



OPEN

DATA DESCRIPTOR

The high-quality telomere-to-telomere genome assembly of the earthworm (*Amyntas aspergillum*)

Guangquan Peng^{1,2,5}, Yanghe Qin^{2,5}, Zhiming Yan², Mingwei He², Yucheng Zhao^{3,4} [✉],
Neng Jiang² [✉] & Changhong Wei¹ [✉]

Earthworms have been extensively studied as ancient soil invertebrates, that are highly diverse. Previous studies of these invertebrates have mainly focused on their ecosystem services, medicinal value, and ecological habits. However, their genomic analysis remains inadequate. In this study, we generated the first high-quality telomere-to-telomere (T2T) assembly of the genome of the earthworm, *Amyntas aspergillum* (Perrier, 1872), which belongs to the genus *Amyntas* of the family Megascolecidae. The T2T assembly was 758.86 Mb with an N50 contig size of 16.59 Mb. The sequences were anchored to 43 chromosomes ($2n = 2x = 86$) with a coverage of 98.43% (746.95 Mb), and 83 telomeres were detected. In addition, we also predicted 35,723 protein-coding genes with 97.02% being functionally annotated. This T2T genome assembly will establish substantial groundwork for exploring the evolutionary mechanisms of the earthworm genome and enhance the specificity of its pharmacological effects.

Background & Summary

Earthworms, classified under the Oligochaeta genus within the phylum Annelida, are ancient terrestrial invertebrates with significant ecological and medicinal importance¹. There are more than 3,000 species of earthworms worldwide, with more than 600 species found in China alone². These organisms play a pivotal role in soil ecosystems by decomposing organic matter, thereby enhancing microbial activity and soil fertility^{2,3}. Their ecological contributions extend to waste management and environmental remediation, with applications in sewage purification and soil quality improvement^{3–5}. Moreover, earthworms are used in traditional Chinese medicine to treat various diseases such as high fever, dizziness, joint paralysis, and urinary edema⁶, and they are also used as an emerging high-protein feed.

Amyntas aspergillum (Perrier, 1872), also known as geosaurus or “Guang dilong”, is a terrestrial annelid belonging to the genus *Amyntas* in the family Megascolecidae. It is usually 15–20 cm long and 1–2 cm wide (Fig. 1a). This species is predominantly found in Chinese provinces such as Guangxi, Guangdong, and Fujian⁶. As a traditional Chinese medicine, the earthworm first appeared as the “white-necked earthworm” in *Shennong Bencaojing* of the Eastern Han Dynasty, and it began to be called “Guang dilong” in the *Revised Materia Medica* of the Tang Dynasty⁷. *A. aspergillum* is one of 10 varieties of geo-authentic traditional Chinese medicine in Guangxi and has been widely used in traditional Chinese medicine for thousands of years because of its excellent medicinal properties. Its medicinal form is the dried body (Fig. 1b), and its efficacy is better than that of “Hu dilong” and “Tu dilong”⁷. In addition, *A. aspergillum* is widely used in modern clinical medicine, and studies of its chemical composition and pharmacological activity were conducted as early as 1974⁸. It contains diverse chemical components, of which polypeptides, nucleosides, lipids, enzymes, and amino acids are the main

¹Department of Research & Clinical Laboratory, The Fifth Affiliated Hospital of Guangxi Medical University & The First People's Hospital of Nanning, Nanning, China. ²Department of Pharmacy, Guangxi Medical University Cancer Hospital, Nanning, 530021, Guangxi, PR China. ³State Key Laboratory of Natural Medicines, Department of Resources Science of Traditional Chinese Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing, 211198, China. ⁴Institute for Safflower Industry Research, Key Laboratory of Xinjiang Phytomedicine Resource and Utilization (Ministry of Education), School of Pharmacy, Shihezi University, Shihezi, 832002, China. ⁵These authors contributed equally: Guangquan Peng, Yanghe Qin. ✉e-mail: zhaoyucheng1986@126.com; cujinh@163.com; weichanghong1984@163.com

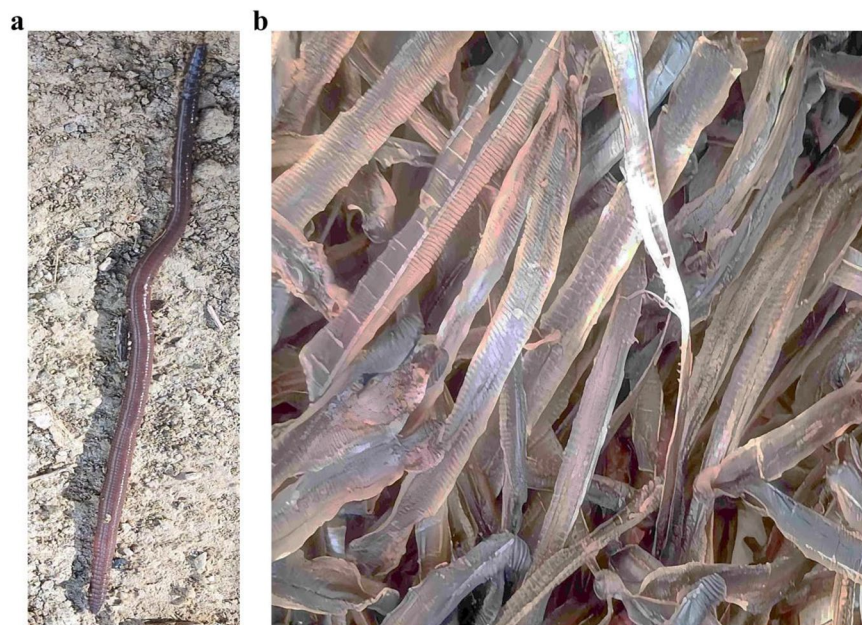


Fig. 1 Photograph of *A. aspergillum*. **(a)** A live sub-adult earthworm. **(b)** The dried bodies used for medicinal purposes.

active ingredients⁹. These components underpin its diverse pharmacological properties, including anti-tumor, anti-thrombotic, anti-hypertensive, immunomodulatory, and wound-healing effects^{10–12}.

