scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A high-quality chromosome-scale genome assembly of the wild sorghum (*Sorghum virgatum*)

Xiaojian Fang[1], Leina Zhou[1] & Zhongwei Lin[1,2] ✉

Sorghum virgatum **(Sv) is a wild subspecies of sorghum (*Sorghum bicolor* ssp. *Verticilliflorum* (L.) Moench). We employed Hi-C sequencing and HiFi technology in this study, successfully assembly a high-quality genome for Sv. This assembled genome is 795 Mb in total and scaffold N50 is 62.47 Mb. Within the Sv genome, 27,851 protein-coding genes were predicted. Repeat annotation detected 563.96 Mb of repetitive elements, making up 71.86% of the genome. This chromosome-level sv genome offers valuable insights to support comparative genomic studies within the *Poaceae* family, and it will aid genome-driven breeding and germplasm enhancement for modern sorghum.**

Background & Summary

Sorghum (*Sorghum bicolor* (L.) Moench) ranks among the world's five staple cereal crops[1]. It holds a key role in global food production and acts as a main calorie source for populations in the semi-arid tropics of Africa and Asia[2,3]. Sorghum comprises four wild ancestral subspecies: *aethiopicum*, *arundinaceum*, *verticilliflorum*, and *virgatum*[4,5]. Primarily distributed in Africa, the *arundinaceum* subspecies is characterized by its large and exposed inflorescence along with flexible branches without basal division. The *aethiopicum* subspecies features a relatively small and contracted panicle and it commonly distributed across the African Sahel. The *verticilliflorum* subspecies, native to Africa and Madagascar, is distinguished by a large, open inflorescence, with spreading branches that divide at the base. The *virgatum* subspecies, characterized by a narrow inflorescence and long, linear leaf blades, is common in northeastern Africa[6]. These four wild subspecies can interbreed with one another and also hybridize with modern cultivated sorghum[4].

Among sorghum's four wild subspecies, *Sorghum virgatum* (Sv) is marked by extremely narrow leaves (usually less than 2 cm in width) and slender inflorescences with short branches (Fig. 1a). This species is mainly distributed along riverbanks and irrigation channels in central Sudan. Sv shows strong stress tolerance, differs significantly in phenotype from cultivated sorghum, and can hybridize with cultivated varieties. This renders it a valuable resource for exploring key domestication genes. Previous research has used recombinant inbred line (RIL) populations from crosses between Sv and the cultivated sorghum Tx430 to successfully map the Sh1 gene linked to seed shattering[7]. A total of 24 QTLs for various traits were found in $F_{2:3}$ populations derived from the crosses between Sv and the cultivated variety Tx623[8]. Furthermore, the *AWN1* gene controls awn development, was cloned in the RIL population between Sv and Tx623[9]. Thus, Sv as a crucial genetic resource, making it necessary to analyze its genomic information.

In this study, we assembled a high-quality genome for Sv. By using HiFi sequencing and Hi-C technology, we successfully build a high-quality genome composed of 10 chromosomes (Fig. 1b). The assembled genome size is 795 Mb, and it scaffold N50 is 62.47 Mb (Supplementary Tables 1, 2). The BUSCO analysis showed 98.02% completeness, and the LAI index is 22.08, reached reference genome-level quality (Supplementary Table 3). About 71.86% of the genome was detected as repetitive sequences, including 54.99% LTR retrotransposons, 7.90% TIR elements and 0.77% LINE elements (Supplementary Table 4). By combining of RNA-seq data, homology-based annotation, and ab initio prediction, we found a total of 27,851 protein-coding genes (Supplementary Table 5). Comparative genomic analysis with *Sorghum bicolor* (v3) revealed syntenic regions and identified a total of 30,472 PAVs (presence/absence variants), 4,279,067 SNPs (single nucleotide polymorphisms) and 725,689 InDels (insertions/deletions) (Fig. 1c, and Supplementary Table 6). This high-quality wild sorghum Sv genome will act as a valuable resource for understanding the shift from wild to cultivated sorghum, providing a solid reference for future research.

