



OPEN

DATA DESCRIPTOR

Haplotype-resolved genome sequences of a springtail species, *Folsomia candida* (Collembola: Isotomidae)

Zhihong Zhan^{1,2}, Hao Yang^{1,2}, Jianfeng Jin³ & Feng Zhang^{1,2}

Folsomia candida (Collembola: Isotomidae) is one of the most important indicator species for soil animal studies and is widely distributed worldwide. Here, we utilized PacBio HiFi long reads, Oxford Nanopore Technologies (ONT) ultralong reads, Hi-C, Illumina short reads, and ONT long-read transcriptome sequencing to assemble and annotate the genome of *Folsomia candida*. Two gapless, telomere-to-telomere, haplotype-resolved genomes (HapA and HapB) were successfully assembled, with genome sizes of 228.62 and 228.45 Mb, respectively. The scaffold N50 sizes are 41.45 Mb for both haplotypes. We manually curated and annotated each gene on HapA, providing comprehensive annotation information for this haplotype. Gene annotation revealed a total of 26,329 protein-coding genes in HapA and 26,768 in HapB. These haplotype-resolved *Folsomia candida* genomes provide a high-quality genomic resource that enables future studies on soil arthropods and springtail genomics.

Background & Summary

Folsomia candida (Collembola: Isotomidae) is a widely used model organism in soil ecotoxicology and environmental monitoring, owing to its global distribution, ease of laboratory cultivation, and established role in standardized soil toxicity assays^{1,2}. As a representative springtail species, *F. candida* has long served as an indicator animal of soil quality and the effects of environmental stressors^{3,4}. These features have made *F. candida* an important reference species in soil biology and environmental research.

The investigation of *F. candida* genomics began in 2017 with the first genome assembly, which comprised 221.7 Mb in 162 scaffolds, representing a significant milestone in springtail genomics⁵. Subsequently, several improved versions of assemblies have been published, enabling comparative analyses and providing useful foundations for ecological and evolutionary studies². However, existing assemblies still contained gaps and did not revolve haplotypes, limiting further utilization for high-resolution genomic applications.

To address these limitations and provide a more complete *F. candida* genomic resource, we integrated PacBio HiFi, ONT ultralong reads, Hi-C data, Illumina short reads, and ONT long-read transcriptome data to generate new high-quality, haplotype-resolved reference genomes. By combining these datasets with our previous Hi-C data, we successfully assembled haplotype-resolved, telomere-to-telomere (T2T), and gapless genomes (HapA and HapB). The haplotype-phased genomes consist of 7 chromosomes for each haplotype, with genome sizes of 228.62 Mb (HapA) and 228.45 Mb (HapB), representing scaffold N50 values of 41.45 Mb for both haplotypes. We manually curated and corrected gene annotations on HapA, providing comprehensive annotation information. This T2T genome assembly delivers complete genomic information for *F. candida* and establishes a high-quality genomic framework for advancing soil arthropod research and comparative genomics studies.

Methods

Sample collection. A single *F. candida* female was taken from the laboratory of Nanjing Agricultural University, Dermark strain (FCDK), and then reared in a special container with peat soil for parthenogenic

¹Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, 210095, China.

²State Key Laboratory of Agricultural and Forestry Biosecurity, College of Plant Protection, Nanjing Agricultural University, Nanjing, 210095, China. ³College of Life Science, Xinyang Normal University, Xinyang, 464000, China.

✉e-mail: fzhang@njau.edu.cn

Libraries	Insert sizes	Raw data (Gb)	Coverage ($\times$)	Source
PacBio*	15Kb	28.00	122	NCBI(SRR35197880)
ONT ultralong*	100Kb	14.30	63	NCBI (SRR35197879)
Hi-C	350bp	48.85	222	NCBI (SRR13435515)
RNA	350bp	6.80	—	NCBI (SRR13452035)
RNA-ONT*	5Kb	25.40	—	NCBI (SRR35197878)

Table 1. Statistics of the sequencing data applied for *Folsomia candida*, T2T assembly, asterisks represent the new data obtained for this study.

reproduction. Around 600 offspring, including the samples of PacBio HiFi, ONT ultralong, and ONT long-read transcriptome, were extracted from the single female reproductive population and washed via phosphate-buffered saline (PBS) solution for five minutes to eliminate any possible external pollutants. Specimens were then transferred into liquid nitrogen, frozen for at least 20 minutes, and kept at -80°C for temporary storage until sequencing.

