scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Agaricogonopus acrotrifoliolatus* (Diplopoda: Spirostreptida)

Chong Liu[1,2,4], JiaJia Wang[1,4], ShiJia Zhu[1,2], ZeWu Han[1,2], Xuankong Jiang[3] ✉ & Yan Dong[1] ✉

Agaricogonopus acrotrifoliolatus **(Diplopoda: Spirostreptida) is the largest millipede species from China. Morphologically, it is characterized by a pair of prominent, fungiform gonopodal processes. Despite the ecological importance of Spirostreptida millipedes as key decomposers in soil ecosystems, genomic resources for this order are unavailable. Herein, we report the first chromosome-level genome assembly for Spirostreptida, represented by *A. acrotrifoliolatus*. The genome was constructed using a combination of PacBio HiFi long reads, Illumina short reads, and Hi-C sequencing. An initial assembly totaling 2.33 Gb was generated. Following Hi-C integration, 96.16% of the assembly was anchored to 11 pseudochromosomes, resulting in a final chromosome-level assembly of 2.24 Gb. The assembly exhibited high contiguity and completeness, with a scaffold N50 of 360.24 Mb and a 96.20% Benchmarking Universal Single-Copy Orthologs (BUSCO) score. Furthermore, 12,492 protein-coding genes and ~1.71 Gb (76.27%) of repetitive elements were identified. As the first reference genome for Spirostreptida, this assembly serves as a crucial genomic resource for advancing millipede genomics and evolutionary studies.**

Background & Summary

Millipedes (Diplopoda), as key decomposers, play vital roles in organic matter decomposition and nutrient cycling within soil ecosystems[1,2]. According to literature, the class Diplopoda comprises >12,000 described species across 145 families and 16 orders[3,4]. Most Diplopoda range 1–100 mm in body length[5–13]. In contrast, members of the order Spirostreptida are known for their exceptional size, often >300 mm in length[5–13]. This large body size enables Spirostreptida to efficiently fragment coarse organic debris, promoting early-stage decomposition and microbial colonization[4,14]. These traits underscore their ecological importance as macrodetritivores in soil ecosystems[15]. Despite their ecological relevance, genomic studies on Spirostreptida are unavailable, a gap also evident in public databases. As of July 2025, only 36 genome assemblies for Diplopoda were available in the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=7553), representing 7/16 recognized orders: Chordeumatida, Glomerida, Julida, Platydesmida, Polydesmida, Polyzoniida, and Spirobolida. Among these, only three species have chromosome-level genome assemblies, including *Cylindroiulus punctatus* (GCA_965125795.1; 354.1 Mb, 7 chromosomes, scaffold N50 = 47,054.77 kb, order Julida), *Choneiulus palmatus* (GCA_965278725.1; 626.5 Mb, 14 chromosomes, scaffold N50 = 42,358.63 kb, order Julida), and *Nanogona polydesmoides*[16] (GCA_965153395.1; 406.3 Mb, 16 chromosomes, scaffold N50 = 25,383.15 kb, order Chordeumatida). To date, no genomic resource of any kind has been reported for Spirostreptida. *Agaricogonopus acrotrifoliolatus*, a member of Spirostreptida, was first described from the tropical rainforests of Xishuangbanna, Yunnan, China[17]. It is characterized by a pair of prominent fungiform processes arising from the gonopods and a tripartite basal plate (gonocoxite)[17]. As the largest millipede species documented in China, it is morphologically distinctive and taxonomically representative. These attributes make *A. acrotrifoliolatus* an excellent candidate for genome sequencing to address

[1]School of Biological Science and Food Engineering, Chuzhou University, Chuzhou, 239000, Anhui, China. [2]School of Life Sciences, Anhui University, Hefei, 230000, Anhui, China. [3]Guizhou Institute of Biology, Guizhou Academy of Sciences, Guiyang, 550009, China. [4]These authors contributed equally: Chong Liu, JiaJia Wang. ✉e-mail: antoma93@gmail.com; dongyan_bio@126.com

