



OPEN

DATA DESCRIPTOR

A complete reference genome of broomcorn millet

Haigang Wang^{1,4}✉, Junjie Wang^{1,4}, Chunhai Chen^{2,4}, Ling Chen^{1,4}, Meng Li^{1,4}, Huibin Qin^{1,4}, Xiang Tian¹, Sen Hou¹, Xukui Yang², Jianbo Jian², Peng Gao³, Lun Wang¹✉, Zhijun Qiao¹✉ & Zhixin Mu¹✉

Broomcorn millet (*Panicum miliaceum* L.), known for its traits of drought resistance, adaptability to poor soil, short growth period, and high photosynthetic efficiency as a C₄ plant, represents one of the earliest domesticated crops globally. This study reports the telomere-to-telomere (T2T) gap-free reference genome for broomcorn millet (AJ8) using PacBio high-fidelity (HiFi) long reads, Oxford Nanopore long-read technologies and high-throughput chromosome conformation capture (Hi-C) sequencing data. The size of AJ8 genome was approximately 834.7 Mb, anchored onto 18 pseudo-chromosomes. Notably, 18 centromeres and 36 telomeres were obtained. The assembled genome showed high quality in terms of completeness (BUSCO score: 99.6%, QV: 61.7, LAI value: 20.4). In addition, 63,678 protein-coding genes and 433.8 Mb (~52.0%) repetitive sequences were identified. The complete reference genome for broomcorn millet provides a valuable resource for genetic studies and breeding of this important cereal crop.

Background & Summary

Broomcorn millet (*Panicum miliaceum* L.), a member of the Paniceae tribe in the Gramineae family, exhibits remarkable adaptability to marginal regions due to its short growing season (60–90 days), low water requirements, high salt tolerance, and efficient nutrient resource utilization^{1,2}. Being a C₄ plant, broomcorn millet demonstrates enhanced carbon fixation and efficient utilization of water and nitrogen resources. Additionally, its grains are characterized by their gluten-free nature and exceptional nutritional value, containing higher protein content, mineral composition, and antioxidant levels compared to most other cereals³. Consequently, broomcorn millet has been extensively cultivated in semiarid regions across Asia, Europe, and other continents and is considered one of the oldest crops worldwide⁴. The cultivation of broomcorn millet holds promise for enhancing food security, diversifying agriculture, and promoting a healthier diet⁵. Broomcorn millet has an allotetraploid genome consisting of 36 (2n = 4x = 36) chromosomes⁶. Although four chromosome-level of broomcorn millet, Jinshu⁷, LM_v1⁸, LM_v2⁹, and Pm_0390⁸, have been made available, there are still missing segments within the genome due to the presence of highly repetitive sequences clustered across the genome, particularly in the telomere and centromere regions. In recent years, T2T and gap-free genomes have been successfully obtained in various important crops, including rice¹⁰, barley¹¹, and maize¹².

In the present study, we assembled the first T2T gap-free genome of broomcorn millet (AJ8) (Fig. 1a), achieved through PacBio HiFi long reads, Nanopore technologies and Hi-C sequencing data. The resulting complete genome assembly has a final size of 834.7 Mb and is organized into 18 pseudo-chromosomes (Table 1; Fig. 1b). Gene annotation identified 52.0% repetitive sequences and 63,678 protein-coding genes (Fig. 1b). This complete reference genome provides a robust foundation for future studies on population and conservation genetics of broomcorn millet.

¹Center for Agricultural Genetic Resources Research, Shanxi Agricultural University, Key Laboratory of Crop Gene Resources and Germplasm Enhancement on Loess Plateau, Ministry of Agriculture and Rural Affairs, Taiyuan, 030031, China. ²BGI Genomics, Shenzhen, 518083, China. ³BGI, Shenzhen, 518083, China. ⁴These authors contributed equally: Haigang Wang, Junjie Wang, Chunhai Chen, Ling Chen, Meng Li, Huibin Qin. ✉e-mail: wanghg@sxau.edu.cn; wanglun976pzs@sina.com; nkyzps@126.com; muzx2008@sina.com