In recent years, with increasing research on earthworms, progress has been made on species identification, chemical composition, pharmacological effects, and quality control^{5,13}, but few studies have investigated earthworm genomics. The lack of high-quality genomes and further research on earthworm genomics seriously hinders our in-depth understanding and exploration of their functions and evolutionary processes. To date, the chromosome-level genomes of six earthworm species (*Eisenia fetida*¹⁴, *Eisenia andrei*¹⁵, *Metaphire vulgaris*¹⁶, *Amyntas corticis*¹⁷, *Lumbricus rubellus*¹⁸, and *Lumbricus terrestris*¹⁹) have been sequenced and assembled, laying a foundation for studies of earthworm ecology, evolutionary mechanisms, and molecular mechanisms of immune defense and earthworm regeneration²⁰. However, this is insignificant compared to the large number of earthworms species. Notably, no telomere-to-telomere (T2T) genome has been assembled to date, leaving significant gaps in our knowledge of their genomic architecture and evolutionary mechanisms. The availability of genomes from more species will provide valuable insights into the genetics and molecular mechanisms of earthworms.

Here, we sequenced and assembled the first complete T2T genome of the earthworm *A. aspergillum* by combining Pacific Biosciences (PacBio) HiFi, Oxford Nanopore Technologies (ONT) ultra-long, Illumina, and high-throughput chromosome conformation capture (Hi-C) sequencing technologies. Sequencing data and completeness assessments suggested that the T2T assembly was superior to the previous chromosome-level assembly. The assembly demonstrates exceptional quality, with an N50 contig length significantly exceeding those of previously sequenced earthworm genomes. In summary, we utilized multiple methods combining genetics and cytology to produce a high-quality T2T genome that will bring new opportunities to identify the unique genes and structural variations in the “dark matter” regions, such as centromeres, transposable elements and segmental duplications.

Methods

Sample collection and pre-treatment. One sexually mature *A. aspergillum* was gathered from a breeding base that is located in Longan Town, Nanning, Guangxi, China (107°47'59"E, 23°4'29"N). The sample has been sequenced by COI gene sequencing and identified as *A. aspergillum* with 100% species similarity (Fig. S1). After rinsing with saline solution to remove any attached dirt, the earthworm was dissected to remove its gut. Then the body wall was washed thoroughly three times using 1 × PBS and was carefully sheared into tissues less than 0.5 cm both in length and width. Eventually, the pre-treatment sample was stored at −80 °C and would be used for further DNA extraction and sequencing.

DNA extraction and sequencing. The body tissue of *A. aspergillum* was used for high-quality DNA isolation, and then the PacBio HiFi library, ONT library, Illumina library, and Hi-C library were constructed following the manufacturer's instructions. Briefly, the genomic DNA was damaged, end-repaired, ligated to adapters, and exonuclease digested. Then the digested DNA was screened for target fragments using BluePippin to obtain the PacBio HiFi library, which was sequenced on the PacBio Revio platform and switched to CCS (Circular Consensus Sequence) data using the smrtlink v9.0 of PacBio program. For ultra-long ONT sequencing, a library was generated with the Oxford Nanopore SQK-ULK001 kit following the standardized protocol and