Libraries	Sequencing platform	raw data	clean data	Depth* (X)
Pacbio HiFi	PacBio Revio	85,310,084,313	85,310,084,313	~37
Illumina	Illumina HiSeq 2000	117,771,967,200	113,519,199,000	~49
Hi-C	Illumina HiSeq 2000	7,970,022,900	7,038,342,633	—
HiFi RNA-seq	PacBio Revio	4,322,382,619	4,214,810,784	—
HiFi RNA-seq	PacBio Revio	2,170,671,856	2,125,263,958	—
Illumina RNA-seq	Illumina HiSeq 2000	7,496,587,500	7,274,303,739	—

Table 1. Statistics of the sequencing data. *For ease of calculation, the genome size of *Agaricogonopus acrotrifoliolatus* is assumed to be 2.3 Gb.

the current lack of genomic data for Spirostreptida. In this study, we report the first chromosome-level genome assembly for the order Spirostreptida, represented by *A. acrotrifoliolatus*. The assembled genome spans ~2.24 Gb and includes 11 chromosome-scale scaffolds, with a scaffold N50 of 360.24 Mb. Individual chromosomes range in length from 53 Mb to 394 Mb. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, based on the arthropoda_odb10 lineage dataset, revealed a high completeness score of 96.20%. The combination of a high scaffold N50 and strong BUSCO results indicates that the assembly is both highly contiguous and complete. As the first genome assembly for Spirostreptida, the genome of *A. acrotrifoliolatus* not only addresses a major gap in genomic data for this order but also provides a valuable foundation for future evolutionary and functional studies.

Methods

Sample collection and sequencing. A male specimen of *A. acrotrifoliolatus* was collected from the Xishuangbanna Tropical Botanical Garden in Yunnan Province, China. Immediately after collection, the sample was preserved at −80 °C to maintain DNA integrity for genome sequencing. To minimize potential contamination from gut microbiota, we dissected the head, thorax, gonopods, and body muscle tissues separately and extracted genomic DNA from these tissues for downstream sequencing. A multi-platform sequencing approach was used to generate a chromosome-level reference genome, incorporating PacBio HiFi long-read sequencing, Illumina short-read sequencing, and Hi-C chromatin conformation capture technologies.

For HiFi sequencing, genomic DNA was extracted using the Blood & Cell Culture DNA Midi Kit (Qiagen, Hilden, Germany). Following circular DNA template preparation, DNA fragments >20 kb were selected for SMRTbell library construction. Then, whole-genome sequencing was conducted on the PacBio Revio platform, yielding ~85 Gb of HiFi long-read data (Table 1).

For Hi-C sequencing, a high-throughput chromatin conformation capture (Hi-C) library was prepared using formaldehyde crosslinking, DpnII restriction enzyme digestion, biotin labeling, proximity ligation, and purification[18]. The resulting libraries, with insert sizes ranging 300–600 bp, were sequenced on an Illumina HiSeq 2000 platform, yielding ~8 Gb of 150-bp paired-end reads (Table 1).

For transcriptome sequencing. A male specimen of A. acrotrifoliolatus was dissected to collect head, legs, gonopods, and body muscle tissues along the anterior–middle–posterior gradient. Total RNA was extracted from these tissues using TRIzol reagent. mRNA was reverse-transcribed to cDNA, and RNA-seq libraries were constructed using the TruSeq RNA v2 kit, while full-length transcriptome libraries were prepared with the SMRTbell Express Template Prep Kit 2.0. Sequencing was performed on Illumina HiSeq 2000 and PacBio Revio platforms, generating both short-read and full-length transcriptome data, which were subsequently quality-controlled and used for genome annotation.

For the genome survey, ~118 Gb of raw data were generated through paired-end sequencing (2 × 150 bp) on the Illumina HiSeq 2000 platform (Table 1).

Genome survey. The raw Illumina reads were processed using fastp (v0.23.4)[19] to obtain high-quality clean reads for downstream analyses. Filtering criteria included (1) reads containing >5 bp of adapter sequence; (2) reads with low sequencing quality (≥50% of bases with Q-scores <19); and (3) reads with high ambiguity (N bases >5%). If either mate in a read pair failed to meet the quality criteria, the entire pair was discarded. Following quality filtering, k-mer analysis (k = 21) was performed using Jellyfish (v2.3.1)[20], and genome features were inferred using GenomeScope2.0[21]. The genome survey estimated a genome size of ~2.28 Gb, with a homozygosity rate of ~0.22%, a total repeat length of ~1.46 Gb, and a model fit of 97.1% (Fig. 1a).