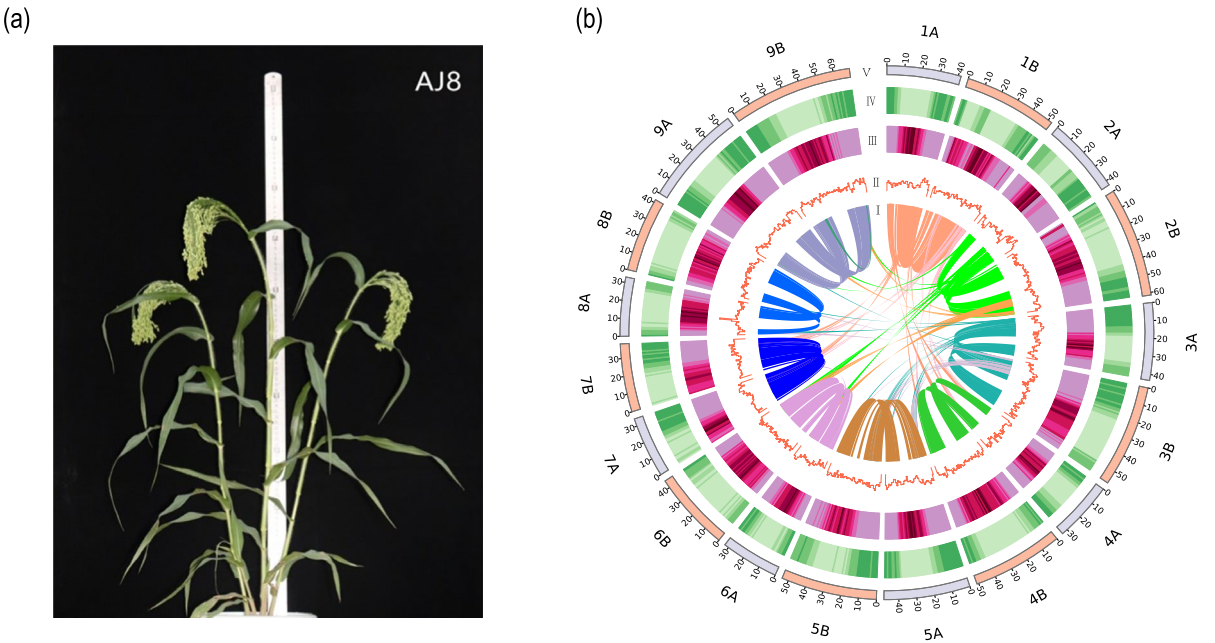


Fig. 1 An overview of the AJ8 genome. (a) The photograph of the AJ8 plant. (b) Circos plot illustrating the genome of the AJ8 genome. The plot includes the following components, arranged from inside to outside: (I) Collinear regions within the AJ8 assembly; (II) GC content in non-overlapping 1 Mb windows; (III) Percentage of repeats in 1-Mb sliding windows; (IV) Gene density in 1-Mb sliding windows; (V) Length of pseudo-chromosome in megabases (Mb).

Item	This study (AJ8)	Wang <i>et al.</i> ⁹ (LM_v2)	Sun <i>et al.</i> ⁷ (Jinshu7)	Zou <i>et al.</i> ³⁷ (Pm_0390)	Shi <i>et al.</i> ⁸ (LM_v1)
Size of assembly (Mb)	834.7	846.0	862.0	854.8	848.4
Number of contigs	18	704	4,062	5,538	1,308
Number of scaffolds	18	444	1,616	1,306	475
Contig N50 (Mb)	48.3	26.2	0.7	0.4	2.6
GC content	46.8%	46.8%	45.7%	45.9%	46.3%
Anchored pseudo-chromosomes (%)	100.0	99.0	93.3	96.2	98.9
Number of gaps in chromosomes	0	304	2,447	4,232	829
LTR assembly Index	20.4	20.8	11.1	9.3	15.9
Genome complete BUSCOs	99.6%	98.9%	98.7%	99.0%	98.8%
Quality value	61.7	—	—	—	—
Repetitive sequences	52.0%	57.9%	57.6%	58.2%	54.1%
Number of protein-coding genes	63,678	60,096	61,782	55,930	63,671

Table 1. Summary statistics of broomcorn millet genome assemblies.

Methods

Plant materials and growth conditions. The broomcorn millet landrace sequenced in this study was originally collected from Center for Agricultural Genetic Resources Research, Shanxi Agricultural University, Taiyuan, Shanxi Province (coordinates: E 112° 34′ 26.66″, N 37° 46′ 37.16″). The plants were planted under controlled conditions with a temperature of 25 °C, humidity of 60%, and a light intensity following a 14-hour day and 10-hour night cycle. Twenty seedlings with consistent growth at the fourth leaf stage were carefully chosen and sampled from various organs, including roots, stems, and leaves. A weight of 2 g was measured for each tissue organ, which was immediately placed in a freezing chamber with liquid nitrogen and subsequently stored at −80 °C.

Long insert libraries preparation and sequencing. Genomic DNA was extracted from leaf tissue using DNeasy Plant Maxi kit (Qiagen). The PacBio long insert libraries were prepared according to manufacturers’ instructions with an insert size of approximately 20 kb (Pacific Biosciences, USA). Subsequently, the libraries were subjected to sequencing using PacBio Sequel II platforms in circular consensus sequencing mode. The subreads were processed using SMRTLink (v11.1.0)¹³ with parameters “–minPasses 3 –minPredictedAccuracy 0.99 –minLength 500”, yielding approximately 77.0 Gb high-fidelity (HiFi) reads with a N50 size of about 18.0 kb (Table 2).

	Reads Number	Total length (bp)	Genome depth	N50 length of reads (bp)
Total ultra-long ONT data	5,657,711	168,590,000,000	196	55,765
Total error corrected ONT data	364,289	34,765,836,136	40	92,975
PacBio subreads	69,659,269	1,153,616,500,153	1338.3	18,495
PacBio CCS reads	4,321,122	77,005,593,171	89.3	17,995
Hi-C raw data	1,700,452,074	255,067,811,100	295.9	150
Hi-C clean data	1,686,533,526	252,980,028,900	293.5	150

Table 2. Summary of sequencing data of AJ8 genome.

Sample	Develoamental stage	Tissue	Total raw reads (Mb)	Total clean reads (Mb)	Total clean bases (Gb)	Clean reads Q20%	Clean reads GC%	Total MappingRatio
SL	seedling	leaf	60.30	57.99	8.70	98.29	57.23	99.33%
SS	seedling	stem	67.87	64.24	9.64	98.48	55.79	99.28%
SR	seedling	root	67.81	64.69	9.70	98.62	55.30	98.98%
TL	tillering	leaf	63.40	60.46	9.07	98.40	56.82	99.21%
TS	tillering	stem	69.89	66.85	10.03	98.53	55.04	99.26%
TR	tillering	root	72.38	67.93	10.19	98.71	54.32	92.81%
JL	jointing	leaf	56.91	54.30	8.14	98.45	56.73	99.33%
JS	jointing	stem	72.59	68.77	10.31	98.54	55.06	99.28%
JR	jointing	root	54.08	52.05	7.81	98.44	54.60	97.20%
BL	booting	leaf	72.88	69.47	10.42	98.43	57.06	99.24%
BS	booting	stem	58.16	55.33	8.30	98.42	55.94	99.28%
BR	booting	root	62.12	60.13	9.02	98.60	55.20	98.32%
BT	booting	tassel	61.17	58.46	8.77	98.53	55.11	99.21%
HL	heading	leaf	64.09	61.15	9.17	98.49	56.79	99.26%
HS	heading	stem	65.65	61.42	9.21	98.44	56.33	99.37%
HR	heading	root	64.52	61.84	9.28	98.59	54.38	97.12%
HT	heading	tassel	68.19	64.74	9.71	98.36	55.19	99.19%
FL	filling	leaf	62.43	59.36	8.90	98.46	55.76	99.34%
FS	filling	stem	60.22	58.35	8.75	98.48	54.05	99.07%
FR	filling	root	75.01	72.73	10.91	98.44	54.34	98.88%
FT	filling	tassel	57.27	54.80	8.22	98.12	53.80	99.19%

Table 3. Summary of RNAseq sequencing data of AJ8 genome.

