–primary mode and wtdbg2 v2.5[25] was supplied with the expected genome size to minimize redundant haplotigs. For each assembly, we first performed standardized post-processing steps that were either implemented within or recommended by the corresponding assembly pipelines. Subsequently, scaffolding was conducted using HapHiC v1.0.7[26] with the Hi-C dataset, and chromosome-scale scaffolds were further manually curated in Juicebox v2.17.00[27]. The assembly generated by wtdbg2, when scaffolded with Hi-C data, exhibited the highest contiguity and best agreement with prior expectations based on the species genome characteristics[17,18], and was therefore selected as the final assembly. In addition, guided by annotation information, read-coverage and alignment evidence, we carefully excluded from the final version any contigs showing clear signatures of redundant haplotypes or repeat-derived sequences (simultaneously lacked any predicted gene annotation, were shorter than 1 Mb in length, and exhibited mean sequencing coverage less than half of the genome-wide average), thereby improving representation of the haploid genome (Table 2).

Genome annotation. A *de novo* repeat library was constructed using RepeatModeler v2.0.2[28]. which was subsequently employed to identify and mask repetitive elements in the *P. murinus* genome assembly via RepeatMasker v4.1.2-p1[29]. In total, repetitive sequences comprised 75.8% of the genome. The most abundant repeat classes were DNA transposons (34.9%), followed by LTR retrotransposons (12.7%), LINEs (11.64%), SINEs (2.91%) and unclassified repeats (13.1%) (Table 1).

Protein-coding gene annotation was conducted using transcriptome evidence obtained from an in-house dataset generated from the same individual used for genome sequencing. RNA-seq reads were aligned to the

	contigs		scaffolds		
	hifiasm	wtdbg2	assembly (from hifiasm)	final assembly (from wtdbg2)	gene set (protein)
BUSCO					
Complete	2847	2799	2846	**2826**	2787
Complete and single-copy	2593	2590	2458	**2503**	1498
Complete and duplicated	254	209	388	**323**	1289
Fragmented	32	53	32	**36**	45
Missing	55	82	56	**72**	102
n=2934					
Assembly (quast)					
# contigs (>=0 bp)	2137	22062	2025	**1151**	
# contigs (>=1000 bp)	2137	22062	2025	**1151**	
# contigs (>=5000 bp)	2137	21704	2025	**1120**	
# contigs (>=10000 bp)	2137	18212	2024	**648**	
# contigs (>=25000 bp)	2100	9862	1988	**194**	
# contigs (>=50000 bp)	1679	6379	1569	**62**	
Total length (>=0 bp)	5745601940	4756414851	5745623240	**4686383565**	
Total length (>=1000 bp)	5745601940	4756414851	5745623240	**4686383565**	
Total length (>=5000 bp)	5745601940	4754905951	5745623240	**4686250580**	
Total length (>=10000 bp)	5745601940	4729420384	5745614428	**4682871981**	
Total length (>=25000 bp)	5744810259	4591046071	5744840490	**4675717425**	
Total length (>=50000 bp)	5728649618	4473788565	5728735689	**4671488878**	
# contigs	2137	22062	2025	**1151**	
Largest contig	179441458	14916085	489926907	**447773322**	
Total length	5745601940	4756414851	5745623240	**4686383565**	
GC (%)	40.12	39.96	40.12	**39.97**	
N50	51545645	1485895	262763734	**274866492**	
N75	19320802	590440	102981035	**160710360**	
L50	29	799	9	**7**	
L75	75	2070	17	**13**	
# N's per 100 kbp	0	0	0.3	**33.29**	
other					
Scaffold number at chromosome level	—	—	21	**22**	
The proportion of sequences on chromosome	—	—	81.50%	**99.60%**	

Table 2. Evaluation results of genome assembly.

genome assembly with STAR v2.7.1[30], and the aligned reads were assembled into transcripts using StringTie v2.2.1[31] with the genome assembly as a reference. Coding sequences (CDSs) were predicted from the assembled transcripts using TransDecoder v5.5.0. An alternative annotation strategy combining homology-based (annotations of *Luthela beijing*, *Atypus karschi*, *A. geniculata*, and *Ectatosticta davidi*)[5,15,32], transcriptomic and *ab initio* prediction using BRAKER3 v3.0.8[33] was evaluated but resulted in an inflated gene count (114,219 protein-coding genes). Consequently, only the transcriptome-supported gene set was retained.

The resulting gene models were functionally annotated by searching against the NCBI non-redundant protein database (Nr)[34] and the EggNOG v5.0[35] database.

The mitochondrial genome of *P. murinus* was assembled and annotated using the MitoHiFi v3.2.2[36] pipeline. First, we used the built-in "findMitoReference.py –species" function to automatically identify an appropriate reference mitogenome, which retrieved that of *Ornithoctonus huwena* (NC_005925.1). We then used this reference as a guide for assembly and annotation, resulting in a complete mitogenome with 35 annotated genes, including 21 tRNA genes (Fig. 1c).