Genome assembly. Prior to the initial genome assembly, raw HiFi long reads were screened for adapter contamination using HiFiAdapterFilt (v2.0.0)[22] with default settings. No adapter-contaminated reads were detected, and the clean reads were retained in FASTQ format for subsequent assembly. The filtered HiFi long reads were assembled using Hifiasm (v0.19.5-r587)[23] to generate the initial draft genome. This assembly spanned ~2.33 Gb and consisted of 1,276 contigs, with a contig N50 of ~31 Mb and a GC content of 41.51%. Assembly completeness was evaluated using BUSCO (v5.4.7)[24] with the *arthropoda_odb10* database, which identified 96.30% of BUSCO genes as complete (Table 2; Fig. 2). To anchor the genome assembly into chromosome-scale scaffolds, Hi-C analysis was performed. Initially, the raw Hi-C reads were quality-filtered using TrimGalore (v0.6.10) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the following parameters: "trim_galore -q 20 --phred33 --fastqc --stringency 1 --length 20 -e 0.1 --paired --dont_gzip --clip_R1 4 --clip_R2 4 --three_prime_clip_R1 4 --three_prime_clip_R2 4 --illumina -o ./." Using the cleaned Hi-C data, the draft genome was

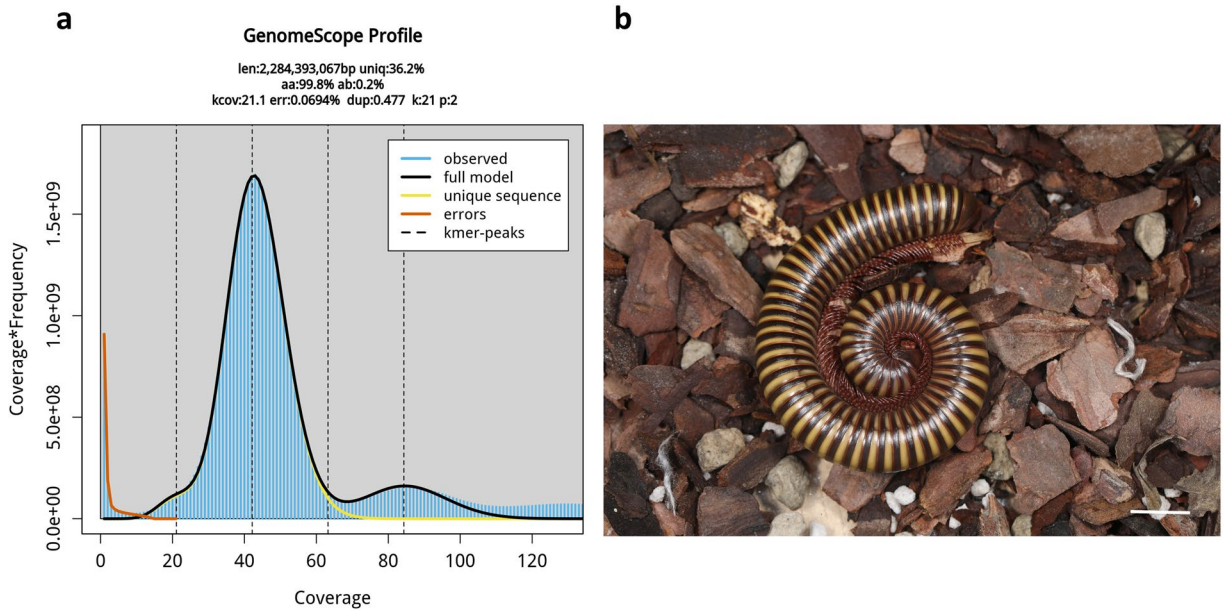

Fig. 1 (**a**) K-mer frequency and genome size estimation of the *Agaricogonopus acrotrifoliolatus* genome. (**b**) Male specimen of *A. acrotrifoliolatus*.Scale bars: 10 mm.

