scientific data

OPEN

DATA DESCRIPTOR

The telomere-to-telomere genome assembly of the wild mulberry, *Morus mongolica*

Jinhong Yang[1,2], Yunwu Peng[1,2], Fang Yang[2], Gang Meng[1,2] & Weiqing Kong[1,2] ✉

Morus mongolica **is a wild mulberry native to China and North Korea. In the current study, we assembled a high-quality telomere-to-telomere genome sequence of *M. mongolica* using NGS, HiFi, ONT, and Hi-C technologies. The genome was determined to be 341.88 Mb in size with a contig N50 of 23.82 Mb. The numbers of telomeres and centromeres were 28 and 14, with average lengths of 9.86 kb and 1.91 Mb, accounting for 0.08% and 7.84% of the total genome, respectively. A total of 21,657 protein-coding genes and 186.50 Mb repeat sequences were annotated. Genome integrity evaluation by BUSCO revealed a completeness score of 99.44% and a quality value of 46.7. Collinearity analysis between *M. mongolica* and either *Morus alba* or *Morus notabilis* showed that the breakage and fusion of chromosomes in *Morus* occurred at the centromere region of *M. notabilis*, which provided important genomic evidence for the evolution and chromosome breakage-fusion mechanism of *Morus* species.**

Background & Summary

Mulberry (family Moraceae, genus *Morus*) is an important economic crop that is widely distributed in Asia, Europe, Africa, Oceania, and the Americas, with over 3,000 varieties[1]. Mulberry is an essential component of the traditional silk industry (sericulture), and planting mulberry to feed silkworms has a history of over 5,000 years in China, strongly influencing the world through the Silk Road[2]. Mulberry is also valued for its fruits, uses in traditional Chinese medicine, and its timber[3]. The genus *Morus* contains 10–16 species according to their morphometric and/or molecular markers[4,5]. There is also controversy over the origin of the *Morus* genus. Systematic phylogenetic analysis, based on the sequences of the internal transcribed spacer (ITS) between the large and small subunit rRNA sequences, and the chloroplast *trn*L-*trn*F intergenic spacer region of 12 *Morus* species confirmed the monophyletic evolution of *Morus*[6], while Nepal (2012) divided the global *Morus* genus into 13 species and emphasized that *Morus* may have evolved from two subgenera from North America and Asia[7], indicating that clarification of the taxonomy of *Morus* needs more and stronger supporting evidence.

The basic chromosome number (n) is crucial for research into the ploidy of plant germplasm and the evolutionary path of populations. It is assumed that most *Morus* species are diploid with n = 6 (sometimes separated into 7 chromosomes after mitosis) or n = 14[8]. *Morus notabilis*[9] and *Morus yunnanensis*[10] have been reported to have chromosome numbers of n = 6, while *Morus alba*, which includes most of the cultivated varieties, has chromosome numbers of n = 14[11]. The other species closely related to *M. alba*, such as *Morus atropurpurea*, *Morus bombycis*, and *Morus indica*, have chromosome numbers of n = 14, too[11]. In addition, there are also a large number of polyploid mulberry species, such as *Morus cathayana*, which has three types of polyploid: triploid, tetraploid and hexaploid[12], and *Morus nigra*, a mulberry of natural decosaploid with 308 chromosomes[13]. With the development of next-generation sequencing technology, the genomes of many mulberry species have been successfully sequenced, such as *M. notabilis*[9], *M. alba*[14], *M. atropurpurea*[15], *M. indica*[16], and *M. yunnanensis*[10], and these data are helpful in more deeply understanding and revealing the complex phylogenetic relationships among mulberry species and identifying valuable mulberry germplasm resources. The genome of *M. notabilis* has been sequenced by telomere-to-telomere (T2T) genome sequencing, but there are no reports of T2T genome sequencing of *Morus* species with n = 14.

Morus mongolica is a wild mulberry native to China and North Korea, and is regarded as a separate species on the basis of various classification methods[17]. In the current study, we identified the ploidy level and basic chromosome number of *M. mongolica*, sequenced its genome using MGI next-generation sequencing (NGS),

[1]Shaanxi key laboratory of sericulture, Ankang University, Ankang, China. [2]School of Modern Agriculture & Biotechnology, Ankang University, Ankang, China. ✉e-mail: weiqing_kongwq@126.com

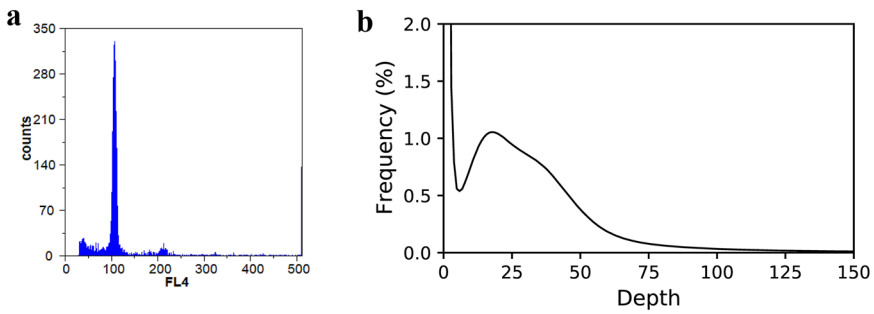

Fig. 1 Ploidy and genome size estimate of *Morus mongolica*. (**a**) Ploidy analysis by Flow cytometry. (**b**) K-mer distribution.