[1]State Key Laboratory of Maize Bio-Breeding; National Maize Improvement Center, Department of Crop Genetics and Breeding, China Agricultural University, Beijing, 100193, China. [2]Sanya Institute of China Agricultural University, Sanya, Hainan, 572025, China. ✉e-mail: zlin@cau.edu.cn

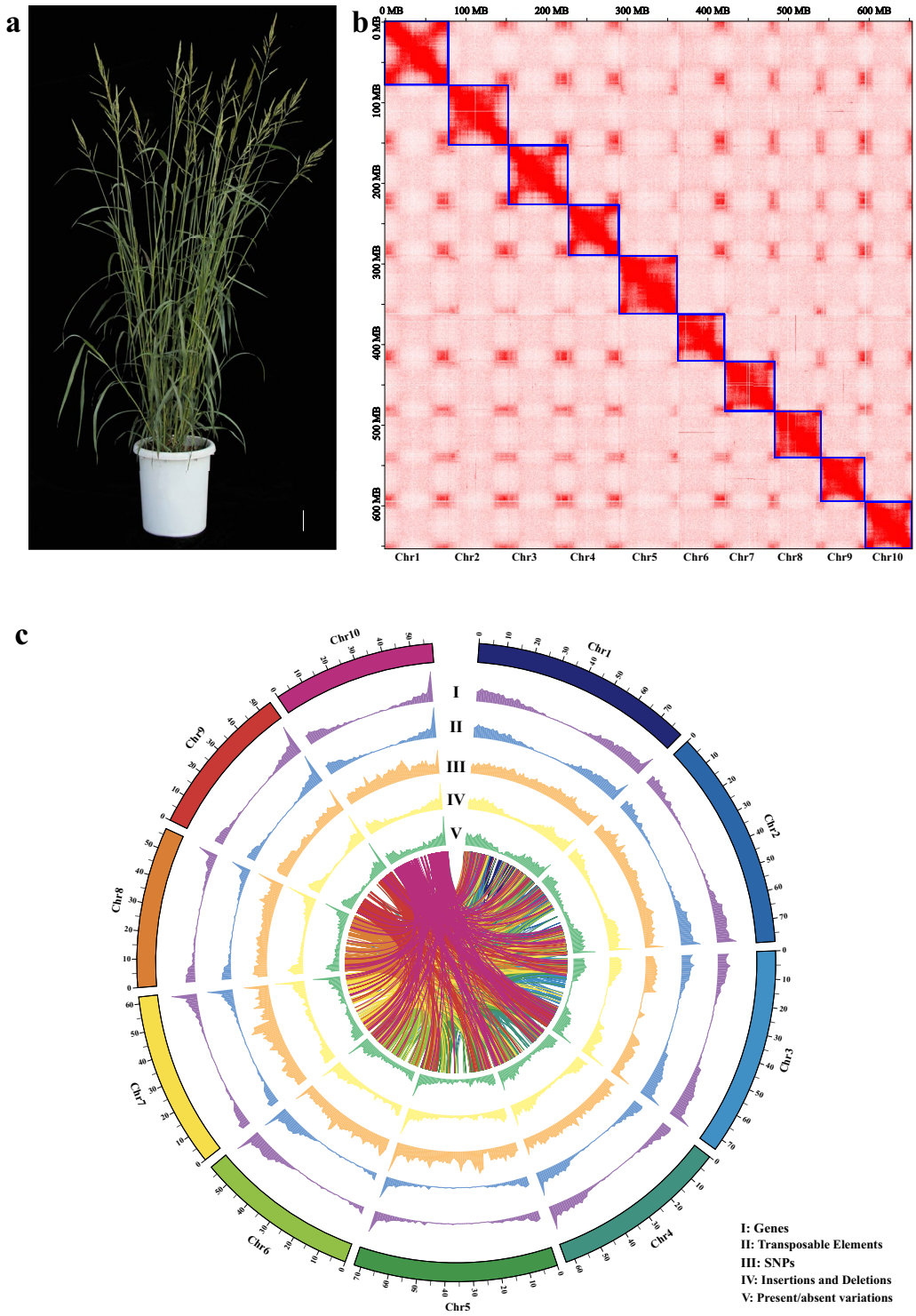

Fig. 1 Phenotypes of Sv and an overview of its genome assembly. (**a**) The phenotypes of Sorghum virgatum (Sv). Scale bar, 10 cm. (**b**) Hi-C contact heatmap for the assembled chromosomes. (**c**) Circular views of the Sv genome. The circles, from outer to inner, represent the densities of genes (I), transposable elements (II), SNPs (III), insertions and deletions (IV), and presence/absence variants (PAVs) (V).