Type	Preliminary Assembly	Chromosome-level Assembly
Assembly size (bp)	2,330,714,204	2,241,317,908
Contig/Scaffold N50 (bp)	31,442,345	360,240,644
Contig/Scaffold N90 (bp)	2,503,738	123,269,442
CC ratio	116	1
Number of contig/pseudochromosome	1,276	11
Longest contig/scaffold (bp)	195,835,358	393,078,066
Contig/Scaffold L50	18	3
Contig/Scaffold L90	103	9
GC content (%)	41.51	41.37
BUSCO completeness (%)	96.30	96.20

Table 2. Genome assembly statistics for *Agaricogonopus acrotrifoliolatus*.

scaffolded to the chromosome level using the AutoHiC (v1.3.3)[25] automated pipeline. This pipeline integrates genome indexing, the Juicer[26] workflow, the 3D-DNA[27] scaffolding protocol, and deep learning–based error correction and structural optimization. In total, 96.16% of the draft genome was anchored to 11 pseudochromosomes, resulting in a final chromosome-level assembly of ~2.24 Gb, with a scaffold N50 of 360.24 Mb (Table 2) and individual chromosome lengths ranging 53–394 Mb (Table 3). Chromosome numbering follows the default AutoHiC output order and does not carry biological significance. The completeness of the final genome assembly was assessed using BUSCO analysis, which identified 96.20% complete, 2.4% fragmented, and 1.4% missing BUSCOs (Fig. 2). These metrics indicate that the anchored genome is highly complete and of sufficiently high quality for downstream genomic and evolutionary analyses.

Genome repeats annotation. A de novo repeat library specific to *A. acrotrifoliolatus* was constructed using RepeatModeler (v2.0.5)[28]. This custom library was integrated with the Dfam3.7[29] and RepBase-20181026[30] databases to generate the final repeat sequence reference. Repetitive elements in the genome were annotated using RepeatMasker (v4.1.5)[31] based on this combined library. The analysis revealed that ~76.27% of the genome consists of repetitive sequences, including long interspersed nuclear elements (LINEs, 34.10%), short interspersed nuclear elements (SINEs, 0.37%), DNA transposons (8.20%), long terminal repeat elements (LTRs, 7.58%), unclassified elements (25.40%), and other repeat types (Table 4; Fig. 3).

Genome annotation. Protein-coding genes were predicted following repeat masking, using three complementary strategies: RNA-seq-based, homology-based, and ab initio gene prediction. For transcriptome-based gene prediction, Illumina RNA-seq reads were aligned to the genome using HISAT2 (v2.1.1)[32], and transcripts were assembled with StringTie (v2.1.0)[33]. Full-length transcriptomic reads from PacBio HiFi sequencing were filtered and clustered using IsoSeq. 3 (v4.0.0)[34], and the polished HiFi isoforms were aligned to the genome using Minimap2 (v2.26)[35]. Both Illumina RNA-seq and PacBio full-length transcriptome data were