The ONT ultra-long insert libraries were generated using the Oxford Nanopore SQK-LSK109 kit, and then sequenced on a PromethION flow cell (Oxford Nanopore Technologies, Oxford, UK). A total of 165.6 Gb of ONT data with 196x coverage was generated, and the N50 value was 55,765 bp (Table 2). After error correction and length filtering of the data, 34.8 Gb ultra long ONT reads with the N50 value 92, 975 bp were obtained (Table 2).

Short insert libraries preparation and sequencing. For chromosomal conformational capture (Hi-C) sequencing, Hi-C libraries based on *DpnII* restriction enzymes were generated as previously described¹⁴, and sequenced on the MGISEQ-2000 platform. A total of 253.0 Gb of clean data were obtained from 255.1 Gb of sequencing data using software SOAPnuke (v2.0)¹⁵ with parameters “-n 0.01 -l 20 -q 0.1 -i -Q 2 -G 2 -M 2 -A 0.5” (Table 2).

RNA-seq libraries from leaf tissues were constructed using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, Ipswich, MA, USA) following the manufacturer’s protocol. Then the RNA libraries were sequenced on a MGISEQ-2000 instrument and generated 150 bp paired-end reads. After quality control by fastp (0.19.5)¹⁶ with parameters of “-adapter_sequence AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA-adapter_sequence_r2 AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG-average_qual 15 -l 150”, each library contained more than 7.8 Gb of clean data. More than 98.1% of the clean data had scores greater than Q20 in each library (Table 3).

Genome assembly. Using HiFi reads, ultra-long ONT reads, and Hi-C clean data, the primary contigs were assembled by Hifiasm (v 0.19.5)¹⁷ with default parameters. To anchor contigs onto chromosomes, we used BWA (v 0.7.12)¹⁸ to align the Hi-C clean data to the assembled contigs. Low-quality reads were filtered out using the HiC-Pro pipeline¹⁹ with default parameters. The remaining valid reads were employed to anchor chromosomes with Juicer²⁰ and 3d-dna pipeline²¹. Excitingly, our results showed that the hifiasm assembly consists of contiguous sequences covering the entire length of all chromosomes. This achievement can be attributed to the remarkable

Type	Repbse TEs		TE protiens		De novo		*Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
Class I	209,289,045	25.07	108,867,424	13.04	347,247,904	41.60	379,892,897	45.51
LTR	199,720,006	23.93	102,637,122	12.30	341,764,732	40.95	367,377,405	44.01
Gypsy	167,578,773	20.08	83,383,157	9.99	303,670,978	36.38	321,394,835	38.51
Copia	31,228,868	3.74	19,222,359	2.30	32,348,058	3.88	44,339,339	5.31
LINE	9,720,644	1.16	6,231,875	0.75	5,688,415	0.68	13,406,696	1.61
SINE	151,212	0.02	0	0.00	39,826	0.00	186,663	0.02
Class II: DNA	30,371,375	3.64	6,938,097	0.83	13,865,556	1.66	38,069,361	4.56
†Unclassified	11,035	0.00	0	0.00	2,576,970	0.31	2,588,002	0.31
Total	239,065,404	28.64	115,803,473	13.87	362,426,030	43.42	414,237,138	49.63

Table 4. Interspersed repeat contents in AJ8 genome assembly. Note: This statistical table does not contain Tandem Repeats, some elements may partly include another element domain. *Combined: the non-redundant consensus of all repeat prediction/classification methods employed. †Unclassified: the predicted repeats that cannot be classified by RepeatMasker. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat.

Type		Gene number	Percentage
Total		63,678	100.00%
NR		61,980	97.33%
SwissProt		44,342	69.63%
KEGG		43,826	68.82%
KOG		42,801	67.21%
TrEMBL		61,821	97.08%
Interpro	All	60,143	94.45%
	GO	35,547	55.82%
Annotated		62,885	98.75%
Unannotated		793	1.25%

Table 5. Number of functional annotations for predicted genes in AJ8 assembly.

Primary ID	New ID after subgenome phaser
Chr01	9B
Chr02	2B
Chr03	3B
Chr04	9 A
Chr05	5B
Chr06	1B
Chr07	4B
Chr08	5 A
Chr09	6B
Chr10	3 A
Chr11	2 A
Chr12	1 A
Chr13	8B
Chr14	7B
Chr15	7 A
Chr16	4 A
Chr17	8 A
Chr18	6 A

Table 6. Correspondence between chromosome identifications of AJ8 and Jinshu7. *Consistency of AJ8 new ID with Jinshu7 Genome.

accuracy of HiFi data, the utilization of ultra-long ONT data, and the ongoing enhancements in assembly algorithms. Analogous to the T2T genome of rapeseed²², the hifiasm assembly comprises continuous sequences spanning the entirety of nine chromosomes. For further refinement of the genome, the T2T assembly was polished