A protocol inspired by the pipeline described in previous articles[37] was adapted to identify lncRNA loci in the *P. murinus* assembly. In brief, final transcript assemblies were first compared against the annotated genome GFF using gffcompare v0.12.8[38], and candidate transcripts $\geq$ 200 bp lacking overlap with annotated protein-coding exons were selected. These were then evaluated for coding potential using CPC2_standalone v1.0.1[39], and only those predicted as non-coding were retained. To ensure expression support and remove likely artifacts, transcripts with sequence coverage $\geq$ 5 and FPKM $\geq$ 0.1 were retained as high-confidence lncRNA models.

Data Records

The sequencing datasets generated in this study, including HiFi reads (SRR34259886, SRR34259885), Hi-C reads (SRR34259884), and transcriptome data (SRR34259883), have been deposited in the NCBI Sequence Read Archive (SRA) under the accession number SRP595452[40]. The chromosome-scale genome assembly has been deposited in GenBank under the accession number GCA_053477515.1[41]. In addition, all raw sequencing data,

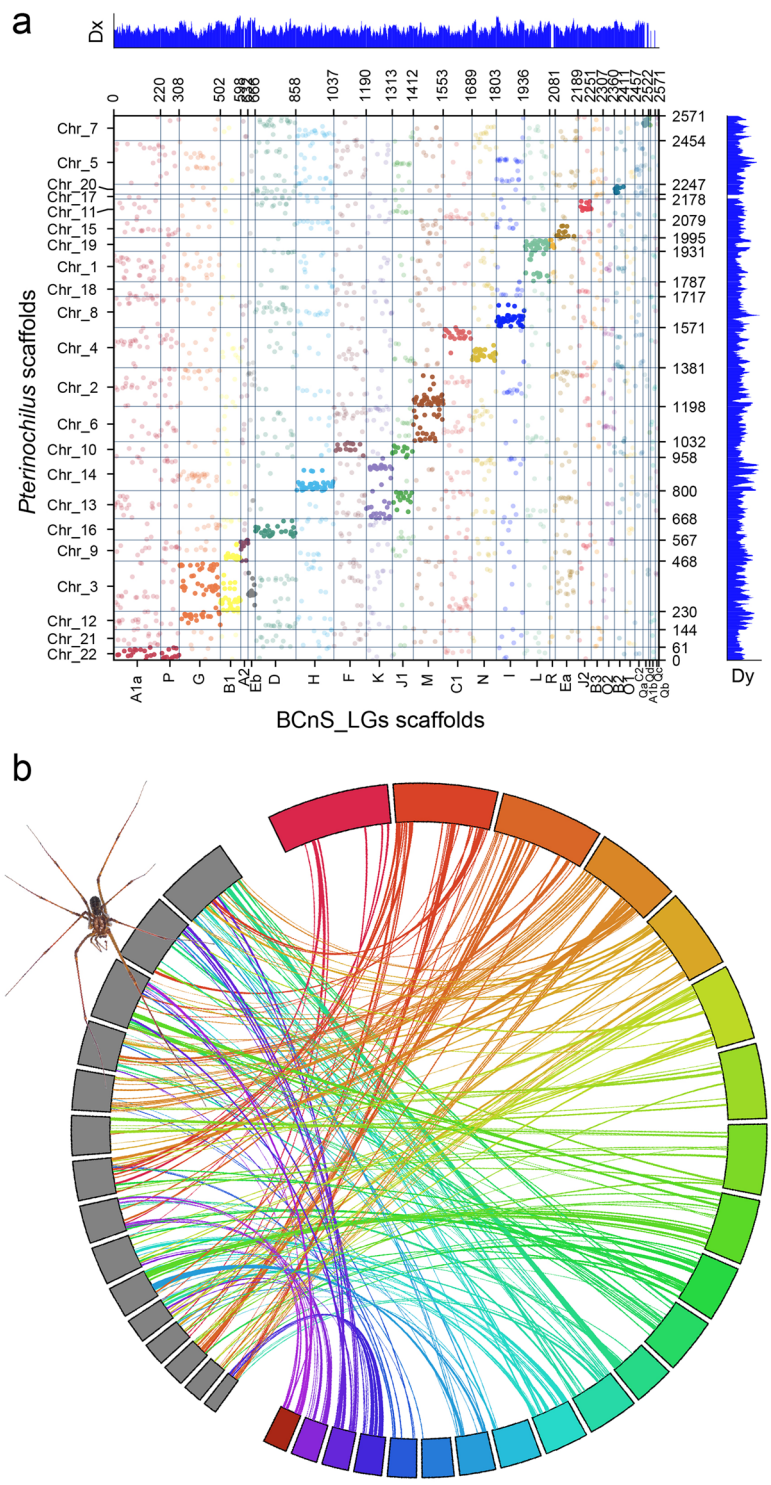

Fig. 2 Oxford dot plot and diagram of Chromosomal collinearity. (**a**) Oxford dot plots representing orthologous genes shared on the previously reported ancestral linkage groups (bilaterian–cnidarian–sponge, BCnS) and *Pterinochilus murinus*. Each dot in the plot represents an ortholog, specifically a reciprocal best diamond blastp match between two species. The color of the dots represents different ancestral linkage groups. Dots that are a solid color are in cells with a FET p-value less than or equal to 0.05. Dots that are translucent are in cells with a FET p-value greater than 0.05. (**b**) Chromosomal collinearity between *P. murinus* (colored) and reference species (*Ectatosticta davidi*, gray). Photographer: Tongyao Jiang.

along with the assembled genome and associated annotation files, have been deposited in Science Data Bank (ScienceDB, sciencedb.27183)[42] to ensure long-term preservation and accessibility.

Technical Validation