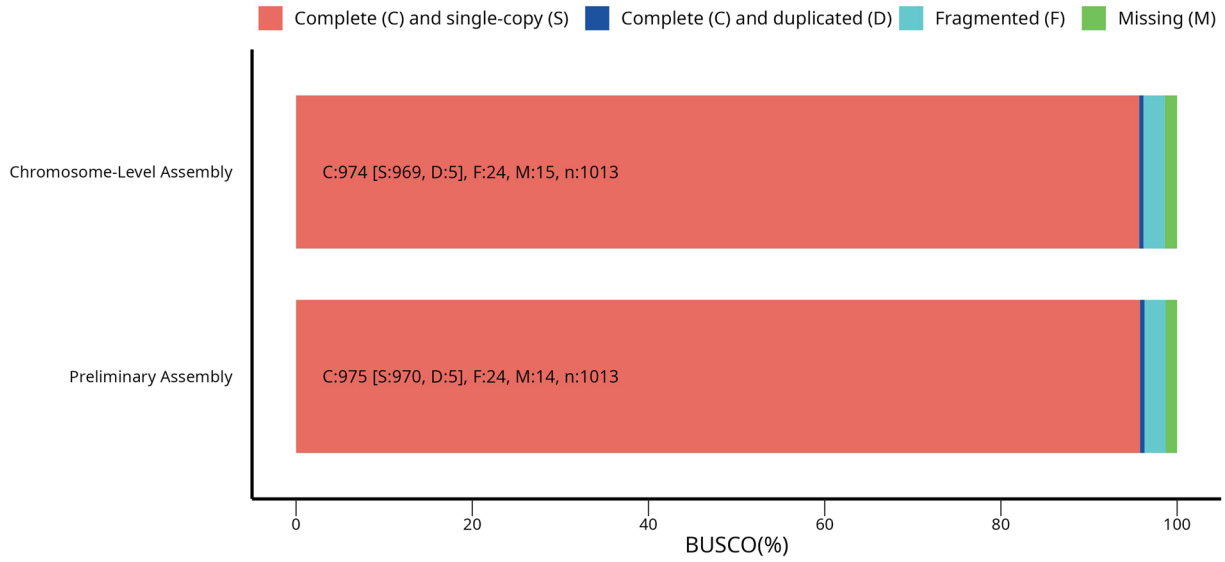

Fig. 2 BUSCO summary of the *Agaricogonopus acrotrifoliolatus* genome. The x-axis represents the percentage of BUSCOs; the y-axis indicates different genome assembly stages.

Chromosome	Length(bp)	GC(%)	% of assembly
Chromosome 1	388,542,410	41.83	16.67
Chromosome 2	53,865,104	40.08	2.31
Chromosome 3	128,125,110	39.84	5.5
Chromosome 4	277,036,024	41.69	11.89
Chromosome 5	123,269,442	40.02	5.29
Chromosome 6	149,133,426	40.29	6.4
Chromosome 7	56,380,136	40.65	2.42
Chromosome 8	126,334,156	42.46	5.42
Chromosome 9	360,240,644	41.49	15.46
Chromosome 10	185,313,390	40.18	7.95
Chromosome 11	393,078,066	42.38	16.87
Total	2,241,317,908	41.37	96.17

Table 3. Statistical summary of the 11 pseudochromosomes in the *Agaricogonopus acrotrifoliolatus* genome.

Type	Number of elements	Length (bp)	% of genome
LINE	743,929	764,347,600	34.10
SINE	58,965	8,240,454	0.37
LTR	120,064	169,814,354	7.58
DNA Transposon	590,883	183,792,974	8.20
Unknown	1,579,163	569,189,018	25.40
Other	31,647	14,055,354	0.62
Total	3,124,651	1,709,439,754	76.27

Table 4. Statistics of repeat elements identified in the *Agaricogonopus acrotrifoliolatus* genome.

subsequently used as evidence for transcriptome-based gene prediction. For homology-based annotation, protein sequences from five representative Diplopoda species, *Anaulaciulus tonginus*, *Glomeris maerens*, *Helicorthomorpha holstii*, *Niponia nodulosa*, and *Trigoniulus corallinus*, were retrieved from Figshare[36] (https://doi.org/10.6084/m9.figshare.15088722.v4). The final protein-coding genes were predicted using BRAKER3 (v3.0.3)[37] in ETP mode, which integrates RNA-seq data, homologous proteins, and ab initio predictions to produce evidence-supported, high-confidence gene models. In total, 12,492 protein-coding genes were predicted. The average gene length was 26,861 bp. The BUSCO analysis indicated high annotation completeness, with 95.8% complete, 0.8% fragmented, and 3.5% missing BUSCOs. To functionally annotate the predicted protein-coding genes, the resulting protein sequences were aligned against multiple public databases, including eggNOG[38].