Type	Clean reads	Number of bases	Mean (bp)	N50 (bp)	Maximum (bp)	Coverage (%)/Depth (x)
DNBSEQ-T7RS	136,449,202	18,914,708,912	150	150	150	99.90/51.73
PacBio HiFi	2,002,536	31,096,955,435	15,528	15,806	56,381	99.75/84.64
Hi-C	430,581,442	64,545,834,332	150	150	150	—
ONT	2,278,593	59,794,270,254	26,241	74,755	658,192	99.23/165.83
Illumina RNA-seq	42,997,036	6,446,479,955	150	150	150	—

Table 1. Statistics for the sequencing data of the *Morus mongolica* genome.

PacBio HiFi, ultra-long Oxford Nanopore Technologies (ONT), and Hi-C sequencing, bridged all assembly gaps in the currently available reference genomes, and assembled its high-quality T2T genome. The study provided a foundation for further analysis of the evolutionary relationship and chromosome recombination between mulberry species.

Methods

Sample collection. The male *M. mongolica* plant used in this study was transplanted from Langao County, Shaanxi Province, China (32.40°N, 108.75°E), and re-planted at the research base of Shaanxi Key Laboratory of Sericulture. In May 2023, leaves were collected and immediately snap-frozen in liquid nitrogen for subsequent library preparation, genome sequencing, and ploidy analysis. The related sequencing services were performed by Grandomics Biosciences Co. Ltd. (Wuhan, China).

Ploidy analysis. The ploidy of *M. mongolica* was identified using CyFlow Space Flow Cytometer and CyStain UV Precise P Stain (Sysmex Partec, Norderstedt, Germany), according to the manufacturer's operating procedure. *M. alba* 'Heyebai', a domesticated diploid mulberry, with 2n = 28, was handled simultaneously as the standard[14]. Comparison of the mean fluorescence intensity between them indicated that *M. mongolica* is a diploid mulberry (Fig. 1a).

Genome size estimation. The genomic DNA (gDNA) was extracted using the cetyltrimethylammonium bromide (CTAB) method. The gDNA library was constructed and sequenced on the DNBSEQ-T7RS Genetic Sequencer platform (MGI, Shenzhen, China) with 150-bp paired-end reads. As a result, a total of 18.91 Gb clean data was obtained and used for further analysis (Table 1). The genome size and heterozygosity of *M. mongolica* was first estimated using KMC software[18] with 17-mer frequency distribution (Fig. 1b). As a result, a total of 12,866,189,456 k-mers, with a depth of 36, was obtained, the estimated genome size was 357.40 Mb, and the heterozygosity ratio was 2.00%.

Pacific Biosciences (PacBio) HiFi sequencing. The gDNA to be used in PacBio HiFi Sequencing was evaluated using agarose gel electrophoresis and Qubit® 3.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) to ensure the production of PacBio long reads. Then, the SMRTbell large target size libraries were constructed according to PacBio's standard protocol and sequenced on a PacBio Revio long-read sequencer instrument with the Revio Kit (Pacific Biosciences, Menlo Park, CA, USA). The quality control of raw reads was performed using the PacBio SMRT-Analysis package (https://www.pacb.com), and a total of 31.10 Gb clean data, with an average read length of 15.53 kb and an N50 value of 15.81 kb, was obtained (Table 1).

Oxford nanopore technology (ONT) sequencing. The ultra-long gDNA for ONT sequencing was extracted using the QIAGEN® Genomic DNA Extraction Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. Approximately 8–10 μg gDNA was purified using the SageHLS HMW Library System (Sage Science, Beverly, MA, USA), and subjected to the PippinHT system (Sage Science, MA, USA) for size-selection of long DNA fragments (>50 kb). Then, the DNA was repaired and attached with adapters using the Ligation Sequencing 1D Kit (Catalog No. SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK). The library (approximately 400 ng) was measured with a Qubit® 3.0 Fluorometer (Invitrogen, Carlsbad, CA, USA)

Assembly step	Total length (bp)	Number of scaffold (Chromosome)	Longest scaffold (bp)	N50 length (bp)	N90 length (bp)
HiFiasm	357,899,125	125	33,595,736	19,803,500	16,094,918
Hi-C	356,322,934	104(14)	70,892,002	23,816,741	16,872,853
Gap filling	342,063,602	(14)	71,390,041	23,816,741	16,872,853
NextPolish2	341,884,498	(14)	71,327,794	23,816,835	16,870,394

Table 2. Summary of each step in construction of the *Morus mongolica* genome assembly.