Methods

Sv sequencing and assembly. High-quality genomic DNA was extracted from fresh Sv leaf tissue, processed for library preparation following PacBio protocols. Beijing Annuo Biotech conducted HiFi data using the PacBio Sequel II(v4.2.0) to generate circular consensus sequences. Genomic DNA extraction employed the CTAB method. For PacBio HiFi reads, a 15 Kb library was constructed with the SMRT bell Express Template Prep Kit (3.0). Library preparation included DNA shearing, damage repair, end repair, hairpin adapter ligation,

size selection. After quality control, the SMRT bell library was sequenced on a single SMRT Cell using the PacBio Revio platform. For Hi-C reads, the library was prepared from cross-linked chromatin extracted with standard Hi-C protocol. The library was sequenced on the DNBSEQ platform to produce 2 × 150 bp paired-end reads. Total RNA was taken from roots, leaves, stems, seeds, and tassels using TRIzol reagent in line with the standard procedure. For Illumina paired-end sequencing, mRNA was reverse-transcribed into cDNA, and five libraries with a 350 bp insert size were put together using the TruSeq RNA Library Preparation Kit (Illumina, USA) as per the manufacturer's guidelines. Sequencing was done on the NovaSeq 6000 Platform in PE150 mode. Genome assembly use the "-l 0" mode by hifiasm software (v0.16.1)[10]. Hi-C data were matched to the assembled genome by the juicer pipeline (v1.5.7), followed by genome scaffolding using 3D-DNA (v1.8) software, parameter is "-r 0"[11,12]. Assembly mistakes were manually corrected using Juicebox software (v1.11.08)[11]. The final chromosomal-level genome was built by 3D-DNA software[12]. The completeness and quality of the genome assembly were checked using BUSCO (v5.2.2), using "embryophyta_odb10" database[13]. Besides, the LTR Assembly Index (LAI) was calculated to assess assembly quality using LTR Harvest, based on set standards: $0 \leq LAI < 10$, Draft; $10 \leq LAI < 20$, Reference; $20 \leq LAI$, Gold[14].

Genome annotation. Transposable element annotation is using EDTA pipeline (v1.9.5) and repeat libraries used is from the Maize Transposon Element Consortium (MTEC; https://github.com/oushujun/MTEC)[15]. Following repeat masking of the Sv genome through RepeatMasker (v4.0.8) software, parameter is "-pa 30 -s -html -gff". The treated sequence was applied to subsequent protein-coding gene prediction. Gene model construction combined a multiple analytical framework. *De-novo* prediction via Augustus (v2.4) with standard parameters[16]; comparative genomic analysis against six representative plant species (*Sorghum bicolor*[17], *Arabidopsis thaliana*[18], *Oryza sativa*[19], *Setaria italica*[20], *Triticum aestivum*[21], and *Zea mays*[22]) by GenomeThreader (v1.7.1), parameters is "-intermediate -gff3out", with redundant sequences filtered by CD-HIT (v4.6) with standard parameters[23]; transcriptomic evidence from RNA-seq data encompassing five Sv tissues (roots, leaves, stems, seeds, tassels). Transcriptome assembly was performed with Trinity (v2.13.2), parameters is "--seqType fq --max_memory 50 G --CPU 6", alignment to the genome was executed with PASA (v2.0.2), using standard parameters[24,25]. Final gene model integration was combination with EvidenceModeler (v1.1.1) software, the parameters is "-weights ABINITIO_PREDICTION 1, TRANSCRIPT 8, PROTEIN 5", and then structural refinement through the PASA pipeline to produce consensus annotations[24,26].

SNPs and SV calling and genomic comparison. Single nucleotide polymorphisms(SNPs) and short insertions/deletions(Indels) (<50 bp) were identified using the MUMmer (v4.0) software[27,28]. Initial whole-genome alignments were created via NUCmer, parameters is "-mum -g 1000 -c 90 -l 40", then strict filtering using delta-filter, parameters is "-r -q", to establish a clear 1:1 genomic correspondence. Variant calling was conducted via show-snp, parameters is "-Clr", analysis on filtered alignment blocks. For detecting large-scale structural variations (100–100,000 bp), the assembly platform was used to process NUCmer outputs[29]. Gene collinearity patterns were identified using MCScanX (v1.63) with strict orthologous gene pairs[7]. Multidimensional genomic features (including gene counts, structural variants, SNPs, and indels) were visualized using Circos (v0.69-8) to plot genome architecture relationships (https://circos.ca/).

Data Records

Data supporting the findings of this work are available within the paper and its Supplementary Information files. The PacBio HiFi long-reads, Hi-C reads, and RNA-Seq datasets have been deposited in the National Genomics Data Center (NGDC; BioProject PRJCA035881[30]) under accession numbers CRA022781[31], CRA025558[32], CRA022779[33]. These data have also been uploaded to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with accession identifiers SRR33597079[34], SRR33613270[35] and SRR33585381[36]. The chromosome-scale genome assembly has been submitted to NGDC under accession GWHFJJD00000000.1[37] and has also been deposited in NCBI GenBank GCA_051168655.1[38]. Genome annotation files, including structural gene models and functional predictions, are publicly available through FigShare (https://doi.org/10.6084/m9.figshare.29063117.v1)[39].