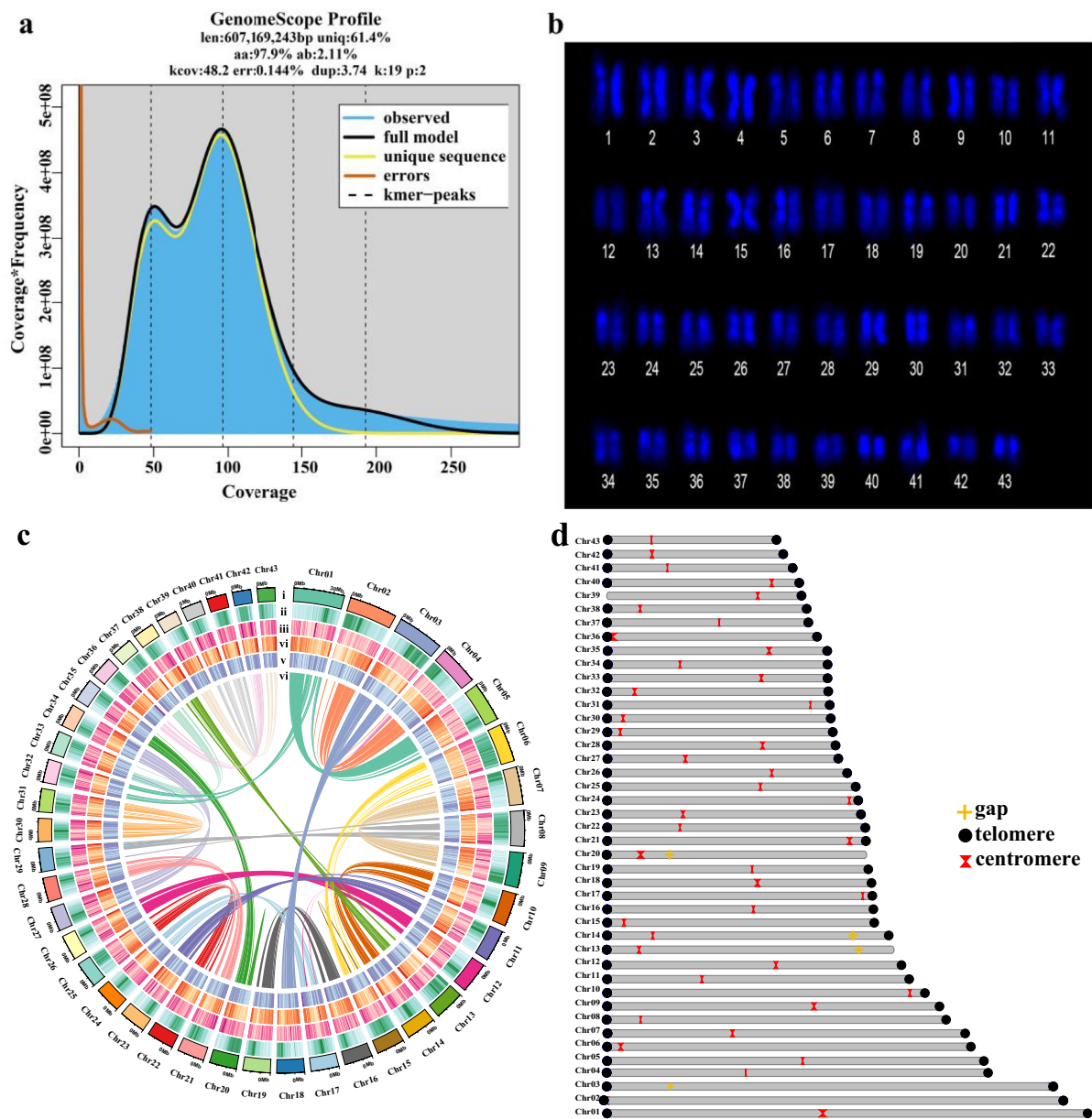


Fig. 2 Genome and chromosome features of *A. aspergillum*. **(a)** K-mer distribution ($k = 19$) in the *A. aspergillum* genome. The graph serves as an indicator of the estimation of genome complexity and provides a strategy for subsequent genome assembly. **(b)** The image of each chromosome in one individual *A. aspergillum* produced in karyotype analysis ($2n = 2x = 86$). **(c)** Genome characteristics of *A. aspergillum*. Circos plot from the outer to the inner layers represents the following: (i) Pseudo-chromosomes (Chr1-Chr43); (ii) TEs density; (iii) SSR density; (iv) Gene density; (v) GC content; (vi) Syntenic blocks within the *A. aspergillum* genome. **(d)** The distributions of chromosome features. The numbers on the left are the numbers of each chromosome (Chr1-Chr43). The circle in yellow and the triangle in black represent gaps and telomeres of chromosomes in the genome, respectively. The crossover site in each chromosome represents centromeres.

then sequenced on the PromethION platform. The Illumina library was constructed through DNA breaking, end repairing, adding A tail, ligating adapters, selecting target fragments, and expanding with PCR. After using Qseq. 400 and Qubit to detect fragment size and quality, the library was sequenced on the Illumina NOvaSeq. 6000 platform (PE150). As a result, we generated 71.00 Gb ($\sim 92\times$) Illumina short reads, 80.34 Gb ($\sim 101\times$) of CCS reads with an average length of 16.21 kb, 25.75 Gb ($\sim 30\times$) of ultra-long ONT reads with an average length of 97.68 kb (Table S1).

For Hi-C sequencing, the experiment type of library construction was *in situ* Hi-C, including crosslinking DNA, cutting with restriction enzyme Hind III, filling ends and marking with biotin, ligating, purifying, and shearing DNA into 300 bp $\sim$ 700 bp fragments and pulling down biotin. The concentration and insert size

Assembly	<i>A. aspergillum</i>	<i>E. fetida</i>	<i>E. andrei</i>	<i>M. vulgaris</i>	<i>A. corticis</i>	<i>L. rubellus</i>	<i>L. terrestris</i>
Total length	758.86 Mb	1.47 Gb	1.3 Gb	728.57 Mb	1.19 Gb	787.5 Mb	1.06 Gb
Contig number	205	463,133	2,970	559	16,882	2,261	1,587
Contig N50	16.59 Mb	967 kb	740 kb	4.20 Mb	117 kb	0.7 Mb	1.6 Mb
GC content	40.76%	40%	35–45%	40%	40.34%	40.5%	40.8%
Complete BUSCOs	95.18%	18.2%	92.1%	94.3%	91.2%	90.6%	91.4%
Protein-coding genes	35,723	—	31,817	—	29,256	33,426	—

Table 1. Basic information on genome assembly in earthworms.