During both the contig assembly and the Hi-C–assisted scaffolding stages, we continuously monitored assembly quality using BUSCO v5.2.2[43,44] (-m genome) and QUAST[45,46]. Based on these evaluations, we selected the Hi-C assembly derived from wtdbg2-generated contigs as the final genome assembly. The final assembly contained 2,826 complete hits (96.32%), including 323 duplicated hits (11.00%), with only 36 genes classified as fragmented (1.23%) and 72 orthologs missing (2.45%) entirely according to BUSCO evaluation. This assembly achieved the expected chromosome number, exhibited the fewest scaffold fragments, and showed the highest proportion of sequences anchored to chromosomes (Table 2).

To further assess assembly accuracy and completeness, all PacBio HiFi reads were mapped back to the final assembly, resulting in a mapping rate of 99.96% and an average sequencing depth of 31.81×, indicating highly consistent coverage and excellent assembly integrity.

Based on the final assembly, we conducted comparative analyses of gene structure with three other spider species—*L. Beijing*, *A. karschi*[15] and *E. davidi*[32]. The comparisons included metrics such as exon length per gene, exon count per gene, and intron length per gene. In all cases, the annotation metrics for *P. murinus* were consistent with expectations and within the range observed in the other species (Fig. 1e–g).

Furthermore, an Oxford dot plot analysis[47] of ancestral linkage groups across bilaterian–cnidarian–sponge (BCnS)[48] confirmed the overall structural integrity of the final assembly (Fig. 2a). We then performed a synteny analysis with *E. davidi*—the only closely related species with a fully annotated genome currently available (Fig. 2b). These analyses provide a comparative framework to assess the quality of our genome assembly and annotation.

Data availability

All sequencing data generated in this study are available in the NCBI SRA under the accession number SRP595452. The chromosome-scale genome assembly has been deposited in GenBank under the accession number GCA_053477515.1. All associated data are also available in the Science Data Bank: https://doi.org/10.57760/sciencedb.27183.

Code availability

All computational analyses adhered strictly to the published documentation and standard workflows provided by the software developers. We exclusively used officially released versions of the programs, without any custom scripting or algorithmic modifications. This practice ensures full transparency and reproducibility of our results.

Received: 29 August 2025; Accepted: 16 December 2025;
Published online: 03 January 2026