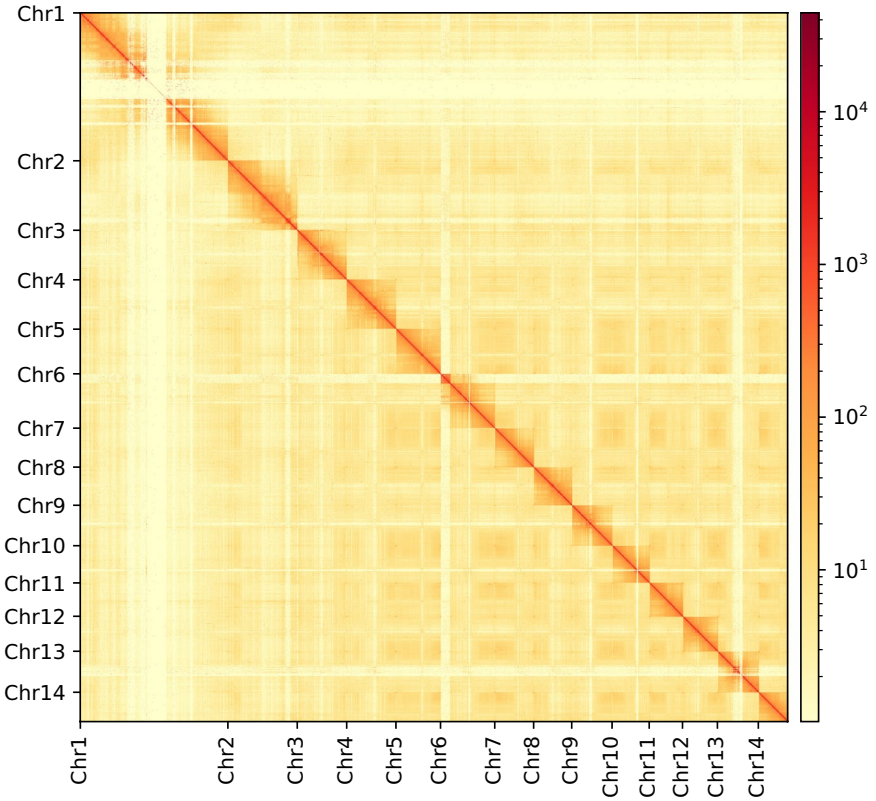

Fig. 2 Hi-C interaction matrix for *Morus mongolica*. In general, intra-chromosomal interactions (blocks on the diagonal line) were strong, whereas inter-chromosomal interactions were weaker. Color indicates Hi-C interaction frequency.

and sequenced on the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). In total, 2,278,593 long-sequence reads, with 59.79 Gb clean data and an N50 value of 74.76 kb, were obtained from the ONT platform (Table 1).

Hi-C library construction and sequencing. Leaves of *M. mongolica* were cut into 1- to 2-mm-wide strips and immersed in formaldehyde to fix and crosslink the DNA, which was then incubated with the restriction enzyme *Dpn*II (New England Biolabs, MA, USA) to produce sticky ends. The DNA fragments were then ligated to biotin-14-dCTP (TriLINK Bio Technologies, San Diego, CA, USA) by terminal DNA repair, and blunt-end ligation was carried out using T4 DNA ligase. Next, proteinase K (Thermo Scientific, Waltham, MA, USA) was used to release and dissociate the proteins from the crosslinked DNA. Finally, the DNA was purified, assessed, and used to construct the sequencing library. The sequencing was performed on the DNBSEQ-T7RS platform (MGI, Shenzhen, China) with 150-bp paired-end reads. The raw data obtained by Hi-C library sequencing were filtered using fastp[19] to exclude low-quality Hi-C reads (quality scores < 20), adapter sequences, and sequences shorter than 30 bp, and a total of 430,581,442 clean reads was ultimately generated (Table 1).

Genome assembly. We assembled the *M. mongolica* genome via four steps: initial package, assisted assembly, gap filling, and genome correction. The initial package was a mixed assembly of HiFi and ONT sequencing data using hifiasm (v0.19) software with the default parameters[20]. A draft genome of 357.90 Mb with 125 contigs and an N50 value of 19.80 Mb was assembled (Table 2).

For assisted assembly, the data from Hi-C sequencing were mapped to the draft genome using Bowtie2 (v2.3.2) (-end-to-end–very-sensitive -L 30) to screen for the valid interaction pairs[21], which were then hierarchically clustered using LACHESIS software[22]. Then, the placement and orientation errors, with obvious

Chromosome			Telomere length (bp)			Mean depth		
Number	Qv	Length (bp)	Left	Right	Centromere boundary (Left : Right)	NGS	HiFi	ONT
1	45.99	71,327,794	6,693	4,950	23,483,868 : 44,562,486	62.13	74.83	157.14
2	48.40	33,523,929	11,453	4,531	28,011,170 : 28,833,076	50.93	78.62	152.24
3	44.77	23,896,502	1,568	4,840	11,287,972 : 11,754,481	52.06	84.75	163.78
4	47.64	23,816,835	13,245	13,463	13,445,130 : 13,974,273	50.64	87.98	170.23
5	47.74	21,538,889	11,856	9,884	12,302,833 : 12,867,965	48.65	94.00	178.80
6	48.24	26,215,542	9,981	21,433	13,510,193 : 13,618,719	47.18	86.83	158.14
7	45.71	18,793,284	17,267	6,168	11,013,318 : 11,374,285	49.92	95.09	177.64
8	46.54	18,409,326	12,800	2,553	8,695,975 : 9,216,490	45.55	85.49	162.79
9	44.26	19,502,757	10,424	7,661	8,237,504 : 9,037,692	52.04	92.31	173.19
10	47.79	17,960,080	7,679	9,299	11,337,099 : 11,682,852	47.80	89.47	169.91
11	48.00	16,094,925	10,076	15,322	8,776,576 : 8,955,822	48.19	96.10	177.96
12	43.94	16,870,394	14,833	395	7,580,578 : 7,736,432	45.59	87.43	168.95
13	46.94	19,782,385	6,050	12,069	8,256,790 : 8,938,486	59.71	90.75	172.83
14	48.37	14,151,856	11,527	18,104	11,156,939 : 11,335,505	48.68	97.65	181.80