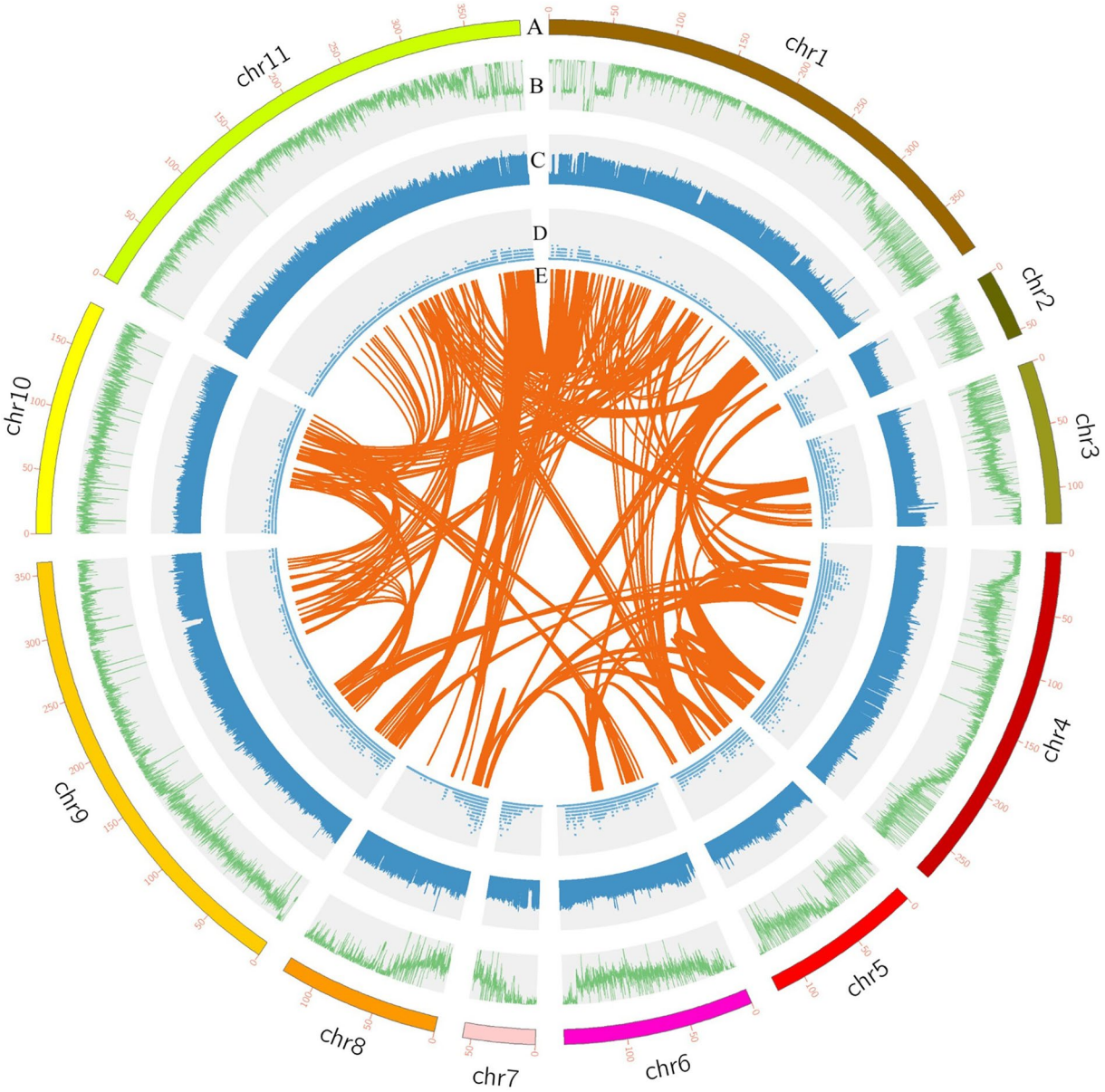

Fig. 3 Genomic features of the assembled *Agaricogonopus acrotrifoliolatus* genome: (**A**) chromosomes, (**B**) repetitive sequence density, (**C**) GC content, (**D**) gene density, and (**E**) colinear relationships among chromosomes.

Type	Total	EggNog	KOG	GO	KEGG	Pfam
Number	12,492	10,757	10,356	8,797	7,823	10,202
Percentage	100%	86.11%	82.90%	70.42%	62.62%	81.67%

Table 5. Statistics of functional gene annotations in the *Agaricogonopus acrotrifoliolatus* genome.

KOG[39], GO[40], KEGG[41], and Pfam[42], using BLASTP[43]. Consequently, 10,757 genes were functionally annotated in eggNOG, 10,356 in KOG, 8,797 in GO, 7,823 in KEGG, and 10,202 in Pfam (Table 5). These findings indicate that the majority of predicted genes have robust functional annotations across multiple databases, reflecting the gene set's high quality and completeness. The annotation of noncoding RNAs was conducted independently. Transfer RNA (tRNA) genes were predicted using tRNAscan-SE (v2.0)[44]. and ribosomal RNA (rRNA) genes were identified using RNAmmer (v1.2)[45] under the eukaryotic model. MicroRNAs (miRNAs), small nuclear RNAs (snRNAs), and other noncoding RNAs (ncRNAs) were annotated using INFERNAL (v1.1.2)[46] via its cmscan program, based on the Rfam database (v14.6)[47]. In total, four major classes of noncoding RNAs were identified: 9,740 rRNAs, 1,046 snRNAs, 51 miRNAs, and 16,342 tRNAs (Table 6).