Type		Sequence length (bp)	Percentage of the genome (%)
Transposable elements	LINEs	58,870,972	7.76
	SINEs	2,441,270	0.32
	LTR-RTs	86,816,877	11.44
	DNA transposon	151,509,405	19.97
	Other	2,088	0
	Unknown	32,312	0
Tandem repeats	microsatellite (1–9 bp units)	28,471,601	3.75
	minisatellite (10–99 bp units)	20,117,127	2.65
	satellite (>= 100 bp units)	26,075,171	3.44
Total		374,336,823	49.33

Table 2. Statistics of repeat annotations of *A. aspergillum*.

of the library were detected by Qubit2.0 and Agilent 2100, respectively. Next, the library was sequenced by using the Illumina NOvaSeq. 6000 platform (PE150)²¹. Finally, 130.81 Gb (~171×) of Hi-C reads were generated (Table S1). All DNA isolation, library construction, and gene sequencing procedures were processed by the BIOMARKER Company (Beijing, China) according to the manufacturer's protocols.

Genome survey, *de novo* genome assembly, and telomeres identification. A genome survey of *A. aspergillum* based on 19 K-mer frequencies of Illumina short reads using jellyfish²² v2.1.4 (-h 1000000000) and Genomescope²³ v2.0 (-k 19 -p 2 -m 100000) indicated that the genome was approximately 607.17 Mb with a high level of repetitive sequence content (~38.64%) and heterozygosity (~2.11%) (Fig. 2a). The results showed that the genome of *A. aspergillum* is highly heterogeneous and complex, and we speculated that the earthworm is diploid, which is consistent with the results of the karyotype analysis ($2n = 2x = 86$) (Fig. 2b and S2).

All of the sequencing data were subjected to quality control to filter out adapter sequences and low-quality reads, ensuring that clean data were generated. With the CCS, ultra-long ONT, and Hi-C high-accuracy data, initial contig assembly was generated using hifiasm v0.19.5-r587 (hifiasm -t40-ul). After removing plasmids and contaminating sequences, the contigs were clustered, ordered, and oriented onto chromosomes based on the high-quality Hi-C data by using LACHESIS²⁴ v2.0.1 (CLUSTER_MIN_RE_SITES = 342; CLUSTER_MAX_LINK_DENSITY = 2; ORDER_MIN_N_RES_IN_TRUNK = 652; ORDER_MIN_N_RES_IN_SHREDS = 634), generating a chromosome-level genome. The filtered CCS and ONT reads were initially assembled into 205 contigs with a total length of 758.86 Mb, an N50 of 16.59 Mb, and the largest length of 30.87 Mb (Table 1). A total of 746.95 Mb (98.43%) of genomic sequences were anchored to 43 chromosomes (Fig. 2c,d), and the largest chromosome reached a length of 30.87 Mb (Table S2). Among the sequences anchored to the chromosomes, the length of the sequences that could determine the order and orientation was 746.95 Mb, accounting for 100.0% of the anchored sequences. Furthermore, we identified 72 telomeres, 43 centromeres, and 5 gaps in the assembled chromosome-level genome using TIDK (<https://github.com/tolkitt/telomeric-identifier>), FindTelomeres (<https://github.com/JanaSperschneider/FindTelomeres>), and Centromics software (<https://github.com/ShuaiNIEgithub/Centromics>), of which 14 lacked telomeres at single ends and only 38 chromosomes were gapless T2T chromosomes.

Subsequently, the ultra-long ONT data was used for filling gaps via TGS-GapCloser²⁵ and quarTeT software, and chromosomes missing telomeric regions were aligned to the ends of chromosomes using ultra-long ONT reads to detect the telomeric regions. The transcriptomic data also assisted in optimizing genome structure in three ways: identification of exon-intron boundaries, discovery of new exons and variable splicing, and discovery of non-coding RNAs and UTR regions. Eventually, we obtained a complete T2T genome of *A. aspergillum*, where a total of 83 telomeres, 43 centromeres, and only 4 gaps were identified. It was worth noting that telomeres were detected on both ends of 40 chromosomes, and 38 chromosomes (88.4%) were gapless T2T chromosomes (Fig. 2d and Table S2). Compared to the previous chromosome-level assembly of earthworm, *A. aspergillum*, the first T2T earthworm genome showed significant improvements in genome accuracy and continuity (Table 1).