Table 3. The statistics for each chromosome of the *Morus mongolica* genome assembly.

type		Number	Length (bp)	Percentage (%)
TE	LINE	39,495	8,949,225	2.62
	LTR	306,907	117,391,460	34.34
	SINE	3,845	408,287	0.12
	DNA	201,808	38,840,793	11.36
	RC	12,497	3,074,846	0.90
	MITE	3,462	857,394	0.25
TR	SSR	45,636	608,346	0.18
	Tandem repeat	55,777	8,440,595	2.47
Other		46,031	7929348	2.32
Total Repeats		715,458	186,500,294	54.55

Table 4. Statistics of repetitive sequence for the *Morus mongolica* genome.

Type		Copy Number	Total Length (bp)	Percentage (%)
tRNA		485	36,960	0.0108
miRNA		93	12,623	0.0037
rRNA	18S	528	957,605	0.2801
	28S	527	3,716,825	1.0872
	5.8S	1	157	0
	5S	85	9,794	0.0029
snRNA	snRNA	898	94,995	0.0278
	splicing	68	9,531	0.0028
	cis-regulatory	7	425	0.0001
	other	10	1,997	0.0006

Table 5. Non-coding RNA annotation of the *Morus mongolica* genome.

discrete chromatin interaction patterns, were manually adjusted, and the interaction heat map was plotted with HiCExplorer v3.6[23] (Fig. 2). After this step, an assisted assembled genome of 356.32 Mb with 104 contigs was obtained. 96.03% of them (342.18 Mb) could be assembled onto 14 different chromosomes (Table 2 and Fig. 2).

For gap filling, minimap2 was used to align the ONT data and the assisted assembled genome[24]. Any reads that were unable to align with the genome or that aligned to the end of contigs were extracted, iteratively assembled locally, aligned back to the genome, and replaced the corresponding gap area. Then, NextPolish2[25] was used to correct base errors by HiFi reads and NGS reads to generate the final genome, which was 341.88 Mb in size (Table 2) and had been deposited at GenBank.

Identification of telomeres and centromeres. The telomeres were identified using Telomere Identification ToolKit (Tidk) v0.2.31[26] with search string 'CCCTAAA/TTTAGGG'. The centromeres were detected

Gene set	Total number of genes	Average gene length (bp)	Average CDS length (bp)	Average exons number per gene	Average exon length (bp)	Average intron length (bp)
RNA-seq/PASA	22,741	5,777.4	2,104.42	7.03	299.43	609.32
homo/GeMoMa	48,535	7,595.84	879.03	3.55	247.72	2,635.58
denovo/AUGUSTUS	23,159	3,612.18	1,292.01	5.65	228.86	499.45
EVM	21,657	3,808.9	1,316.25	5.62	234.02	539.01

Table 6. Statistics of gene predictions in the *Morus mongolica* genome.

Annotation	Number	Percent (%)
Swissprot	17,359	80.15
KEGG	8,328	38.45
KOG	11,368	52.49
GO	13,631	62.94
NR	21,107	97.46
total	21,115	97.50

Table 7. Summary of functional annotations for predicted genes of the *Morus mongolica* genome.