Type		Number	Average Length (bp)	Totoal Length (bp)
miRNA		51	79.47	4,053
tRNA		16,342	75.53	1,234,293
rRNA	18S	2,990	1,746.61	5,222,353
	28S	3,548	3,781.99	13,418,528
	5S	372	118.73	44,168
	5.8S	2,830	153.87	435,458
snRNA	U1_snRNA	290	161.75	46,908
	U2_snRNA	120	183.98	22,078
	U5_snRNA	70	123.43	8,640
	U6_snRNA	22	107.73	2,370
	U4_snRNA	11	131.64	1,448
	U12_snRNA	1	153.00	153
	Other_snRNA	532	158.43	84,286

Table 6. Statistics of noncoding RNA annotations in the *Agaricogonopus acrotrifoliolatus* genome.

Data Records

All sequencing data reported in this paper, including Illumina, PacBio HiFi, Hi-C, RNA-seq, and PacBio IsoSeq data, have been deposited in the Genome Sequence Archive (GSA) under accession number CRA028901[48]. The assembled genome has been deposited at GenBank under the accession number GCA_052040765.1[49].

Technical Validation

The final genome assembly spans 2.24 Gb, with a scaffold N50 of 360.24 Mb. This size is consistent with the genome size estimated through k-mer analysis and the genome survey (Fig. 1a). Assessment using BUSCO, based on the *arthropoda_odb10* database, identified 96.2% of universally conserved genes, indicating that the *A. acrotrifoliolatus* genome assembly is nearly complete. Similarly, evaluation of the protein-coding gene set using BUSCO in protein mode (-m protein) revealed a completeness score of 95.8%.

Data availability

All sequencing data reported are avaoilable at https://ngdc.cncb.ac.cn/gsa/browse/CRA028901. The assembled genome and annotations are available at https://identifiers.org/ncbi/insdc.gca:GCA_052040765.1.

Code availability

No custom scripts were developed in this study. All analyses were conducted according to the official manuals and standard protocols of the respective bioinformatics tools. Detailed software versions and parameters are provided in the Methods section.

Received: 22 August 2025; Accepted: 11 December 2025;
Published online: 16 December 2025

References

1. Enghoff, H. *et al*. A mountain of millipedes XI. The trachystreptoform spirostreptids of the Udzungwa Mountains, Tanzania (Diplopoda, Spirostreptida, Spirostreptidae). *Eur. J. Taxon.* **918**, 1–50 (2024).
2. Horváthová, T. *et al*. Methanogenesis in the Digestive Tracts of the Tropical Millipedes Archispirostreptus gigas (Diplopoda, Spirostreptidae) and Epibolus pulchripes (Diplopoda, Pachybolidae). *Appl. Environ. Microbiol.* **87**, e0061421 (2021).
3. Edward, K. L. & Harvey, M. S. A review of the australian millipede genus atelomastix (diplopoda: spirostreptida: iulomorphidae). *Zootaxa* **2371**, 1–63 (2010).
4. Golovatch, S. I. & Kime, R. D. Millipede (diplopoda) distributions: a review. *Soil Org.* **81**, 565–597 (2009).
5. Wang, Y., Jin, A., Gao, S., Wang, J. & Dong, Y. Descriptions of four species of polyxenida verhoeff, 1934 (diplopoda, penicillata) from China, including one new species and one new record. *Zookeys* **1223**, 149 (2025).
6. Stoev, P., Moritz, L. & Wesener, T. Dwarfs under dinosaur legs: a new millipede of the order Callipodida (Diplopoda) from Cretaceous amber of Burma. *Zookeys* **841**, 79–96 (2019).
7. Moritz, L., Blanke, A., Hammel, J. U. & Wesener, T. First steps toward suctorial feeding in millipedes: comparative morphology of the head of the platydesmida (diplopoda: colobognatha). *Invertebr. Biol.* **140**, e12312 (2021).
8. Valkay Halľková, B., Tajovský, K., Grego, J., Žurovcová, M. & Mock, A. Geoglomeris subterranea (diplopoda, glomerida, glomeridae), the first morphologically non-specialized semiaquatic glomerid millipede? *Invertebr. Biol.* **141**, e12376 (2022).
9. Antić, D. Karadenizia, a new monospecific pachyiuline genus (diplopoda: julida: julidae) from Turkey. *Zootaxa* **5315**, 456–468 (2023).
10. Huynh, C. & Veenstra, A. Taxonomy and biology of a new species of pincushion millipede of the genus monographis (diplopoda: polyxenidae) from Australia. *Zootaxa* **3721**, 573–588 (2013).
11. Liu, W. Taxonomy of Cavernicolous Millipedes in Southern China and Their Morphological Adaptations to Subterranean life (Myriapoda, Diplopoda). (PhD thesis, South China Agricultural University, 2017).
12. Marek, P. E. *et al*. The first true millipede—1306 legs long. *Sci. Rep.* **11**, 23126 (2021).
13. Wesener, T. The Giant Pill-Millipedes, order Sphaerotheriida - An annotated species catalogue with morphological atlas and list of apomorphies (Arthropoda: Diplopoda). *Bonn zoological Bulletin* **63**, 1–104 (2016).
14. Enghoff, H., Dohle, W. & Blower, J. G. Anamorphosis in millipedes (diplopoda)—the present state of knowledge with some developmental and phylogenetic considerations. *Zool. J. Linn. Soc.* **109**, 103–234 (1993).