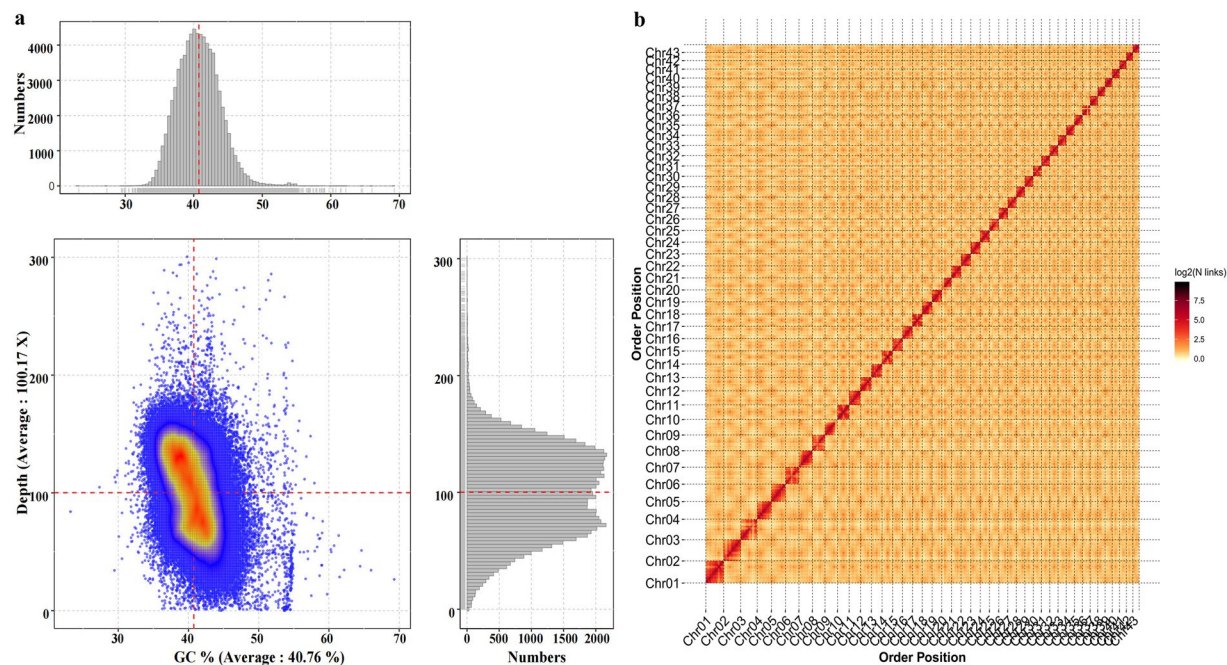


Fig. 3 Completeness assessment of the *A. aspergillum* genome. **(a)** GC content and sequencing depth distribution density plot. The horizontal coordinate indicates the GC content and the vertical coordinate indicates the coverage depth. The right side shows the distribution of contig coverage depth, and the top side shows the distribution of GC content. The large figure in the center shows the scatter plot based on the GC distribution of contigs and the coverage depth information, in which the color shades are used to reflect the density of the points in the scatter plot. **(b)** The Hi-C interaction heatmap illustrates the quality of the Hi-C assembly and the interaction frequencies among 43 chromosomes in *A. aspergillum*. Chromosomes are represented by the squares. The strength of the interaction is defined by the color from yellow (low) to red (high).

Repeat element annotation. Transposon elements (TEs) and tandem repeats were annotated by the following methods. TEs were identified by combining homology-based and *de novo* approaches²⁶. We first customized a *de novo* repeat library of the genome using RepeatModeler²⁷ v2.0.1 (BuildDatabase -name & RepeatModeler -pa 12), which can automatically execute two *de novo* repeat finding programs, including RECON²⁸ v1.0.8 and RepeatScout²⁹ v1.0.6. The long terminal repeat retrotransposons (LTR-RTs) were identified using both LTRharvest³⁰ v1.5.10 and LTR_FINDER³¹ v1.07. The high-quality intact LTR-RTs and non-redundant LTR library were produced by LTR_retriever³² V2.9.0 with default parameters and used to examine the insertion time of LTR-RTs. Subsequently, a non-redundant species-specific TE library was developed by integrating the *de novo* TE sequences library above with the known Dfam v3.5 database. The final TE sequences in the *A. aspergillum* genome were identified and classified through a homology search against this library using RepeatMasker v4.1.0³³ (repeat masker -nolow -no_is -norna -engine wublast -parallel 8 -qq). Moreover, we also identified tandem repeats, including microsatellites, minisatellites, and satellites, using the MICRoSATellite identification tool (MISA)³⁴ v2.1 and Tandem Repeat Finder (TRF)³⁵ v409 (2 7 7 80 10 50 500 -d -h). As a result, approximately 49.33% (374,336,823 bp) of the *A. aspergillum* genome sequences were identified as repetitive, of which 39.49% were transposable elements (TEs) and 9.84% were tandem repeats. Among these, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeat retrotransposons (LTR-RTs) accounted for 7.76%, 0.32%, and 11.44% of the genome, respectively (Table 2).

Gene prediction and genome functional annotation. In addition to the repeat sequences, we predicted 35,723 protein-coding genes from the repeat-masked genome of *A. aspergillum* through a combined strategy of *de novo*, homologous, and RNA-sequencing-based predictions (Table 1 and S3). Specifically, *de novo* prediction was performed by using Augustus³⁶ v3.1.0 and SNAP³⁷ (2006-07-28) with default parameters. For homology-based prediction, we utilized Miniport v1.7 (run. sh mmseqs) to determine a comparative analysis of the sequences from model organisms and closely related species, including *Caenorhabditis elegans*³⁸, *A. corticis*, *E. andrei*, and *L. terrestris*. These sequences were downloaded from the National Center for Biotechnology Information (NCBI) database and compared to the *A. aspergillum* genome to predict gene structure according to homology-based evidence. Moreover, we extracted total RNA and generated RNA reads with a total of 10.21 Gb of clean data from the body tissue of *A. aspergillum* (Table S4). Then GeneMarkS-T³⁹ v5.1 and PASA⁴⁰ v2.4.1 with default parameters were used for transcriptome-based prediction with the RNA-seq clean data. Finally, the prediction results obtained from the above three methods were incorporated using EvidenceModeler (EVM)⁴¹ v1.1.1 with default parameters and modified using PASA⁴⁰ v2.4.1 to generate the final coding gene set. In contrast, 3,697