Chromosomes	Length (bp)	Ends identified with telomeres	Total counts of telomere repeats at the left end	Total counts of telomere repeats at the right end
1 A	41,558,686	both	2,002	2,776
1B	52,342,684	both	2,317	2,037
2 A	42,791,896	both	578	504
2B	61,479,593	both	2,843	767
3 A	43,337,656	both	13	1,655
3B	57,374,001	both	2,860	2,883
4 A	33,404,080	both	1,063	1,428
4B	50,983,745	both	1,828	2,685
5 A	48,261,264	both	14	2,623
5B	53,707,716	both	3,011	16
6 A	32,475,619	both	2,582	2,535
6B	44,372,877	both	179	3,933
7 A	34,166,222	both	12	2,776
7B	38,174,571	both	13	788
8 A	33,047,006	both	1,857	3,687
8B	41,728,197	both	2,085	2,425
9 A	56,346,356	both	1,265	1,706
9B	69,126,039	both	1,563	1,747

Table 7. The identified telomeres in AJ8 assembly.

Chromosomes	Length (bp)	Start	End	Centro Length (bp)	TR percentage
1 A	18,342,834	21,040,163	2,697,330	1,583,268	58.70%
1B	20,653,421	26,624,068	5,970,648	1,298,067	21.74%
2 A	18,772,875	21,884,089	3,111,215	836,409	26.88%
2B	29,257,159	32,333,321	3,076,163	808,436	26.28%
3 A	25,401,874	27,185,391	1,783,518	878,557	49.26%
3B	31,759,679	35,045,521	3,285,843	985,466	29.99%
4 A	16,454,400	18,471,716	2,017,317	1,152,860	57.15%
4B	23,164,795	29,986,713	6,821,919	2,358,195	34.57%
5 A	25,993,962	30,048,990	4,055,029	1,449,363	35.74%
5B	29,311,126	31,111,890	1,800,765	378,448	21.02%
6 A	12,733,711	14,619,504	1,885,794	1,351,381	71.66%
6B	15,559,124	17,754,000	2,194,877	775,006	35.31%
7 A	8,527,110	11,079,922	2,552,813	969,677	37.98%
7B	10,357,744	11,349,473	991,730	190,239	19.18%
8 A	13,176,836	15,335,732	2,158,897	1,114,728	51.63%
8B	19,654,938	22,202,119	2,547,182	484,580	19.02%
9 A	27,161,481	30,534,321	3,372,841	1,953,960	57.93%
9B	28,766,226	36,231,841	7,465,616	910,724	12.20%

Table 8. The distribution of centromeres in AJ8 assembly.

using a similar method described by Mc Cartney *et al.*²³. In brief, the HiFi reads were aligned to the T2T assembly using Winnowmap2 (v 2.03)²⁴. The resulting alignments were filtered to exclude secondary alignments and alignments with excessive clipping by using ‘falconc bam-filter-clipped’ tool. Finally, racon (v 1.5.0)²⁵ was performed with the filtered alignments. The final assembled genome had a length of 834,678,208 bp and a contig N50 of 48.3 Mb (Table 1). The assembled sequences were successfully anchored to 18 pseudo-chromosomes (Table 1). The completeness of the assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (v 5.5.0)²⁶ with the embryophyta_odb10 (parameters: -m genome -l embryophyta_odb10). We applied Merqury (v1.3)²⁷ using PacBio HiFi long reads with a *K*-mer value of 17-bp to estimate the quality value.

Annotation of repetitive sequences. Tandem repeats and interspersed repeats were identified using the method described in Qu *et al.*²⁸. In brief, RepeatModeler (v1.0.4)²⁹ and LTR-FINDER (v1.0.7)³⁰ were employed to construct of a *de novo* repeat sequence library. This compiled library was subsequently deployed for the detection of interspersed repeats and low-complexity sequences utilizing RepeatMasker (v4.0.7)³¹. DNA and protein transposable elements (TEs) were identified through RepeatMasker (v4.0.7) and RepeatProteinMasker (v4.0.7), respectively. Tandem repeats were discerned utilizing Tandem Repeat Finder (v4.10.0)³². A total of 433.8 Mb