15. Brousseau, P., Gravel, D. & Handa, I. T. Traits of litter-dwelling forest arthropod predators and detritivores covary spatially with traits of their resources. *Ecology* **100**, e02815 (2019).
16. Owen, C., Sivell, O., Sivell, D., Twitchett, R. J. & Edgecombe, G. D. The genome sequence of the eyed flat-backed millipede, nanogona polydesmoides (leach, 1814). *Wellcome Open Res.* **10**, 332 (2025).
17. Zhang, C. & Zhang, N. An interesting new millipede, Agaricogonopus acrotrifoliolatus gen. et sp. nov., from the tropical rain-forest region of Xishuangbanna, Yunnan (Diplopoda: Spirostreptida: Harpagophoridae). *Acta Zootaxonomica Sinica* **22**(4), 349–352 (1997).
18. van Berkum, N. L. *et al*. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.: Jove* 1869, https://doi.org/10.3791/1869 (2010).
19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
21. 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).
22. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).
23. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
24. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
25. Jiang, Z. *et al*. A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes. *Nucleic Acids Res.* **52**, e92 (2024).
26. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
27. Dudchenko, O. *et al*. De novo assembly of the aedes aegypti genome using hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
29. Hubley, R. *et al*. The dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–89 (2016).
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
31. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **Chapter 4**, 4.10.1–4.10.14 (2009).
32. 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).
33. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
34. Gordon, S. P. *et al*. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLOS One* **10**, e0132628 (2015).
35. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
36. So, W. L. *et al*. Myriapod genomes reveal ancestral horizontal gene transfer and hormonal gene loss in millipedes. *Nat. Commun.* **13**, 3010 (2022).
37. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol. Biol. (clifton N.J.)* **1962**, 65–95 (2019).
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. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinf.* **4**, 41 (2003).
40. Gene Ontology Consortium The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
41. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
42. El-Gebali, S. *et al*. The pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
43. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
44. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
45. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
46. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
47. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–124 (2005).
48. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA028901 (2025).
49. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052040765.1 (2025).

Acknowledgements

This work was supported by National Natural Science Foundation of China (31401971), The Outstanding Youth Research Project of Colleges and Universities in Anhui Province (No. 2022AH030110), Scientific Research Starting Foundation of Chuzhou University China (2024qd42, 2025qd17).

Author contributions

Conceptualization: Yan Dong; Sample collection: JiaJia Wang, ShiJia Zhu, and ZeWu Han; DNA, RNA extraction and sequencing: JiaJia Wang, ShiJia Zhu, and ZeWu Han; Bioinformatics analysis: Chong Liu, JiaJia Wang and ShiJia Zhu; Visualization: Chong Liu and JiaJia Wang; Original draft preparation: Chong Liu; Manuscript review & editing: Chong Liu, JiaJia Wang, Xuankong Jiang and Yan Dong; Funding acquisition: Yan Dong; Final approval of the manuscript: All authors.

Competing interests

The authors declare no competing interests.

Additional information

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