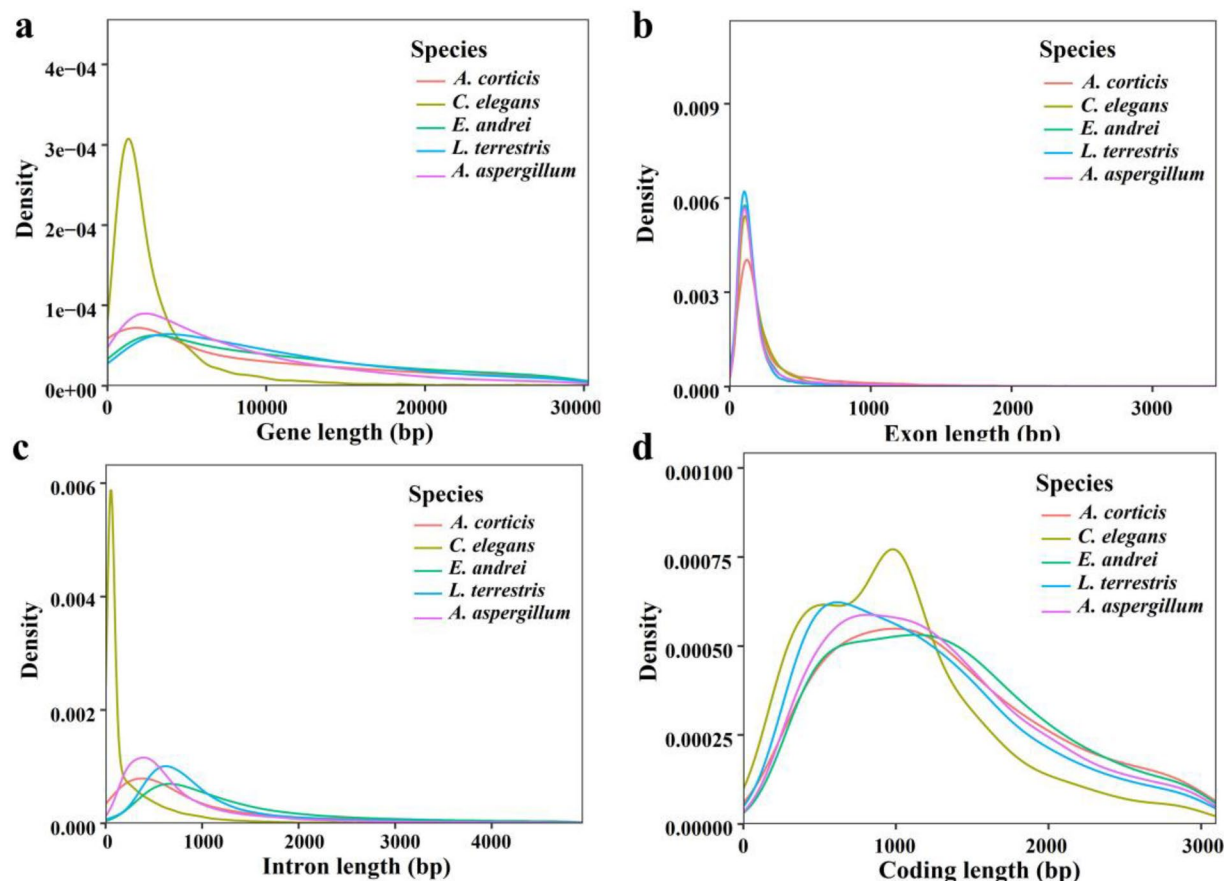


Fig. 4 Comparison of the distribution of annotated gene features in *A. aspergillum* with four closely related species. (a) Gene length; (b) Exon length; (c) Intron length; (d) Coding length. The horizontal axis represents the length and the vertical axis represents the gene density.

noncoding RNAs, 138 pseudogenes whose biological functions were lost, 1,959 conserved motifs, and 65,549 domains were identified based on the respective annotation method (Table S5).

After gene prediction, we conducted gene functional annotation by aligning the protein-coding gene sequences obtained from the preceding methods against the Non-Redundant (NR)⁴², EggNOG⁴³, TrEMBL⁴⁴, KOG, SWISS-PROT⁴⁴ and Pfam⁴⁵ protein databases using diamond v0.9.29.130 (diamond blastp—masking 0 -e 0.001) and the Kyoto Encyclopedia of Genes and Genomes (KEGG)⁴⁶ database (<http://www.genome.jp/kegg/>) with an E-value threshold of 1E-3. Gene Ontology (GO)⁴⁷ IDs (<http://www.geneontology.org/>) for each gene were obtained from TrEMBL⁴⁴, InterPro⁴⁸, and EggNOG⁴³. A total of 31,657 protein-coding genes were annotated, accounting for 88.62% of all predicted genes in *A. aspergillum*. The specific functional annotation statistics are presented in Table S3. In the eggNOG function classification, the unknown function group (S) accounted for the largest proportion, reaching approximately 23.88% (Fig. S3).

Data Records

All sequencing raw data have been deposited into the NCBI Sequence Read Archive (SRA) database at SRR316 56544-SRR31656548^{49–53}. In addition, the T2T Genome data have been deposited at the NCBI database under the accession JBJUSN000000000⁵⁴, and the genome annotation files have been submitted to Figshare dataset⁵⁵.