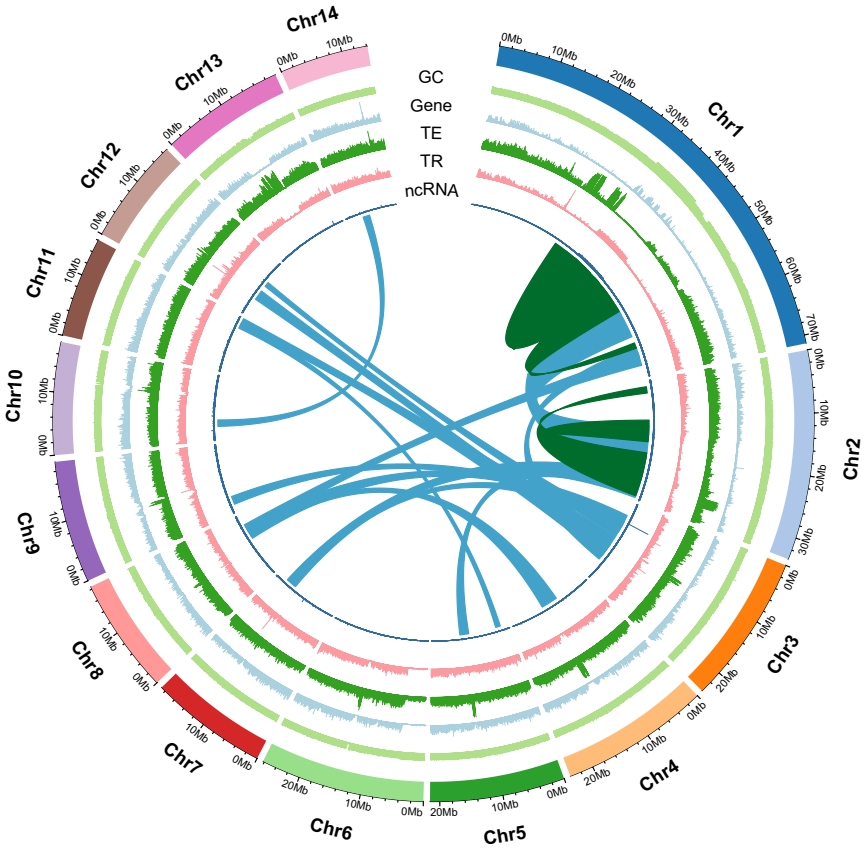

Fig. 3 Genome characteristics of *Morus mongolica*. Circos plot from outer to inner layers depicts the following: GC content; gene density; TE retroelement density; TR density and ncRNA density. The plot was drawn in 50 kb sliding windows, the homologous region is displayed in the center.

using the quarTeT toolkit[27], a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification, with the default parameters. As a result, 28 telomeres with an average length of 9,861.57 bp, accounting for 0.08% of the total genome, and 14 centromeres with an average length of 1,913,759.50 bp, accounting for 7.84% of the total genome, were identified in *M. mongolica* genome v1.0 (Table 3).

Repetitive DNA element annotation. For tandem repeats (TR), we used GMATA[28] and Tandem Repeats Finder[29] to search for simple sequence repeat (SSR) and TR sequences in *M. mongolica* genome v1.0, respectively.

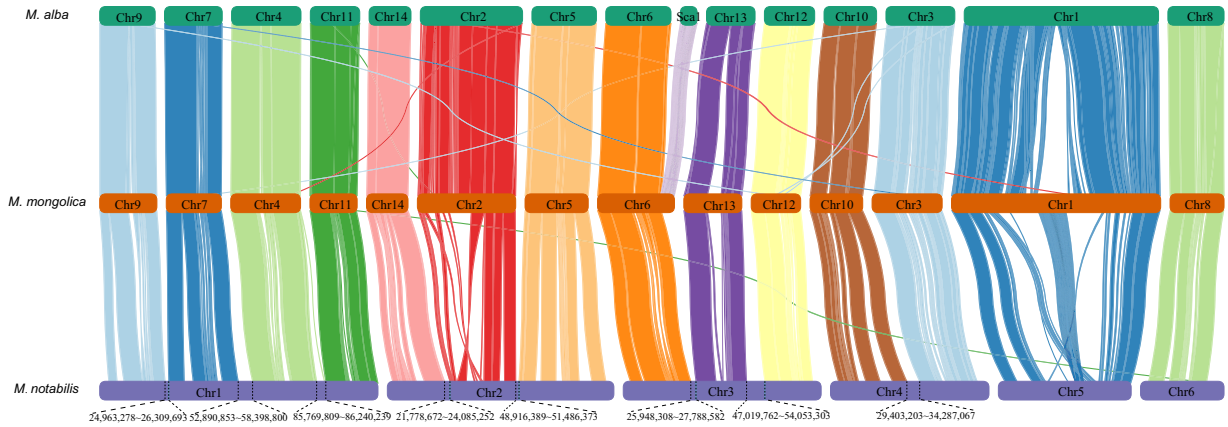

Fig. 4 Genome Collinearity analysis between the three *Morus* genome (*Morus alba*, *Morus mongolica* and *Morus notabilis*). The numbers indicated the putative rupture and fusion sites. Sca1 of *Morus alba* indicated a scaffold could not be assembled into a chromosome.

Type	Number	Percent (%)
Complete BUSCOs (C)	1,605	99.44
Complete and single-copy BUSCOs (S)	1,589	98.45
Complete and duplicated BUSCOs (D)	16	0.99
Fragmented BUSCOs (F)	2	0.12
Missing BUSCOs (M)	7	0.43

Table 8. BUSCO analysis of the *Morus mongolica* genome completeness.

A total of 608,346 bp of SSR sequences and 8,440,595 bp of TR sequences, accounting for 0.18% and 2.47% of the *M. mongolica* genome, respectively, were annotated. We then merged these two datasets, removed duplications between them, and named the resulting data TR.lib.

We identified the transposable elements (TE) in the TR.lib soft-masked genome using two methods, namely *de novo* and homology-based strategies. For *de novo* prediction, RepeatModeler[30] was used to construct a *de novo* library, which was named RepMod.lib. Then, the MITE-hunter[31], LTR_FINDER[32], LTRharvest[33] and LTR_retriever[34] were used to generate a long terminal repeat (LTR) library, named TE.lib. For homology-based strategies, the Repbase library was used[35]. Finally, we merged the rebase library, TE.lib, and RepMod.lib, and used RepeatMasker to identify the repeat content in the *M. mongolica* genome. As a result, a total of 169,522,005 bp repeat sequences was obtained, of which LTRs (34.34%), DNA (11.12%), and long interspersed nuclear elements (LINEs, 2.36%) were the major repetitive elements (Table 4).