Library construction and sequencing. Specimens' genomic DNA (gDNA) was extracted using the FastPure[®] Blood/Cell/Tissue/Bacteria DNA Isolation Mini Kit (Vazyme Biotech Co., Ltd, Nanjing, China). High molecular weight (HMW) gDNA was sheared into 15 kb fragments using the MegaruptorTM device (Diagenode, Liege, Belgium) and purified using AMPurePB Beads. A PacBio HiFi 15 kb library was prepared using the SMRTbell Express Template Prep Kit 2.0, and the resulting library was sequenced on the PacBio Sequel II platform. For ultralong sequencing, HMW gDNA was prepared without shearing, and libraries were constructed using the Oxford Nanopore Technologies (ONT) Ligation Sequencing Kit, followed by sequencing on the PromethION platform. ONT long-read transcriptome libraries were constructed using the Direct RNA Sequencing Kit and sequenced on the PromethION platform. All library constructions and sequencing were performed by Berry Genomics (Beijing, China). Consequently, we newly obtained 161.40 Gb of sequencing data, including 28.00 Gb ($123\times$) of PacBio HiFi reads, 14.30 Gb ($63\times$) of ONT ultralong reads, and 24.90 Gb of ONT long-read transcriptome data (Table 1).

Haplotype-resolved genome assembly. Raw PacBio HiFi reads were assembled into the primary assembly using Hifiasm v0.19.8⁶. The outcome HiFi data was then combined with ONT ultralong reads and Hi-C data using default parameters to separate each haplotype. Consequently, a total of 7 chromosomes were assembled for each haplotype, and N50 values were 41.45 Mb for each, respectively.

Hi-C refinement and polishing. We applied the Hi-C data for genome assembly. Raw Hi-C data were processed for quality control to remove duplicates with Chromap v0.2.5-r473⁷. The clean Hi-C data were then used to align the primary assembly for haplotype identification and division. By applying YaHS v1.2⁸ and Juicer v1.6.2⁹, contigs were anchored and oriented onto chromosomes successfully. We manually checked outcome results under Juicebox v1.11.08¹⁰, and corrected any potential assembly errors to generate the final Hi-C assembly (Fig. 1).

To ensure high-quality genome assembly, we applied a comprehensive contamination detection and removal pipeline targeting sequences from humans, bacteria, viruses, and plants. Primary contamination screening was performed using MMseq. 2 v11.28¹¹ with BLASTN-like searches against the UniVec database, while vector contamination was specifically investigated through BLASTN searches (BLAST + v2.11.029). Our stringent filtering criteria automatically removed sequences showing $>90\%$ similarity to known contaminants, while sequences with 80–90% similarity underwent additional verification through online BLASTN analysis against the NCBI nucleotide database. The final assembled genome was checked through NCBI's automatic contamination screening pipeline, which revealed no significant contaminant sequences Table 2.

The final genome assembly comprised two distinct gapless haplotypes, and we named them HapA (228.62 Mb) and HapB (228.45 Mb), respectively. Both haplotypes exhibited larger genome size and fewer contigs (Table 3), representing our high-quality T2T haplotype-resolved outcome.

Telomere and centromere detection. Telomeres for each haplotype were detected by identifying the animal telomeric repeat sequences (AATCCG/AAATTT/AAACCT). We performed the *F. candida* telomere identification by using the TeloExplorer model of quarTeT v1.2.5¹². A total of 14 telomeres (7 chromosomes) for each haplotype were eventually recognized (Fig. 3). For centromere detection, we utilized RepeatObserverV1¹³ on both haplotypes to identify centromeric tandem repeat clusters. Consequently, 7 centromeres were identified for each haplotype (Fig. 3).

Collinearity analysis. Genome collinearity analysis between our two haplotypes and the previous assembly genome FCDK was performed using a dual-validation approach. Initially, D-Genies¹⁴ was employed to provide a comprehensive whole-genome dotplot visualization, offering an intuitive overview of large-scale syntenic relationships and potential structural variations between assemblies (Fig. 4a,b). To further validate these findings and obtain more detailed syntenic block information, we performed complementary collinearity analysis between HapA and FCDK using MCScanX¹⁵, which utilizes gene-based collinearity detection algorithms. The results were subsequently visualized through SynVisio¹⁶ (Fig. 4c). Consequently, collinearity analysis between HapA and FCDK maintained the consistent result, and exhibited the previous assembly errors in FCDK (Fig. 4c). Both analysis approaches demonstrated consistent results, revealing remarkable collinearity between our haplotype-resolved assembly and the FCDK genome. Based on these unambiguous syntenic relationships, we further established a clear one-to-one correspondence between HapA chromosomes and the FCDK chromosome identifiers. This mapping is summarized in Table 4 and is provided to facilitate cross-reference between assemblies and ensure consistency with previous literature.