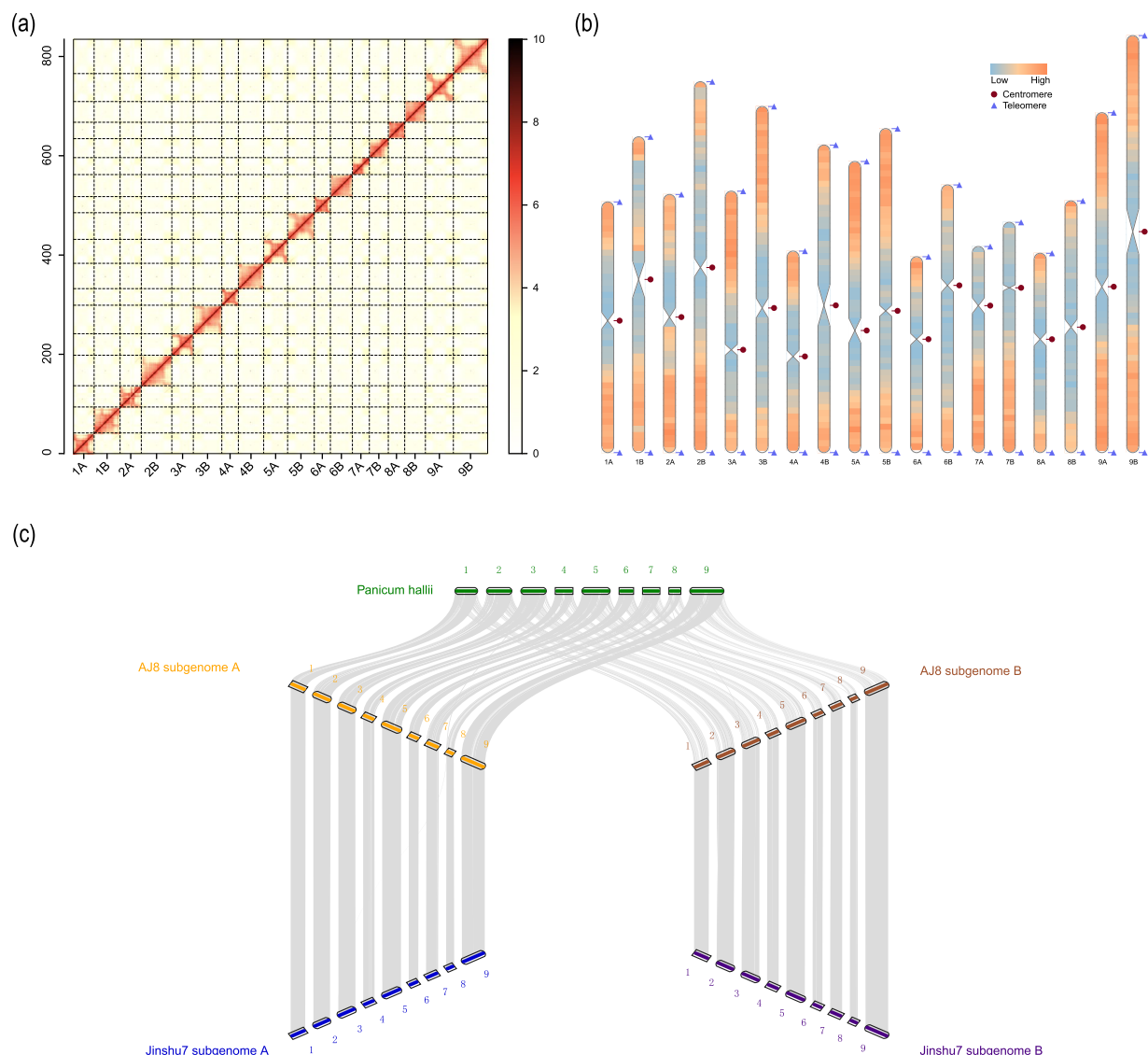


Fig. 2 The high-quality of the AJ8 genome. **(a)** Heatmap displaying Hi-C interactions of AJ8 pseudomolecules. **(b)** Telomere detection map. Triangles and circles represent telomeres and centromere within the AJ8 assembled chromosomes. The orange color represents regions with high gene density, while the dark sky blue color represents regions with low gene density. **(c)** Synteny analysis of *Panicum hallii*, AJ8 and Jinshu7.

(~52.0%) of repetitive sequences were obtained. Among the interspersed repeats, three types of repetitive elements, namely class I (retrotransposons), class II (DNA transposons), and unclassified elements, accounted for 49.6% of the genome assembly (Table 4). The telomeric sequences and centromere region in the AJ8 genome assembly were identified using quartet (v1.1.36)³³ with “-c plant”.

Protein-coding genes prediction and functional annotation. Gene prediction in this study was conducted using a combination of transcriptome-based prediction, homology-based prediction, and *ab initio* prediction methods. For transcriptome-based prediction, 194.2 Gb Illumina clean reads from 21 samples were assembled by Trinity (v 2.15.1)³⁴ with parameters of ‘-max_memory 200 G-CPU 40-min_contig_length 200-genome_guided_bam merged_sorted.bam-full_cleanup-min_kmer_cov 4-min_glue 4-bfly_opts ‘-V 5-edge-thr = 0.1-stderr’-genome_guided_max_intron 10000’, which generated 227,501 transcripts with a N50 of 2,698. These assembled transcripts were aligned against the AJ8 T2T assembly using Program to Assemble Spliced Alignment (PASA) (v 2.4.1)³⁵. Gene structures were generated from valid transcript alignments (PASA-set). RNA-seq clean reads were also mapped to the AJ8 T2T assembly using Hisat2 (v 2.0.1)³⁶. Stringtie (v 1.2.2)³⁷ and TransDecoder (v 5.7.1) (<https://github.com/TransDecoder/TransDecoder>) were employed to assemble the transcripts and identify candidate coding regions into gene models (Stringtie-set). Homologous genomes from seven plants, including rice (T2T-NIP)¹⁰, maize (T2T-Mo17)¹², *A. thaliana* (Col-PEK)³⁸, sorghum (GCA_000003195.3)³⁹, pearl millet (Tift23D2B1-P1-P5)⁴⁰, foxtail millet (GCA_000263155.2)⁴¹, and two previous versions of broomcorn millet (Jinshu7: GCA_026771285.1; Longmi4: GCA_002895445.3)^{7,9} were downloaded