Non-coding RNA annotation. The tRNA sequences in the *M. mongolica* genome were predicted using tRNAscan-SE[36]. Small nuclear RNAs (snRNAs), microRNAs (miRNAs) and rRNAs were predicted based on the alignment with the Rfam database using Infernal (v1.1.4)[37] and RNAmmer (v1.2)[38] software. A total of 2,702 non-coding RNAs (ncRNAs), accounting for 1.42% of the total genome, was identified (Table 5) in the *M. mongolica* genome.

Gene prediction. The total RNA from leaves of *M. mongolica* was extracted and used for cDNA library construction. High-throughput sequencing was carried out on the Illumina HiSeq 2000 (Illumina, San Diego, CA, USA) platform and a total of 6.45 Gb RNA sequencing (RNA-seq) data was obtained (Table 1). Genes were predicted by three methods: RNA-seq-based prediction, *de novo* prediction, and homology-based prediction. RNA-seq-based prediction was executed using PASA[39] software and 22,741 genes were predicted. For *de novo* prediction, the genes obtained by RNA-seq-based prediction were compared with the SWISS-PROT database and those with a consistency greater than 95% were selected and used in GeneMark-ST for further self-training of gene starts[40]. The 3,000 genes with the highest scores were then retained as the AUGUSTUS model training set and used to perform the *de novo* gene prediction[41]. For homology-based prediction, GeMoMa v1.730[42] was run using gene models from *M. atropurpurea*[15], *M. notabilis*[9], *Ficus hispida*[43], and *M. yunnanensis*[10]. Totals of 23,159 and 48,535 genes were obtained by *de novo* prediction and homology-based prediction, respectively (Table 6).

EVidenceModeler[44] was then used to integrate the results of the three methods and generate an initial gene set of *M. mongolica* (Table 6). Then, TransposonPSI[45] was used to remove any genes containing transposable elements or with coding errors from the initial gene set, and the resulting genes represented the final gene set. A total of 21,657 genes, with an average length of 3808.9 bp, was ultimately predicted. The average length of coding sequences (CDS) was 1316.25 bp (Table 6), the longest of all the assembled mulberry genomes[9,10,14–16]. The average number and length of exons were 5.62 and 234.02 bp, respectively, and the average length of introns

was 539.01 bp. Collinearity within the *M. mongolica* genome was analyzed with the Multiple Collinearity Scan toolkit X version (MCScanX)[46], using the default parameters. All the annotations, namely GC content, gene, TE elements, TR elements, and ncRNAs, were visualized, using circlize (Circular Visualization in R) v0.4.15[47] software package with a sliding window distance of 50 kb (Fig. 3).

Gene function annotation. For gene function annotation, the protein sequences were comprehensively aligned with the NCBI Non-Redundant-Protein-Database (NR) and SWISS-PROT, using the BLASTp[48] program with parameters E-value 1e−5 and max_target_seqs. 1. Then, the KEGG[49] and KOG[50] pathway information was annotated according to the KEGG (Kyoto Encyclopedia of Genes and Genomes) and KOG (Eukaryotic Orthologous Groups of protein) databases. The Gene Ontology (GO) annotation was performed by comparing protein sequences against the Pfam database, using InterproScan software[51]. The Interpro numbers obtained were converted into corresponding GO annotation and classified into the ontology aspects, namely biological process, cellular component, and molecular function. In total, 97.50% of the predicted genes were functionally annotated (Table 7).

Collinearity analysis. We conducted a comparative genome analysis among *Morus* species to assess their collinearity. *M. alba*[14], a model species of *Morus*, and *M. notabilis*[9], the only species for which a T2T genome had previously been completed, were selected. Minimap2 was used to identify homologous syntenic blocks between *M. mongolica* and either *M. alba*[52] or *M. notabilis*[53]. The relationships among the blocks were visualized using NGenomeSyn[54] with parameter MinAlnLen 20000. The results showed that there was high collinearity between *M. mongolica* and *M. alba* (Fig. 4), while the collinearity between *M. mongolica* and *M. notabilis* was interesting. Chromosomes 1 to 4 of *M. notabilis* corresponded to different chromosome numbers of *M. mongolica*. Chromosome 1 of *M. notabilis* corresponded to four chromosomes of *M. mongolica*, which are chromosomes 4, 7, 9 and 11. Each of the chromosomes 2 and 3 of *M. notabilis* was homologous with three chromosomes of *M. mongolica*, which are chromosomes 2, 5, 14 and chromosomes 6, 12, 13, respectively. *M. notabilis* chromosome 4 matched with chromosomes 3 and 10 of *M. mongolica*. Chromosome 5 of *M. notabilis*, which underwent rupture and fusion in its replication process, corresponded to chromosome 1, the largest chromosome of *M. mongolica*. Chromosome 6 of *M. notabilis* corresponded to chromosome 8 of *M. mongolica*. Thus, eight breakage points were formed on chromosomes 1, 2, 3, and 4 of *M. notabilis*, and further analysis on them showed that they were all located at the centromere regions[9]. This finding is consistent with the view that chromosome breakage often occurs in DNA repeat regions, such as near telomeres or centromeres[55,56].