Technical Validation

Completeness assessment of the assembled T2T genome. To evaluate the completeness of the *A. aspergillum* genome assembly, we utilized Benchmarking Universal Single-Copy Orthologs (BUSCO)⁵⁶ v5.2.2 (busco -m genome -c 24 -e 1e-3—Augustus) with the OrthoDB10 database to identify complete BUSCOs in the assembly. The BUSCO assessment identified a total of 954 BUSCO genes, of which 908 (95.18%) were completely captured, only six genes (0.63%) were fragmented, and 40 (4.19%) were missing from the genome, indicating the high integrity of the T2T genome assembly (Table 1 and S6). The percentage of complete BUSCOs was greater than that of *A. corticis* (91.2%) and *M. vulgaris* (94.3%) from Megascolecidae^{16,17}, indicating that the integrity of the T2T assembly was higher than that of the chromosome-level assembly (Table 1). Additionally, bwa⁵⁷ v0.7.10 (bwa index & bwa mem -t 16) and Minimap2⁵⁸ v2.24-r1122 (-I 20 G—MD -ax map-hifi/-I 20 G—MD -ax map-ont) software was used for aligning the Illumina short reads, CCS and ultra-long ONT reads to the assembled

genome. The mapping rates of the Illumina, CCS, and ultra-long ONT reads were 99.72%, 99.86%, and 98.04%, respectively (Table S7). The average depth and coverage are shown in Tables S1, S7, and the sequencing data were analyzed for GC content and sample contamination (Fig. 3a). Moreover, the consensus quality value score of 46.89 obtained from the K-mer-based Merquary analysis⁵⁹ (github: <https://github.com/marbl/merquary>), indicated high accuracy of the T2T genome.

To examine the quality of Hi-C assembly and the interaction frequencies among different chromosomes, the genome was isotropically cut into 100 kb bins, and then the number of Hi-C Read Pairs between any two bins was used as a signal of the interaction between the two bins to make a Hi-C heatmap⁶⁰. As shown in the Hi-C interaction heatmap, the strength of the correlation was higher at the diagonal position than at the non-diagonal position in each chromosome group (Fig. 3b).

To evaluate the gene prediction quality, accuracy, and reliability, we produced an annotated gene feature chart of the distribution of gene length, coding DNA sequence (CDS) length, exon length, and intron length in *A. aspergillum* with a model organism (*C. elegans*) and three closely related species (*A. corticis*, *E. andrei*, and *L. terrestris*). The consistent distribution among all closely related species further emphasized the ideal annotated gene dataset for *A. aspergillum* (Fig. 4). Complete orthologs for 95.60% of the conserved BUSCOs were identified, indicating high completeness of the predicted protein-coding genes.

Code availability

All bioinformatics tools and pipelines were conducted by the prescribed guidelines from the respective manufacturer. The versions and corresponding parameters of software used in the study were described in the methods section. No custom package code was used during the analysis.

Received: 17 January 2025; Accepted: 23 April 2025;

Published online: 02 June 2025

References

- Phillips, H. R. P. *et al.* Global distribution of earthworm diversity. *Science* **366**, 480–485 (2019).
- Cs, C. Earthworm species, a searchable database. *Opuscula Zoologica Instituti Zoosystematici et Oecologici Universitatis Budapestensis* **43**, 97–99 (2012).
- Toor, M. D. *et al.* Earthworms as Catalysts for Climate-Resilient Agriculture: Enhancing Food Security and Water Management in the Face of Climate Change. *Water, Air, & Soil Pollution* **235**, 779 (2024).
- Medina-Sauza, R. M. *et al.* Earthworms Building Up Soil Microbiota, a Review. *Frontiers in Environmental Science* **7**, (2019).
- Parolini, M., Ganzaroli, A. & Bacenetti, J. Earthworm as an alternative protein source in poultry and fish farming: Current applications and future perspectives. *Science of the Total Environment* **734**, 139460 (2020).
- Commission, C. P. Pharmacopoeia of the People's Republic of China 2020 (Chinese). Beijing: China Medical Science Press; (2020).
- Guan, S. *et al.* Research on the Historical Evolution of the Herbal Research of Earthworm and its Processing Method. *Asia-Pacific Traditional Medicine* **19**, 167–175 (2023).
- Ge, X. *et al.* DNA Sequencing to Identify Zoological Origin of Commercial Pheretima from Chinese Herbal Markets and Discussion on Its Herbal Textual Research. *Modern Chinese Medicine* **21**, 1206–1214 (2019).
- Zhang, J. *et al.* An intelligentized strategy for endogenous small molecules characterization and quality evaluation of earthworm from two geographic origins by ultra-high performance HILIC/QTOF MS(E) and Progenesis Q1. *Analytical and bioanalytical chemistry* **408**, 3881–3890 (2016).
- Yang, W. *et al.* Bioevaluation of Pheretima vulgaris Antithrombotic Extract, PvQ, and Isolation, Identification of Six Novel PvQ-Derived Fibrinolytic Proteases. *Molecules* **26**, 4946 (2021).
- Huang, C. *et al.* Anti-inflammatory activities of Guang-Pheretima extract in lipopolysaccharide-stimulated RAW 264.7 murine macrophages. *BMC complementary and alternative medicine* **18**, 46 (2018).
- Yang, J. *et al.* Earthworm extract attenuates silica-induced pulmonary fibrosis through Nrf2-dependent mechanisms. *Laboratory investigation; a journal of technical methods and pathology* **96**, 1279–1300 (2016).
- Nazeer, A. & Awadh, A. K. Earthworms Effect on Microbial Population and Soil Fertility as Well as Their Interaction with Agriculture Practices. *Sustainability* **14**, 7803 (2022).
- Bhambri, A. *et al.* Large scale changes in the transcriptome of Eisenia fetida during regeneration. *PloS one* **13**, e0204234 (2018).
- Shao, Y. *et al.* Genome and single-cell RNA-sequencing of the earthworm Eisenia andrei identifies cellular mechanisms underlying regeneration. *Nature communications* **11**, 2656 (2020).
- Jin, F. *et al.* High-quality genome assembly of Metaphire vulgaris. *PeerJ* **8**, 10313 (2020).
- Wang, X. *et al.* Amynthes corticis genome reveals molecular mechanisms behind global distribution. *Communications Biology* **4**, 135 (2021).
- Short, S., Etxabe, A. G., Robinson, A., Spurgeon, D. & Kille, P. The genome sequence of the red compost earthworm, Lumbricus rubellus (Hoffmeister, 1843). *Wellcome open research* **8**, 354 (2023).
- Blaxter, M. L., Spurgeon, D. & Kille, P. The genome sequence of the common earthworm, Lumbricus terrestris (Linnaeus, 1758). *Wellcome open research* **8**, 500 (2023).
- Zhai, J. *et al.* Advances in earthworm genomics: Based on whole genome and mitochondrial genome. *Biodiversity Science* **30**, 1–11 (2022).
- Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **162**, 687–688 (2015).
- Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)* **27**, 764–770 (2011).
- 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 (2020).
- Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology* **31**, 1119–1125 (2013).
- Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, gaa094 (2020).
- Chengzhe, Z. *et al.* The chromosome-scale genome assembly of Jasminum sambac var. unifoliatum provides insights into the formation of floral fragrance. *Horticultural Plant Journal* **9**, 1131–1148 (2023).
- Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457 (2020).

28. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269–1276 (2002).
29. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **Suppl 1**, i351–i358 (2005).
30. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
31. Xu, Z. & Wang, H. LTRFINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–W268 (2007).
32. Ou, S. & Jiang, N. LTRretriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant physiology* **176**, 1410–1422 (2018).
33. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* Chapter 4, 4.10.11–14.10.14 (2009).
34. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics (Oxford, England)* **33**, 2583–2585 (2017).
35. Behboudi, R., Nouri-Baygi, M. & Naghibzadeh, M. RPTRE: A rapid perfect tandem repeat finder tool for DNA sequences. *Bio Systems* **226**, 104869 (2023).
36. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)* **24**, 637–644 (2008).
37. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
38. Qadota, H. *et al.* A novel protein phosphatase is a binding partner for the protein kinase domains of UNC-89 (Obscurin) in *Caenorhabditis elegans*. *Molecular biology of the cell* **19**, 2424–2432 (2008).
39. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic acids research* **43**, e78 (2015).
40. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654–5666 (2003).
41. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
42. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61–D65 (2007).
43. 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 research* **47**, D309–D314 (2019).
44. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**, 365–370 (2003).
45. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic acids research* **34**, D247–D251 (2006).
46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457–D462 (2016).
47. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research* **47**, D419–D426 (2019).
48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236–1240 (2014).
49. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31656544> (2025).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31656545> (2025).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31656546> (2025).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31656547> (2025).
53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR31656548> (2025).
54. Peng, G. Q. *et al.* *Amyntas aspergillum* isolate GP-2024, whole genome shotgun sequencing project. NCBI GenBank <https://identifiers.org/ncbi/insdc:JBJSUN000000000> (2025).
55. Peng, G. Q. *et al.* The annotation file of the T2T genome assembly of earthworm (*Amyntas aspergillum*). *Figshare* <https://doi.org/10.6084/m9.figshare.28638413> (2025).
56. 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 (Oxford, England)* **31**, 3210–3212 (2015).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760 (2009).
58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094–3100 (2018).
59. 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).
60. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).

Acknowledgements

This work is supported by the Guangxi Science Fund for Distinguished Young Scholars (Grant No. 2022GXNSFFA035030), the National Natural Science Foundation of China (Grant No. 82473808), and the Guangxi Medical University Training Program for Distinguished Young Scholars. It was also supported by the Xinjiang safflower industry development fund, the Fundamental Research Funds for the Central Universities (2632024TD04), and the specialized research funds from the State Key Laboratory of Natural Medicines, China Pharmaceutical University (SKLNMZZ2024JS37). The authors are grateful to Dr. Yongji Huang from Minjiang University for performing the karyotype analysis and the bioinformatics analysis.

Author contributions

Yucheng Zhao, Neng Jiang, and Changhong Wei designed and coordinated the study. Neng Jiang, Zhiming Yan, Mingwei He, and Guangquan Peng collected and prepared the earthworm samples. Guangquan Peng and Yanghe Qin assembled the genome and wrote the manuscript, and Yucheng Zhao revised the article. All authors have reviewed 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/s41597-025-05058-w>.

Correspondence and requests for materials should be addressed to Y.Z., N.J. or C.W.

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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2025