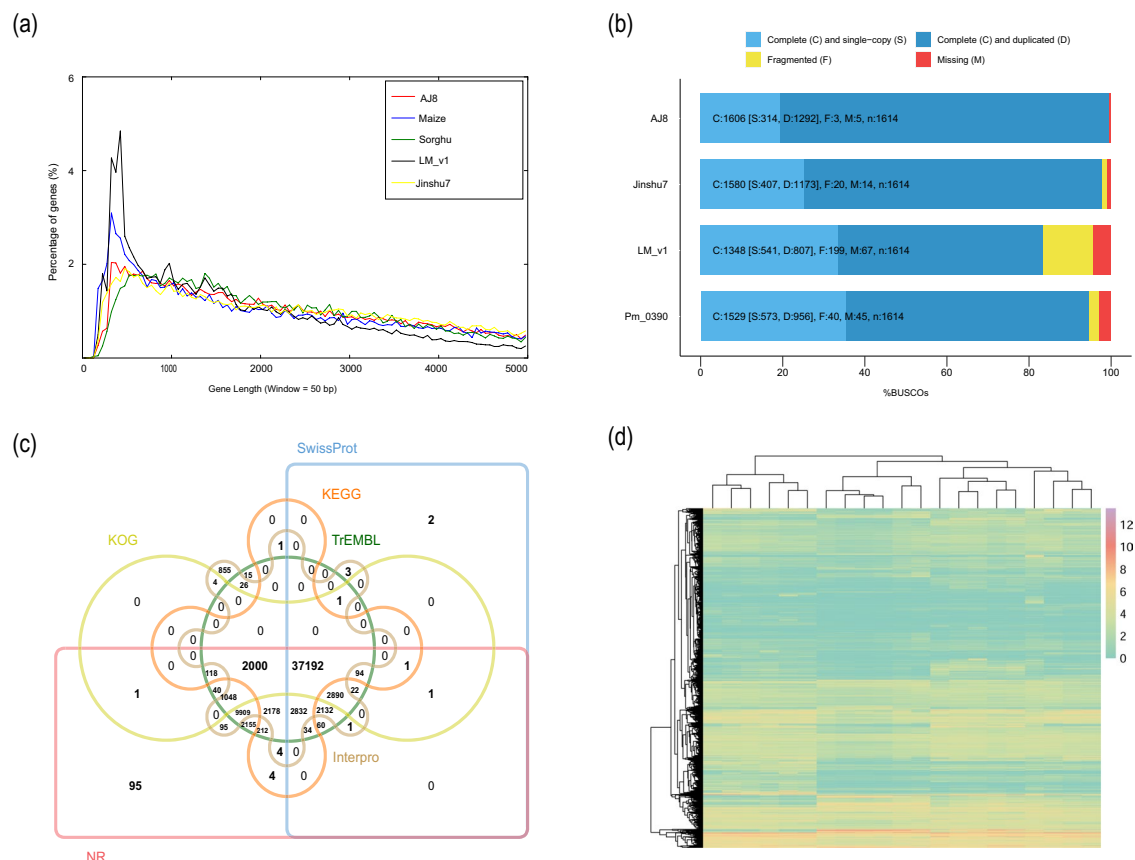


Fig. 3 Gene prediction quality assessment. **(a)** The composition of gene length in the AJ8 genome compared to the genomes of other species. **(b)** BUSCO assessments of the AJ8 gene. **(c)** Venn diagram showing the number of genes with homology or functional classification by each method. **(d)** The expression heatmap illustrates the expression levels among 21 RNA-seq samples. The color bar in the lower right corner represents log2-transformed FPKM values. Blue and red boxes indicate genes with lower and higher expression levels, respectively.

and used as queries to search against the AJ8 T2T assembly using GeMoMa (v 1.9)⁴². These homology predictions were referred to as “Homology-set”. For *ab initio* prediction methods, AUGUSTUS (v 3.2.3)⁴³ was used to predict coding regions in the repeat-masked genome. All gene models were combined using EvidenceModeler (v 2.1.0)⁴⁴ with different weight parameters assigned to evidence from different sources (10 for PASA-set, 5 for Stringtie-set, 5 for Homology-set, and 1 for AUGUSTUS gene prediction). The resulting protein-coding genes that were only derived from *ab initio* prediction were filtered out. Finally, the produced gene models were further refined with PASA (v 2.4.1)³⁷ to generate untranslated regions and alternative splicing variation information. The final comprehensive gene set comprised 63,678 genes.

The integrated gene set was translated into amino-acid sequences and annotated using the method described in Zhou *et al.*⁴⁵. As a result, 98.8% of the predicted protein-coding genes were functionally annotated (Table 5). In addition, we utilized the Diamond (v 0.9.30)⁴⁶ software (E value $\leq 1e-05$) to perform homologous comparison analyses between AJ8 and *Arabidopsis* (TAIR10)⁴⁷, as well as between AJ8 and rice (Osativa_v7_0)⁴⁸. Among the AJ8 genes, 48,284 (75.83%) genes exhibited homology with *Arabidopsis*, while 56,777 (89.16%) genes showed homology with rice (Table S1).

Gene expression analysis. The gene expression analysis was used the same method as previously reported⁴⁹. The expression heatmap was constructed using heatmap R package. The expression matrix of genes in different transcriptome samples was displayed in Table S2.

Syntenic analysis. The identification of syntenic regions was based on conducting homology searches using MCScan (Python version)⁵⁰, with a minimum requirement of 30 genes per block.

Subgenome phasing. By employing repetitive k-mers as “differential signatures” and utilizing the SubPhaser software, we successfully phased the subgenomes of AJ8. The results obtained from SubPhaser⁵¹ were found to be consistent with the Jinshu7 genome, as indicated in Table 6 and Fig. 4.