Data Records

The genomic MGI DNBSEQ-T7RS sequencing data and Hi-C sequencing data were deposited in the NCBI Sequence Read Archive (SRA) database under accession No. SRR31045530[57] and No. SRR31045529[58].

The genomic Pacbio sequencing data were deposited in SRA database under accession No. SRR31188304[59].

The genomic Oxford Nanopore Technology (ONT) sequencing data were deposited in SRA database under accession No. SRR31066347[60].

The transcriptome Illumina sequencing data were deposited in SRA database under accession No. SRR31045527[61].

The assembled genome was deposited in the GenBank at NCBI under accession No. JBIQNV000000000[62].

Technical Validation

The completeness of the *M. mongolica* genome was primarily assessed from two perspectives, namely the genome sequences and the annotated protein sequences. For the genome sequences, we employed the data generated from two methods. Firstly, the data from the DNBSEQ-T7RS platform and the meryl tool[63] were used to generate a 17 k-mer database and used Merqury (v1.3)[64] to evaluate the quality value (qv) of the genome based on the k-mer. The qv ranged from 43.94 (chromosome 12) to 48.40 (chromosome 2) (Table 3). Secondly, the sequences from DNBSEQ-T7RS, HiFi, and ONT were mapped to the assembled genome. The mapping rate and genome coverage values were 99.90% and 51.73%, 99.75% and 84.64% and 99.23% and 165.83%, respectively. For the annotated protein sequences, the BUSCO[65] assessment was employed based on the embryophyta_odb10 database, and 1,605 universally conserved genes, accounting for 99.44% of the total, were successfully identified (Table 8). These results indicate that the assembly of the *M. mongolica* genome is high quality.

Code availability

The manuscript did not use custom code to generate or process the data described.

Received: 4 December 2024; Accepted: 22 April 2025;
Published online: 25 April 2025

References

1. Liu, L., Zhang, L., Zhao, W. G. & Pan, Y. L. Comparison of mulberry germplasm resources between China and overseas. *J. Plant Genet. Resour.* **5**, 285–289, https://doi.org/10.13430/j.cnki.jpgr.2004.03.016 (2004).
2. He, N. *et al*. Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat. Commun.* **4**, 2445, https://doi.org/10.1038/ncomms3445 (2013).
3. Chan, E. W., Lye, P. Y. & Wong, S. K. Phytochemistry, pharmacology, and clinical trials of *Morus alba*. *Chin. J. Nat. Med.* **14**, 17–30, https://doi.org/10.3724/SP.J.1009.2016.00017 (2016).
4. Zeng, Q. *et al*. Comparative and phylogenetic analyses of the chloroplast genome reveal the taxonomy of the *Morus* genus. *Front. Plant Sci.* **13**, 1047592, https://doi.org/10.3389/fpls.2022.1047592 (2022).