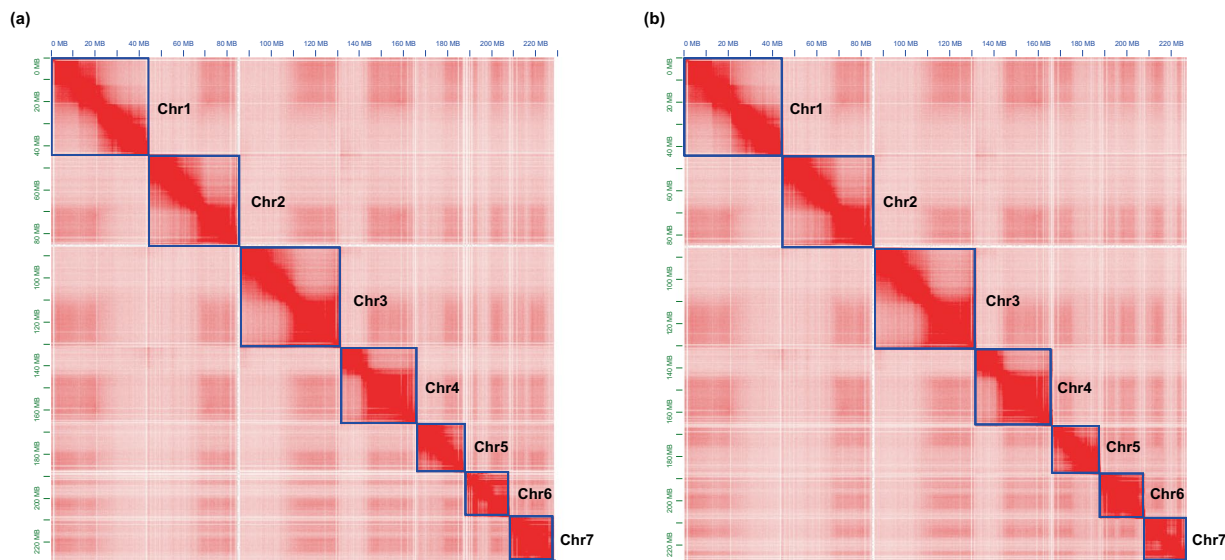


Fig. 1 Genome-wide Hi-C contact heatmaps of the *Folsomia candida* assemblies. The heatmaps display normalized Hi-C interaction frequencies, with higher contact intensities shown as darker colors. Each heatmap illustrates the chromosomal organization and interaction patterns supporting the chromosome-scale assembly. Chromosomes are framed in blue, and contigs (if present) are framed in green. **(a)** Hi-C contact map for HapA, showing seven well-defined chromosome interaction domains with strong intra-chromosomal contacts and minimal inter-chromosomal signal. **(b)** Hi-C contact map for HapB, similarly exhibiting seven chromosome-scale interaction domains consistent with a complete haplotype assembly.

HapA		HapB	FCDK	
Chromosome Number	Length (Mb)	Length (Mb)	Chromosome Number	Length (Mb)
1	44.61	44.63	3	43.32
2	41.45	41.45	4	38.47
3	45.87	45.70	1	44.87
4	34.62	34.61	7	32.24
5	21.62	21.62	5	20.37
6	20.34	20.34	6	18.19
7	20.08	20.08	2	19.19

Table 2. Chromosome lengths of *Folsomia candida*’s two haplotypes and the previous assembly genome, FCDK.

Gene prediction and annotation. Three strategies were combined for protein-coding genes (PCGs) annotation for *F. candida* HapA, including transcriptome-assisted prediction, homology-based proteins, and *ab initio* gene predictions. Transcribed RNA alignment prediction was performed by HISAT2 v2.2.1¹⁷, and relative production was then used as a genome-guided assembly by StringTie v2.1.6¹⁸. We used BRAKER v3.0.3¹⁹ to acquire the *ab initio* gene predictions with the combination of GeneMark-ETP²⁰ and Augustus v3.4.0²¹, and automatically trained based on RNA sequence alignments and reference proteins obtained from the OrthoDB v11 database. The GeMoMa v1.9²² was applied to protein-homology alignments from seven arthropod species, including *Drosophila melanogaster* (GCA_000001215.4²³), *Tribolium castaneum* (GCF_000002335.3²⁴), *Apis mellifera* (GCA_003254395.2²⁵), *Bombyx mori* (GCF_030269925.1²⁶), *Daphnia magna* (GCF_003990815.1²⁷), *Folsomia candida* (GCF_002217175.1⁵), and *Sinella curviseta* (GCA_004115045.4²⁸). Results from BRAKER and GeMoMa were finally combined and applied as the *ab initio* input for MAKER v3.01.03²⁹ to generate a GFF3 file.