References

1. Herzig, V. & King, G. F. The Neurotoxic Mode of Action of Venoms from the Spider Family Theraphosidae. in *Spider Ecophysiology* (ed. Nentwig, W.) 203–215. https://doi.org/10.1007/978-3-642-33989-9_15 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
2. Mammola, S. *et al.* Record breaking achievements by spiders and the scientists who study them. *PeerJ* **2017** (2017).
3. Ibler, B., Michalik, P. & Fischer, K. Factors affecting lifespan in bird-eating spiders (Arachnida: Mygalomorphae, Theraphosidae) - A multi-species approach. *Zool Anz* **253**, 126–136 (2013).
4. *NMBE - World Spider Catalog*. https://wsc.nmbe.ch/.
5. Sanggaard, K. W. *et al.* Spider genomes provide insight into composition and evolution of venom and silk. *Nat Commun* **5**, 3765 (2014).
6. Foley, S., Saranathan, V. & Piel, W. H. The evolution of coloration and opsins in tarantulas: Tarantula Coloration and Opsin Evolution. *Proceedings of the Royal Society B: Biological Sciences* **287** (2020).
7. Chatzaki, M. & Komnenov, M. Description of a new Chaetopelma (Araneae, Theraphosidae) species from Crete and a re-description of *Macrothele cretica* Kulczyński, 1903 (Araneae, Macrothelidae). *Zootaxa* **4544**, 269–284 (2019).
8. Foley, S. *et al.* Tarantula phylogenomics: A robust phylogeny of deep theraphosid clades inferred from transcriptome data sheds light on the prickly issue of urticating setae evolution. *Mol Phylogenet Evol* **140**, 106573 (2019).
9. Zhang, Y. *et al.* A trade-off in evolution: the adaptive landscape of spiders without venom glands. *Gigascience* **13** (2024).
10. Zhu, B. *et al.* Chromosomal-level genome of a sheet-web spider provides insight into the composition and evolution of venom. *Mol Ecol Resour* **22**, 2333–2348 (2022).
11. Zhu, B. *et al.* Genomic and transcriptomic analyses support a silk gland origin of spider venom glands. *BMC Biol* **21**, 1–17 (2023).
12. Miles, L. S. *et al.* Insight into the adaptive role of arachnid genome-wide duplication through chromosome-level genome assembly of the Western black widow spider. *Journal of Heredity* **115**, 241–252 (2024).
13. 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** (2021).
14. Zhang, Y. *et al.* An annotated chromosome-level genome of a troglophilic spider (*Pimoa clavata*). *Scientific Data* **12**, 1–8 (2025).
15. Jin, P. *et al.* Single-cell transcriptomics reveals the brain evolution of web-building spiders. *Nature Ecology & Evolution 2023 7:12* **7**, 2125–2142 (2023).
16. Schöneberg, Y. *et al.* Three Novel Spider Genomes Unveil Spidroin Diversification and Hox Cluster Architecture: *Ryuthela nishihirai* (Liphistiidae), *Uloborus plumipes* (Uloboridae) and *Cheiracanthium punctorium* (Cheiracanthiidae). *Mol Ecol Resour* **25**, e14038 (2025).
17. Yorke, H. L. *Exploring Genome Size Diversity in Arachnid Taxa*. http://hdl.handle.net/10214/17735.
18. Král, J. *et al.* Evolution of karyotype, sex chromosomes, and meiosis in mygalomorph spiders (Araneae: Mygalomorphae). *Biological Journal of the Linnean Society* **109**, 377–408 (2013).
19. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology 2019 37:10* **37**, 1155–1162 (2019).
20. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
21. Hardie, D. C. *et al.* From Pixels to Picograms: A Beginners' Guide to Genome Quantification by Feulgen Image Analysis Densitometry. *Journal of Histochemistry & Cytochemistry* **50**, 735–749 (2002).
22. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
23. Ranallo-Benavidez, T. R. *et al.*GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. https://doi.org/10.1038/s41467-020-14998-3.

24. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
25. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155–158 (2020).
26. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants 2024 10:8* **10**, 1184–1200 (2024).
27. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99 (2016).
28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **25**, 4.10.1–4.10.14 (2009).
30. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
31. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 1–13 (2019).
32. Fan, Z. *et al*. Lampshade web spider *Ectatosticta davidi* chromosome-level genome assembly provides evidence for its phylogenetic position. *Communications Biology 2023 6:1* **6**, 1–12 (2023).
33. Gabriel, L. *et al*. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
34. *NCBI Non-Redundant Protein Sequences Database*. https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz.
35. Cantalapiedra, C. P. *et al*. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
36. Uliano-Silva, M. *et al*. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 1–13 (2023).
37. Jia, H. *et al*. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* **16**, 1478–1487 (2010).
38. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, 304 (2020).
39. Kang, Y. J. *et al*. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* **45**, W12–W16 (2017).
40. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP595452 (2025).
41. *NCBI GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_053477515.1 (2025).
42. Zhang, Y. Genome assembly of *Pterinochilus murinus*. V2. *Science Data bank* https://doi.org/10.57760/sciencedb.27183 (2025).
43. Manni, M. *et al*. 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).
44. Manni, M. *et al*. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc* **1** (2021).
45. Mikheenko, A. *et al*. Icarus: Visualizer for de novo assembly evaluation. *Bioinformatics* **32**, 3321–3323 (2016).
46. Mikheenko, A. *et al*. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
47. Schultz, D. T. *et al*. Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
48. Simakov, O. *et al*. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv* **8**, 5884 (2022).

Acknowledgements

This research was supported by the Science and Technology Project of Hebei Education Department (BJ2025152), the National Natural Science Foundation of China (NSFC 32570534), and the Open Project of the Hebei Key Laboratory of Animal Diversity (HKLAD001). We sincerely thank Professor Shuqiang Li (Anhui Normal University) for his continuous guidance and insightful suggestions throughout this study. We also acknowledge Chenkai Tang for his assistance in genome annotation.

Author contributions

Y.Z., X.Y. and J.L. designed the project. F.L., P.J. and T.J. 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

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06507-2.

Correspondence and requests for materials should be addressed to J.L. or X.Y.