5. Zhou, Z. K. & Gilbert, M. G. Moraceae. in *Flora of China* Vol. 5 (eds. Wu, Z. Y., Raven, P. H. & Hong, D. Y.) pp. 21–73 (Science Press & Missouri Botanical Garden Press, 2003).
6. Zhao, W. *et al*. Phylogeny of the genus *Morus* (Urticales: Moraceae) inferred from ITS and *trnL-F* sequences. *Afr. J. Biotechnol.* **4**, 563–569, https://doi.org/10.1186/1475-2859-4-18 (2005).
7. Nepal, M. P. & Ferguson, C. J. Phylogenetics of *Morus* (Moraceae) inferred from ITS and *trnL-trnF* sequence data. *Syst. Bot.* **37**, 442–450, https://doi.org/10.1600/036364412X635485 (2012).
8. Xuan, Y. *et al*. FISH-based mitotic and meiotic diakinesis karyotypes of *Morus notabilis* reveal a chromosomal fusion-fission cycle between mitotic and meiotic phases. *Sci. Rep.* **7**, 9573, https://doi.org/10.1038/s41598-017-10079-6 (2017).
9. Ma, B. *et al*. The gap-free genome of mulberry elucidates the architecture and evolution of polycentric chromosomes. *Hortic. Res.* **10**, uhad111, https://doi.org/10.1093/hr/uhad111 (2023).
10. Xia, Z. *et al*. Chromosome-level genomes reveal the genetic basis of descending dysploidy and sex determination in *Morus* plants. *Genom. Proteom. Bioinf.* **20**, 1119–1137, https://doi.org/10.1016/j.gpb.2022.08.005 (2022).
11. Jiao, F., Xue, Z. M. & Su, C. Research status quo in polyploid identification and breeding in mulberry (*Morus* spp.). *Acta Agric. Boreali-Occident. Sin.* **24**, 1–13, https://doi.org/10.7606/j.issn.1004-1389.2015.12.001 (2015).
12. Han, M. Z., Wang, S. X., Zhu, G. Y., Su, C. & Jiao, F. Chromosome ploidy of *Morus* plants in Shaanxi. *North Sericult.* **15**, 17–20 (1994).
13. Chu, R. Y. & Sun, X. X. The studies on cytogentics of plants of the genus *Morus*, I. the number of chromosome of some mulberry varieties. *Acta Sericol. Sin.* **12**, 199–202 (1986).
14. Jiao, F. *et al*. Chromosome-level reference genome and population genomic analysis provide insights into the evolution and improvement of domesticated mulberry (*Morus alba*). *Mol. Plant* **13**, 1001–1012, https://doi.org/10.1016/j.molp.2020.05.00 (2020).
15. Dai, F. *et al*. Genomic resequencing unravels the genetic basis of domestication, expansion, and trait improvement in *Morus atropurpurea*. *Adv. Sci.* **10**, 2300039, https://doi.org/10.1002/advs.202300039 (2023).
16. Jain, M. *et al*. Draft genome sequence of Indian mulberry (*Morus indica*) provides a resource for functional and translational genomics. *Genomics* **114**, 110346, https://doi.org/10.1016/j.ygeno.2022.110346 (2022).
17. Meng, Y. C. Induction of polyploidy germplasm resources of *Morus* mongolica and preliminary evaluation of resistance of polyploid *Morus notabilis* (Southwest University, 2019).
18. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761, https://doi.org/10.1093/bioinformatics/btx304 (2017).
19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
20. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
21. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
22. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125, https://doi.org/10.1038/nbt.2727 (2013).
23. Ramírez, F. *et al*. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189, https://doi.org/10.1038/s41467-017-02525-w (2018).
24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
25. Hu, J. *et al*. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. *Genom. Proteom. Bioinf.* **22**, qzad009, https://doi.org/10.1093/gpbjnl/qzad009 (2024).
26. Brown, M., González, De la Rosa, P. M. & Mark, B. A telomere identification toolkit. *Zenodo* https://doi.org/10.5281/zenodo.10091385 (2023).
27. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
28. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **7**, 1350, https://doi.org/10.3389/fpls.2016.01350 (2016).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
30. Bedell, J. A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041, https://doi.org/10.1093/bioinformatics/16.11.1040 (2000).
31. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199, https://doi.org/10.1093/nar/gkq862 (2010).
32. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
33. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinf.* **9**, 18, https://doi.org/10.1186/1471-2105-9-18 (2008).
34. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
35. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
36. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
37. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
38. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
39. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
40. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
41. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
42. Zhang, X., Wang, G., Zhang, S., Chen, S. & Ming, R. Genomes of the Banyan Tree and Pollinator Wasp Provide Insights into Fig-Wasp Coevolution. *Cell* **183**, 875–889, https://doi.org/10.1016/j.cell.2020.09.043 (2020).
43. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89, https://doi.org/10.1093/nar/gkw092 (2016).
44. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).

45. Urasaki, N. *et al*. Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* **24**, 51–58, https://doi.org/10.1093/dnares/dsw047 (2017).
46. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49, https://doi.org/10.1093/nar/gkr1293 (2012).
47. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. *circlize* implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812, https://doi.org/10.1093/bioinformatics/btu393 (2014).
48. Boratyn, G. M. *et al*. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–W33, https://doi.org/10.1093/nar/gkt282 (2013).
49. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
50. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269, https://doi.org/10.1093/nar/gku1223 (2015).
51. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848, https://doi.org/10.1093/bioinformatics/17.9.847 (2001).
52. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JABXEQ00000000 (2021).
53. *CNCB Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/GWHCBHX00000000 (2023).
54. He, W. *et al*. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. *Bioinformatics* **39**, btad121, https://doi.org/10.1093/bioinformatics/btad12 (2023).
55. Schubert, I. & Lysak, M. A. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **27**, 207–216, https://doi.org/10.1016/j.tig.2011.03.004 (2011).
56. Fishman, L., Willis, J. H., Wu, C. A. & Lee, Y. W. Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkeyflowers (*Mimulus*; Phrymaceae). *Heredity* **112**, 562–568, https://doi.org/10.1038/hdy.2013.143 (2014).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31045530 (2024).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31045529 (2024).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31188304 (2024).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31066347 (2024).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31045527 (2024).
62. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBIQNV000000000 (2024).
63. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
64. Miller, J. R. *et al*. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824, https://doi.org/10.1093/bioinformatics/btn548 (2008).
65. 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, https://doi.org/10.1093/bioinformatics/btv351 (2015).

Acknowledgements

This work was supported by grants from the Scientific Research Program Funded by Education Department of Shaanxi Provincial Government (20JS003), the Key Research and Development Program of Shaanxi Province (2021NY-216).

Author contributions

Jinhong Yang and Weiqing Kong conceived and designed the research. Yunwu Peng, Fang Yang and Gang Meng collected and prepared the samples. Jinhong Yang and Weiqing Kong wrote the manuscript. Jinhong Yang and Gang Meng assembled the genome. Fang Yang modified the manuscript. All authors contributed to the article and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.K.

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.