To improve the accuracy of the annotation results, we manually curated each gene’s length and structure on all chromosomes using IGV-GSaman, integrating evidence from *ab initio* prediction, protein homology, and transcriptome data. The manual curation followed strict validation criteria: 1) A CDS was considered valid only when supported by at least two types of evidence among *ab initio* prediction, protein homology, and transcriptome data; 2) A CDS was required to have supporting evidence from either protein homology or transcriptome data (or both). Finally, 26,329 protein-coding genes were annotated on HapA, with the mean protein length of 475.5 aa. The mean gene length was 4,348.8 bp with an average of 6.8 exons per gene (Table 5). We applied the HapA annotation result as the reference, using LiftOn³⁰ to automatically annotate the HapB and get 26,768 protein-coding genes.

Repetitive sequence annotation. We performed the repetitive sequence annotation, including tandem repeats and interspersed repeats (transposable elements, TEs) for both haplotype genomes. A *de novo* specific repeat library was constructed for HapA and HapB by RepeatModeler v2.0.4³¹. This library was further combined

Name	HapA	HapB	FCDK
Genome assembly			
Size (Mb)	228.62	228.45	219.07
Number of contigs	7	7	75
Number of chromosomes	7	7	7
Scaffold N50 length (Mb)	41.45	41.45	38.47
GC (%)	37.71	37.71	37.48
BUSCO completeness (%)	97.5	97.5	97.6
Protein-coding genes			
Number	26,329	26,768	25,139
Mean gene length (bp)	475.5		470.2
Repetitive elements			
Size (%)	24.15	24.20	22.6
DNA transposons (%)	4.43	5.39	4.23
SINEs (%)	0.02	0.01	0
LINEs (%)	0.9	1.09	1.46
LTRs (%)	2.05	2.28	2.17
Unclassified (%)	14.92	13.69	13.34
ncRNAs			
rRNA	135	142	70
miRNA	33	32	34
snRNA	54	52	54
ribozyme	4	4	6
tRNA	113	122	115
lncRNA	2	2	2
Others	111	111	108
Total number of ncRNAs	452	465	396

Table 3. Genome assembly and annotation statistics for Haplotype-resolved genomes and the previous assembly genome, DK strain (FCDK).

with the RepBase-20230909³², and repeat elements were recognized and masked by RepeatMasker v.4.1.5³³ by aligning the custom library. For HapA, results demonstrated 55.22 Mb of repetitive sequence (24.15%), including DNA (4.43%), LINE (0.9%), LTR (2.05%), SINE (0.02%), and Unclassified elements (14.92%). For HapB, we found 55.29 Mb (24.20%) of repetitive sequence, comprising DNA (5.39%), LINE (1.09%), LTR (2.28%), SINE (0.01%), and Unclassified elements (13.69%). Both HapA and HapB exhibited similar repetitive element results, which reconfirmed our haplotype-resolved genomes' accuracy (Fig. 2).

Non-coding RNA annotation. Non-coding RNAs (ncRNAs) and transfer RNA (tRNA) of two haplotypes in *F. candida* were detected and identified by Infernal v1.1.4³⁴ and tRNAscan-SE v2.0.9³⁵. As a result, 452 ncRNAs were identified in HapA, including two long non-coding RNAs, 4 ribozymes, 33 microRNAs, 54 small nuclear RNAs, 111 other ncRNAs, 113 tRNAs, and 135 rRNAs. 465 ncRNAs were found in HapB, including two long non-coding RNAs, 4 ribozymes, 32 microRNAs, 52 small nuclear RNAs, 111 other ncRNAs, 122 tRNAs, and 142 rRNAs (Table 3).

Functional gene annotation. The functional gene annotation was performed on HapA by searching against the UniProtKB database (SwissProt and TrEMBL, version 20221130) using Diamond v2.0.11.1³⁶. Protein domain identification and functional annotation were conducted using eggNOG-mapper v2.1.9³⁷ and InterProScan 5.60–92.0³⁸ for Gene Ontology (GO) and KEGG pathway annotation analysis. Five databases, including Pfam³⁹, SMART⁴⁰, Superfamily⁴¹, Gene3D⁴², and CDD⁴³, were analyzed through InterProScan. The functional annotation results showed that 25,309 genes (96.13%) matched UniProtKB records, with 5,114 genes labeled as “uncharacterized protein” and 1,040 genes labeled as “unknown function”. Through the integration of InterProScan and eggNOG annotations, HapA contained 16,889 genes assigned to COG functional categories, 14,974 genes with GO annotations, 7,291 genes with KEGG pathway annotations, and 4,719 genes with Enzyme Codes (EC numbers) (Table 5).