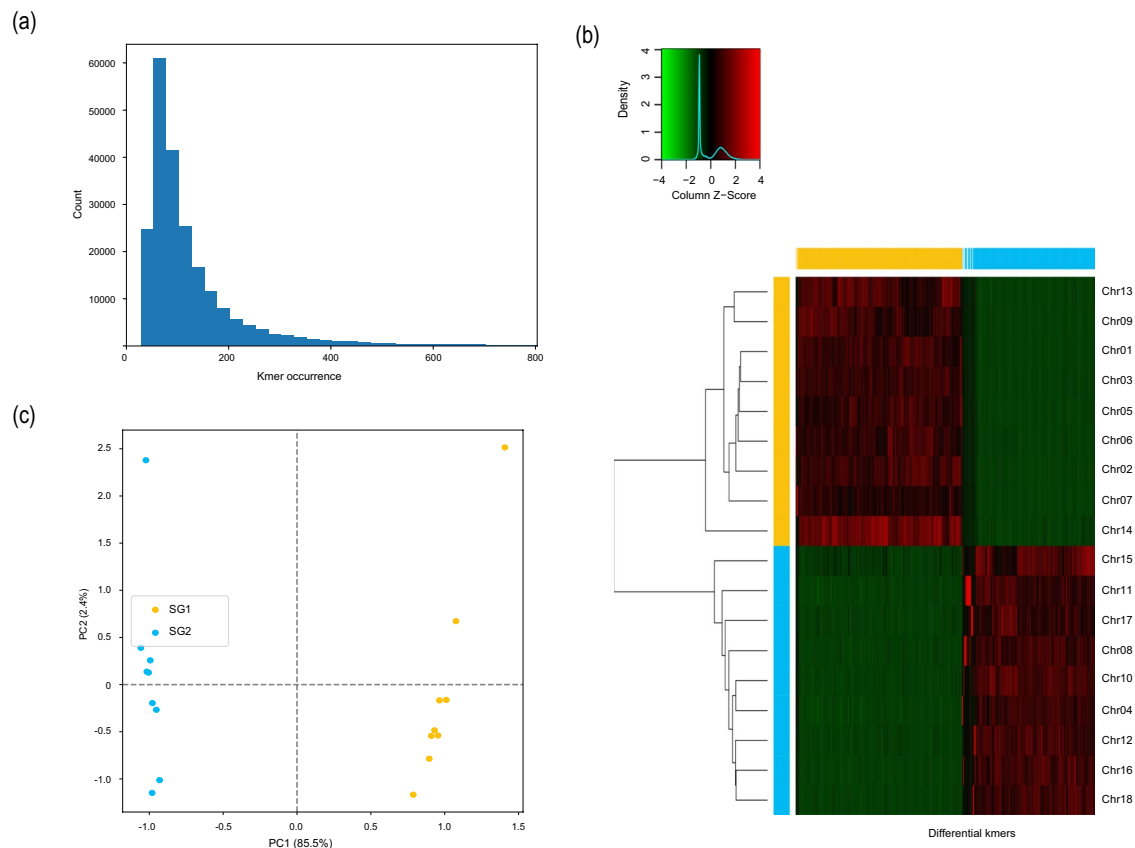


Fig. 4 Phased subgenomes of the AJ8 genome. **(a)** The histogram of differential k-mers among homoeologous chromosome sets. **(b)** Heatmap and clustering of differential k-mers. The x-axis, differential k-mers; y-axis, chromosomes. The vertical color bar, each chromosome is assigned to which subgenome; the horizontal color bar, each k-mer is specific to which subgenome (blank for non-specific kmers). **(c)** Principal component analysis of differential k-mers. Points indicate chromosomes.

Data Records

The sequencing data and assembled genome sequence have been deposited in the Sequence Read Archive with accession numbers SRP482566⁵² under project number PRJNA1059665. The genome assembly has been deposited at GenBank under the WGS accession

GCA_038442765.1⁵³. Files of the gene structure annotation, repeat predictions and gene functional annotation were deposited at Figshare⁵⁴.

Technical Validation

Genome assembly and gene prediction quality assessment. The accuracy and integrity of AJ8 T2T assembly were assessed through several analyses. Firstly, the Hi-C heatmap displayed consistent results across all chromosomes, indicating the correct ordering and orientation of contigs in the assembly (Fig. 2a). Secondly, the assembly successfully captured 18 centromeres and all 36 telomeres, providing strong evidence for their integrity (Fig. 2b; Tables 7, 8). Thirdly, the assembly showed high collinearity with Jinshu7 (GCA_026771285.1)⁷ and *Panicum hallii* (GCA_002211085.2)⁵⁵ (Fig. 2c). Fourthly, alignment results from minimap2 (v 2.24-r1122)⁵⁶ revealed that 100.0% of ONT reads and 99.98% of HiFi reads could be aligned to the AJ8 T2T assembly. Additionally, the average genome alignment rate of the transcriptome was 98.7% (Table 3). Lastly, AJ8 T2T demonstrated an LTR assembly index of 20.4, a quality value of 61.7, and a BUSCO score of 99.6%, indicating its high completeness (Table 1).

We compared the length distribution of genes among the AJ8, maize¹², sorghum (GCF_000003195.3)³⁹, LM_1⁸, and Jinshu7⁷ and found similar patterns (Fig. 3a). The BUSCO analysis showed that 99.5% (single-copy gene: 19.5%, duplicated gene: 80.0%) of 1,614 embryophyta single-copy orthologs were successfully identified as complete, while 0.2% were fragmented and 0.3% were missing in the assembly (Fig. 3b). 62,885 (98.8%) gene models were successfully annotated in diverse databases and 51,958 gene models (81.6%) exhibited detectable transcriptional activity (FPKM value ≥ 1) across 21 RNA-seq samples (Fig. 3c; Table 5; Fig. 3d). Taken together, these results provide strong evidence that a high-quality AJ8 genome has been obtained. The high-quality genome provides a solid foundation for uncovering the drought resistance and adaptive mechanisms of AJ8, and also serves as an important reference for the rapid breeding of AJ8 and other crops.

Code availability

No specific code or script was used in this work. Commands used for data processing were all executed according to the manuals and protocols of the corresponding software.