Technical Validation

Assessment of the genome assembly. Genome integrity assessment was performed using BUSCO (v5.2.2) based on the "embryophyta_odb10" database[13]. Furthermore, we mapped DNA sequencing reads to the assembled genome using HiFiasm[10].

Assessment of the gene annotation. The annotation of the protein-coding genes (PCGs) demonstrated a high degree of completeness, as reflected by a BUSCO score of 98.02%, of which 95.66% were complete single-copy genes, and 2.35% were duplicated (Supplementary Table 3), indicating high-quality annotation of the predicted gene models.

To evaluate the assembly quality, the LTR Assembly Index (LAI) was calculated using LTR Harvest according to established criteria: Draft ($0 \leq LAI < 10$), Reference ($10 \leq LAI < 20$), and Gold ($20 \leq LAI$) classifications[14]. The LAI index reached 22.08 indicating a high-quality genome assembly.

Data availability

The dataset is available at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with accession identifiers SRR33597079, SRR33613270 and SRR33585381.

Code availability

All software and pipelines were executed according to the manuals and protocols of the published bioinformatic tools. All software used in this work are publicly available, with versions and parameters clearly described in Methods. No custom code was used during the compilation of the dataset.

Received: 17 February 2025; Accepted: 29 August 2025;
Published online: 30 September 2025

References

1. Morris, G. P. *et al*. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences* **110**, 453–458 (2013).
2. Mace, E. S. *et al*. A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers. *BMC plant biology* **9**, 1–14 (2009).
3. Mace, E. S. *et al*. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature communications* **4**, 2320 (2013).
4. Ananda, G. K. *et al*. Wild sorghum as a promising resource for crop improvement. *Frontiers in Plant Science* **11**, 1108 (2020).
5. Venkateswaran, K., Elangovan, M. & Sivaraj, N. in *Breeding Sorghum for diverse end uses* 15–31 (Elsevier, 2019).
6. Venkateswaran, K., Muraya, M., Dwivedi, S. & Upadhyaya, H. Wild sorghums—their potential use in crop improvement. (2014).
7. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49–e49 (2012).
8. Liu, H., Liu, H., Zhou, L. & Lin, Z. Genetic Architecture of domestication-and improvement-related traits using a population derived from Sorghum virgatum and Sorghum bicolor. *Plant Science* **283**, 135–146 (2019).
9. Zhou, L. *et al*. Gene duplication drove the loss of awn in sorghum. *Molecular plant* **14**, 1831–1845 (2021).
10. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175 (2021).
11. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
12. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
13. 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).
14. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic acids research* **46**, e126–e126 (2018).
15. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology* **20**, 1–18 (2019).
16. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435–W439 (2006).
17. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000003195.3 (2017).
18. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001735.4 (2018).
19. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_034140825.1 (2023).
20. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000263155.2 (2015).
21. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018294505.1 (2021).
22. *NCBI Genome*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_902167145.1 (2020).
23. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
24. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654–5666 (2003).
25. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644–652 (2011).
26. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22 (2008).
27. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944 (2018).
28. Song, B. *et al*. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proceedings of the National Academy of Sciences* **119**, e2113075119 (2022).
29. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
30. *NGDC Biological Project Library*. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA035881 (2025).
31. *NGDC Genome Sequence Archive*. https://ngdc.cncb.ac.cn/gsa/browse/CRA022781 (2025).
32. *NGDC Genome Sequence Archive*. https://ngdc.cncb.ac.cn/gsa/browse/CRA025558 (2025).
33. *NGDC Genome Sequence Archive*. https://ngdc.cncb.ac.cn/gsa/browse/CRA022779 (2025).
34. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRX28826677 (2025).
35. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRX28841617 (2025).
36. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRX28815137 (2025).
37. *NGDC Genome Warehouse*. https://ngdc.cncb.ac.cn/gwh/Assembly/88415/show (2025).
38. *NCBI GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_051168655.1 (2025).
39. Fang, X. Sorghum virgatum genome annotation. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.29063117.v1 (2025).

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFF1003400 to Z.L.).

Author contributions

Z.L. designed the study. X.F. and L.Z. performed the research. X.F. analyzed the data. X.F. and Z.L. wrote 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/s41597-025-05907-8.

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025