Data Records

The raw sequencing data and genome assembly of *Folsomia candida* have been deposited at the National Center for Biotechnology Information (NCBI). The PacBio HiFi, ONT ultralong, and ONT long-read transcriptome can be found under identification numbers SRR35197880⁴⁴, SRR35197879⁴⁵, and SRR35197878⁴⁶, respectively, under the BioProject accession number PRJNA1040973 and BioSample accession number SAMN17109001. The Hi-C data of *Folsomia candida* DK can be found under the identification number SRR13435515⁴⁷ under the BioProject accession number PRJNA686204. The Illumina transcriptome short reads can be found under the identification number SRR13452035⁴⁸ under the BioProject PRJNA686207. The assembled haplotype-resolved

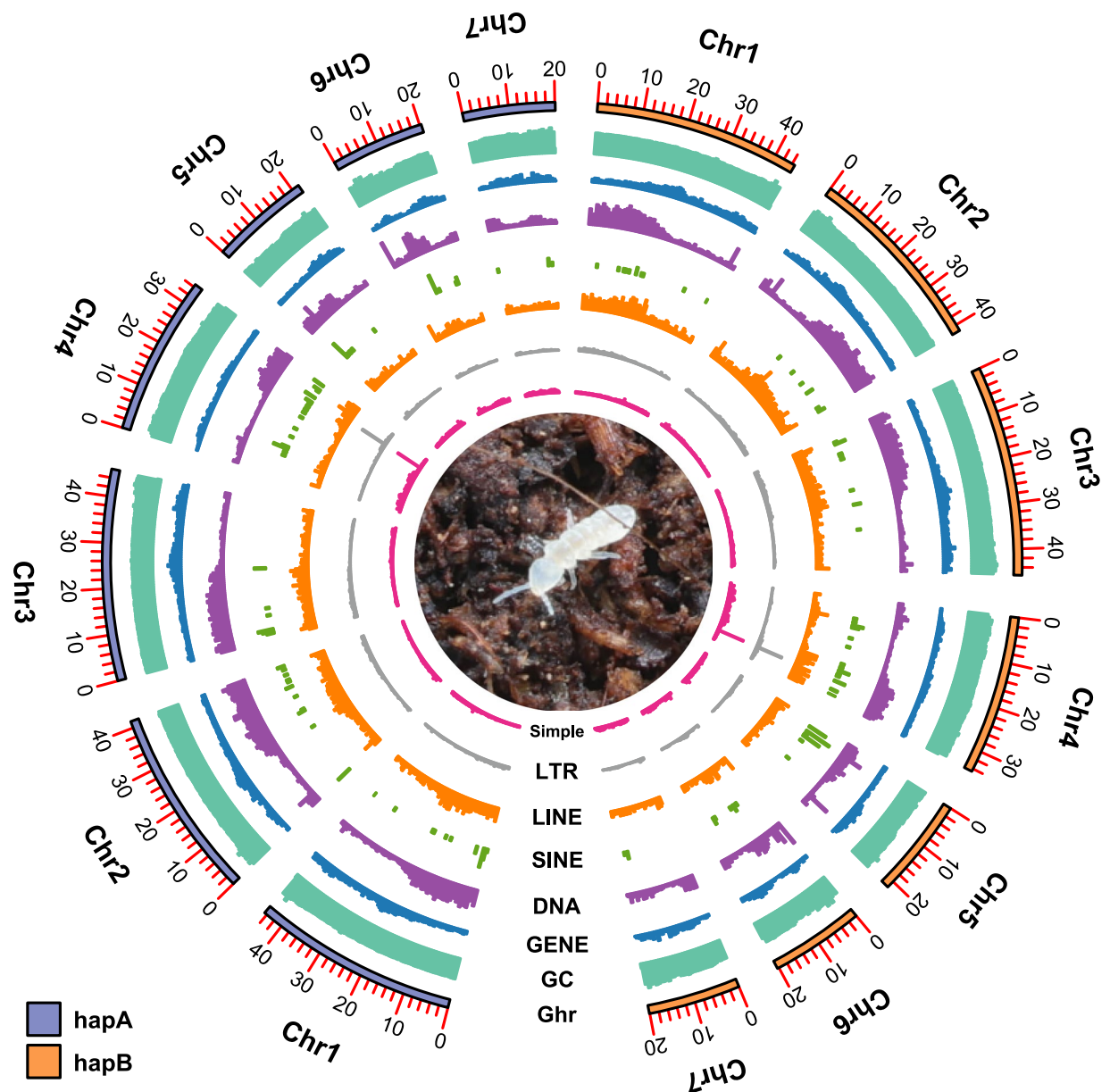


Fig. 2 Genome characteristics of *Folsomia candida*. Circos plot showing the genomic characters of *F. candida* from outer to inner: chromosome length (Chr) (Mb), the density of GC content (GC), the density of protein-coding genes (GENE), the density of TEs (DNA, SINE, LINE, and LTR), and simple repeats (Simple). Blue chromosomes represent HapA, and orange chromosomes represent HapB (The sliding window size is counted for every 10 kb).

genomes have been deposited in the GenBank in NCBI under accession numbers GCA_052721305.1⁴⁹ (HapA) and GCA_052721315.1⁵⁰ (HapB). The annotation results for repeated sequences, gene structure, and functional prediction have been deposited in the Figshare⁵¹ database.