Received: 30 January 2024; Accepted: 6 June 2024;

Published online: 21 June 2024

References

1. Yue, H. *et al.* Transcriptome-wide identification and expression profiles of the WRKY transcription factor family in Broomcorn millet (*Panicum miliaceum* L.). *BMC genomics* **17**, 343 (2016).
2. Rajput, S. G., Santra, D. K. & Schnable, J. Mapping QTLs for morpho-agronomic traits in proso millet (*Panicum miliaceum* L.). *Molecular breeding: new strategies in plant improvement* **36**, 37 (2016).
3. Saleh, A. S. M., Zhang, Q., Chen, J. & Shen, Q. Millet Grains: Nutritional Quality, Processing, and Potential Health Benefits. *Comprehensive Reviews in Food Science and Food Safety* **12**, 281–295 (2013).
4. Lu, H. *et al.* Earliest domestication of common millet (*Panicum miliaceum*) in East Asia extended to 10,000 years ago. *Proceedings of the National Academy of Sciences* **106**, 7367–7372 (2009).
5. Habiyaemye, C. *et al.* Proso Millet (*Panicum miliaceum* L.) and Its Potential for Cultivation in the Pacific Northwest, U.S.: A Review. *Frontiers in Plant Science* **7** (2017).
6. Hunt, H. V. *et al.* Reticulate evolution in *Panicum* (Poaceae): the origin of tetraploid broomcorn millet, *P. miliaceum*. *Journal of Experimental Botany* **65**, 3165–3175 (2014).
7. Sun, Y. *et al.* Biased mutations and gene losses underlying diploidization of the tetraploid broomcorn millet genome. *The Plant journal* **113**, 787–801 (2022).
8. Shi, J. *et al.* Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nature communications* **10**, 464 (2019).
9. Wang, Z. *et al.* A high-quality, phased genome assembly of broomcorn millet reveals the features of its subgenome evolution and 3D chromatin organization. *Plant Communications* **4** (2023).
10. Shang, L. *et al.* A complete assembly of the rice Nipponbare reference genome. *Molecular plant* **16**, 1232–1236 (2023).
11. Navrátilová, P. *et al.* Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. *Plant Biotechnology Journal* **20**, 1373–1386 (2021).
12. Chen, J. *et al.* A complete telomere-to-telomere assembly of the maize genome. *Nature Genetics* **55**, 1221–1231 (2023).
13. Chin, C. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–569 (2013).
14. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**(3), 268–76 (2012).
15. Chen, Y. *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7** (2017).
16. Chen, S., Zhou, Y., Chen, Y. & Jia, G. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
17. 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).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
19. Servant, N. *et al.* HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16** (2015).
20. Durand, N. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
21. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, eaal3327 (2017).
22. Li, B. *et al.* A gap-free reference genome reveals structural variations associated with flowering time in rapeseed (*Brassica napus*). in *Horticulture research* Vol. 10 uhad171 (2023).
23. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature methods* **19**, 687–695 (2022).
24. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature methods* **19**, 705–710 (2022).
25. Vaser, R., Sovic, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**(5), 737–746 (2017).
26. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology (Clifton, N.J.)* **1962**, 227–245 (2019).
27. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21** (2020).
28. Qu, C. *et al.* Comparative genomic analyses reveal the genetic basis of the yellow-seed trait in *Brassica napus*. *Nature Communications* **14**, 5194 (2023).
29. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **5** (2004).
30. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–8 (2007).
31. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **21**(Suppl 1), i351–8 (2005).
32. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
33. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* (2023).
34. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**(7), 644–52 (2011).
35. Haas, B. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360 (2015).
37. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).
38. Hou, X., Wang, D., Cheng, Z., Wang, Y. & Jiao, Y. A near-complete assembly of an Arabidopsis thaliana genome. *Molecular plant* **15**, 1247–1250 (2022).
39. Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses[J]. *Nature* **457**(7229), 551–556 (2009).
40. Ramu, P. *et al.* Improved pearl millet genomes representing the global heterotic pool offer a framework for molecular breeding applications. *Communications Biology* **6** (2023).
41. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant *Setaria*[J]. *Nature biotechnology* **30**(6), 555–561 (2012).

42. Jens *et al.* GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in Molecular Biology* (2019).
43. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465–7 (2005).
44. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
45. Zhou, Y. *et al.* Gap-free genome assembly of Salangid icefish *Neosalanx taihuensis*. *Scientific data* **10**, 768 (2023).
46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).
47. Phytozome database, https://phytozome-next.jgi.doe.gov/info/Athaliana_TAIR10 (2022).
48. Phytozome database, https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0 (2022).
49. Sun, C. *et al.* Chromosome-level genome assembly for the largemouth bass *Micropterus salmoides* provides insights into adaptation to fresh and brackish water. *Molecular ecology resources* **21**, 301–315 (2021).
50. Tang, H. *et al.* Synteny and Collinearity in Plant Genomes. *Science* **320**, 486–488 (2008).
51. Zhang, R.-G. *et al.* Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Nature Communications* **14** (2023).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP482566> (2024).
53. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_038442765.1 (2024).
54. Wang, Haigang. Genome assembly and annotation of *Panicum miliaceum* AJ8. *figshare* <https://doi.org/10.6084/m9.figshare.24947565.v1> (2024).
55. Lovell, J. T. *et al.* The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nature Communications* **9** (2018).
56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2017).
57. Zou, C. *et al.* The genome of broomcorn millet. *Nature communications* **10**, 436 (2019).

Acknowledgements

This work was funded by Key Research and Development Program of Shanxi Province (2022ZDYF110), Project of Conservation and Utilization of Agricultural Germplasm Resources in Shanxi Province (sxzyk202201), the Central Guiding Local Science and Technology Development Fund Project (YDZJSX20231A035), China Agriculture Research System of MOF and MARA (CARS-06-14.5-A16).

Author contributions

Z.X.M., Z.J.Q., and L.W. conceived the study. H.G.W., J.J.W., H.B.Q., and M.L. collected and prepared the samples. C.H.C., H.G.W., X.K.Y., and J.B.J. performed bioinformatics analysis. H.G.W., C.H.C. and L.C. wrote the manuscript with significant contributions. X. T., S.Hou., X.K.Y., P.G. and J.B.J. revised the manuscript. All authors read, edited, and approved the final 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/s41597-024-03489-5>.

Correspondence and requests for materials should be addressed to H.W., L.W., Z.Q. or Z.M.

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) 2024