Technical Validation

Two methods were used to evaluate the quality of the genome assembly. Firstly, BUSCO v5.4.4⁵² was applied for assembly completeness calculation with the reference Arthropoda gene set ($n = 1,013$) with the `euk_genome_met` mode. The final genome assembly showed a BUSCO completeness of 97.5% for both haplotypes, including 972 (96.0%) single-copy BUSCOs, 15 (1.5%) duplicated BUSCOs, 7 (0.7%) fragmented BUSCOs, and 19 (1.8%) missing BUSCOs. The final annotation validation of HapA was also calculated by BUSCOs with a protein mode with the reference Arthropoda gene set ($n = 1,013$). The final annotated HapA also exhibited a BUSCO completeness of 98.1%, including 833 (82.2%) single-copy BUSCOs, 161 (15.9%) duplicated BUSCOs, 7 (0.7%) fragmented BUSCOs, and 12 (1.2%) missing BUSCOs. To investigate the quality of the *de novo* assembly, Merqury v1.3⁵³ was performed to identify possible assembly sequence errors based on efficient k-mer set operations and QV score calculation. Consequently, the k-mer completeness value of the HapA is 99.03%, and the QV score is 55.57. For HapB, the k-mer completeness

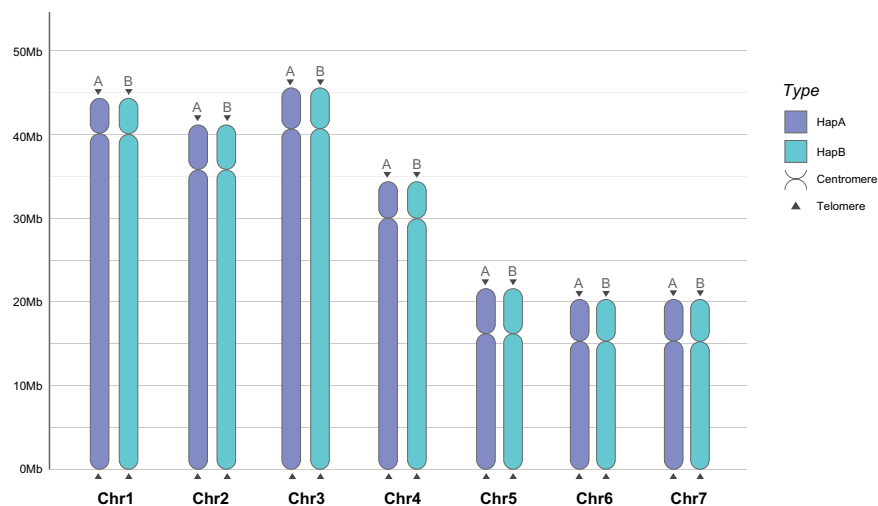


Fig. 3 Telomere and centromere detection of two haplotype-resolved genomes. Darker color chromosomes represent HapA, and brighter color chromosomes represent HapB. All telomeres were marked with triangle symbols, and all centromere positions were exhibited by the bar graph.

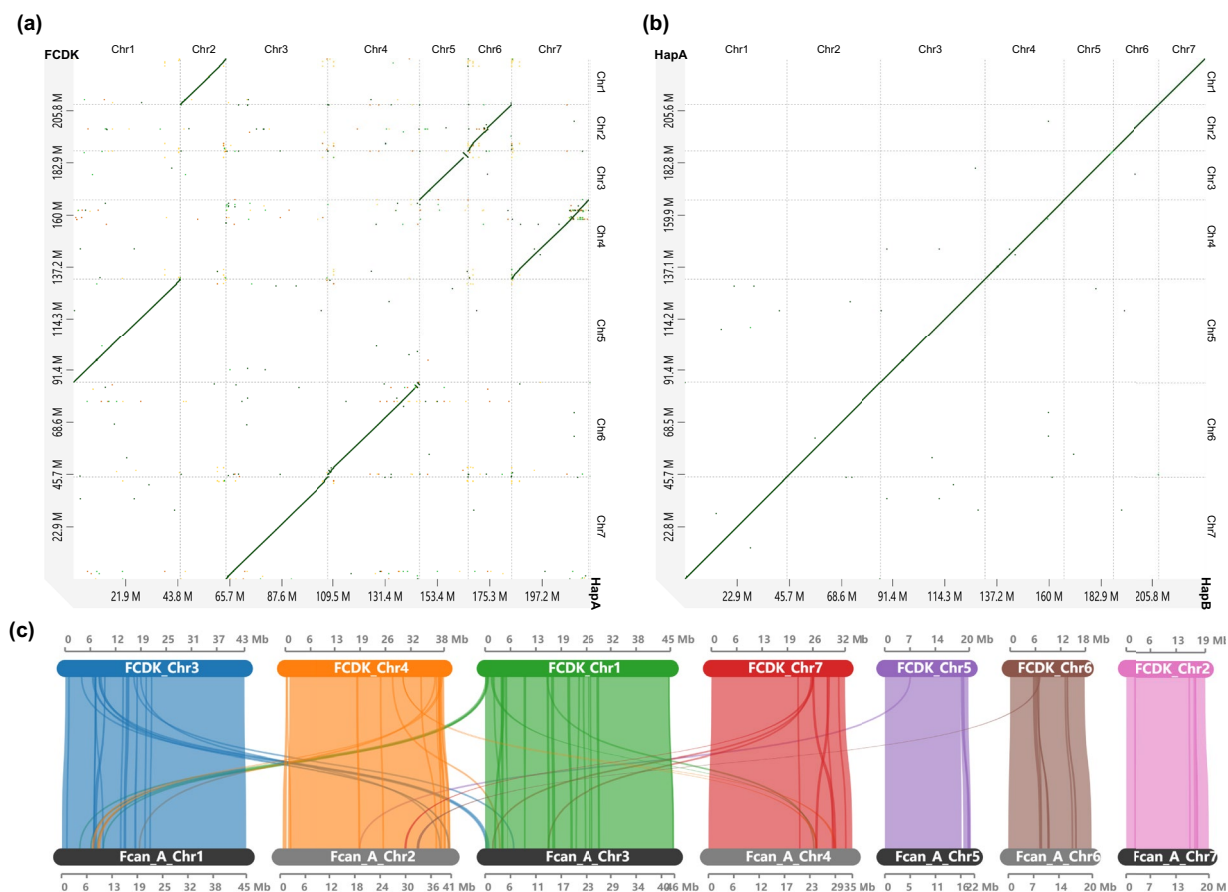


Fig. 4 Genome collinearity analysis of *Folsomia candida*. (a) Synteny test between HapA (right) and the previous genome version, FCDK (upper); (b) Synteny test between HapA (upper) and HapB (right); (c) Collinearity analysis between HapA (below) and FCDK (upper), genes with darker colors representing the assembly errors of FCDK.

value is 98.97% and the QV score is 54.67. Both the k-mer value and the QV score reflect the high accuracy of the base pairs, combined with the BUSCOs, which exhibit the high completeness and accuracy of our genome assembly.

Hap A	FCDK
Chr 1	Chr 3
Chr 2	Chr 4
Chr 3	Chr 1
Chr 4	Chr 7
Chr 5	Chr 5
Chr 6	Chr 6
Chr 7	Chr 2

Table 4. One-to-one correspondence between chromosomes in the HapA assembly and those in the FCDK reference genome. This mapping is provided to ensure consistency with previous literature and to facilitate cross-reference between assemblies.

Structure annotation	
Number of protein-coding genes	26,329
Number of predicted protein sequences	26,345
Mean protein length (aa)	475.5
Mean gene length (bp)	4,348.8
Gene ratio (%)	50.08
Number of exons per gene	6.8
Mean exon length (bp)	286.3
Exon ratio (%)	22.63
Number of CDSs per gene	6.5
Mean CDS length (bp)	219.0
CDS ratio (%)	16.47
Number of introns per gene	5.8
Mean intron length (bp)	406.9
Intron ratio (%)	27.45
Function annotation	
Number of genes matching Uniprot records	25,309
Number of genes labeled as “Uncharacterized protein”	5,114
Number of genes labeled as “unknown function”	1,040
Number of genes with InterProScan annotations	18,184
Number of genes with GO items from InterProScan annotations	10,533
Number of genes with KEGG pathway items from InterProScan annotations	0
Number of genes with eggNOG annotations	17,784
Number of genes with GO items from eggNOG annotations	12,165
Number of genes with Enzyme Codes (EC) from eggNOG annotations	4,719
Number of genes with KEGG ko terms from eggNOG annotations	11,365
Number of genes with KEGG pathway terms from eggNOG annotations	7,291
Number of genes with COG Functional Categories from eggNOG annotations	16,889
Number of genes with GO items (combining InterProScan and eggNOG results)	14,974

Table 5. Summary statistics of genome annotations in the HapA.

Data availability

All data used in this manuscript can be found under the following list: PacBio HiFi reads: SRR35197880⁴⁴, ONT-Ultralong reads: SRR35197879⁴⁵, ONT-long transcript reads: SRR35197878⁴⁶, Hi-C data: SRR13435515⁴⁷, Illumina short transcript reads: SRR13452035⁴⁸, HapA genome assembly: GCA_052721305.1⁴⁹, HapB genome assembly: GCA_052721315.1⁵⁰. Genome annotation information: <https://doi.org/10.6084/m9.figshare.30173248.v1>.

Code availability

No special scripts or code were developed during this work. All computational analyses were conducted using established bioinformatics software with versions and key parameters specified in the Methods section. Default parameters were used unless otherwise indicated.

Received: 29 September 2025; Accepted: 5 March 2026;

Published online: 28 March 2026

References

1. Fountain, M. T. & Hopkin, S. P. *Folsomia candida* (Collembola): a “standard” soil arthropod. *Annual Review of Entomology* **50**, 201–222 (2005).
2. Luan, X. *et al.* High-quality genomes reveal significant genetic divergence and cryptic speciation in the model organism *Folsomia candida* (Collembola). *Molecular Ecology Resources* **23**(1), 273–293 (2022).
3. Hopkins, T. *Biology of the springtails* (Insecta: Collembola). Oxford University Press (1997).
4. ISO (International Organization for Standardization). Soil quality—inhibition for reproduction of collembola (*Folsomia candida*) by soil contaminants (ISO Standard No. 11267:2014) (2014).
5. Faddeeva-Vakhrusheva, A. *et al.* Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics* **18**, 493 (2017).
6. 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).
7. Zhang, H. *et al.* Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Communications* **12**(1), 1–6 (2021).
8. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**(1), btac808 (2023).
9. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
10. Dudchenko, O. *et al.* *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
11. Steinegger, M. & Soding, J. MMseqs. 2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
12. Lin, Y. *et al.* QuarTeT: A telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**(8), uhad127 (2023).
13. Elphinstone, C. *et al.* RepeatObserver: tandem repeat visualization and centromere detection. *bioRxiv* (2023).
14. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
15. Wang, Y. *et al.* MCSScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
16. Bandi, V. & Gutwin, C. Interactive exploration of genomic conservation. In *Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020*. Canadian Human-Computer Communications Society, Waterloo, CAN (2020).
17. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
18. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
19. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *Nar Genom. Bioinform.* **3**, lqaa108 (2021).
20. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. *bioRxiv* **13**, 524024 (2023).
21. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
22. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *Bmc Bioinformatics* **19**, 189 (2018).
23. Hoskins, R. A. *et al.* The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res.* **25**, 445–458 (2015).
24. Herndon, N. *et al.* Enhanced genome assembly and a new official gene set for *Tribolium castaneum*. *BMC Genomics* **21**, 47 (2020).
25. Wallberg, A. *et al.* A hybrid *de novo* genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics* **20**(1), 275 (2019).
26. Kawamoto, M. *et al.* High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology* **107**, 53–62 (2019).
27. Lee, B. *et al.* The genome of the freshwater water flea *Daphnia magna*: A potential use for freshwater molecular ecotoxicology. *Aquatic Toxicology* **210**, 69–84 (2019).
28. Zhang, F. *et al.* A High-quality Draft Genome Assembly of *Sinella curviseta*: A Soil Model Organism (Collembola). *Genome Biology and Evolution* **11**(2), 521–530 (2019).
29. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics* **12**, 491 (2011).
30. Chao, K. H. *et al.* Combining DNA and protein alignments to improve genome annotation with LiftOn. *bioRxiv*, 2024.05.16.593026 (2024).
31. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
32. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. Dna.* **6**, 11 (2015).
33. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. Available online: <http://www.repeatmasker.org> (accessed on 14 October 2023) (2013–2015).
34. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
35. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol.* **1962**, 1–14 (2019).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
37. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
38. Finn, R. D. *et al.* InterPro in 2017—Beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
39. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
40. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
41. Wilson, D. *et al.* SUPERFAMILY—Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
42. Lewis, T. E. *et al.* Gene3D: Extensive Prediction of Globular Domains in Proteins. *Nucleic Acids Res.* **46**, D1282 (2018).
43. Marchler-Bauer, A. *et al.* CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
44. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR35197880> (2025).
45. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR35197879> (2025).
46. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR35197878> (2025).
47. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR13435515> (2025).
48. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR13452035> (2025).
49. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_052721305.1 (2025).
50. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_052721315.1 (2025).
51. Zhan, Z. Gene annotation for the *Folsomia candida* T2T genome hapA and hapB. *figshare* <https://doi.org/10.6084/m9.figshare.30173248.v1> (2025).

52. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
53. Rhie, A., Walenz, P. B., Koren, S. & Philippy, M. A. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

Author contributions

F.Z. contributed to the research design. Z.H.Z. and J.F.J collected and sequenced the sample. Z.H.Z. and F.Z. carried out the analysis, and Z.H.Z. wrote the manuscript. H.Y. revised the manuscript. All co-authors contributed to this manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.Z.

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) 2026