Plant Communications, Volume *4*

Supplemental information

Insights into the evolution and spatial chromosome architecture of jujube from an updated gapless genome assembly

Meng Yang, Lu Han, Shufeng Zhang, Li Dai, Bin Li, Shoukun Han, Jin Zhao, Ping Liu, Zhihui Zhao, and Mengjun Liu

Supplemental Information

Part 1. Supplemental results

1.1 Telomere to Telomere (T2T) gapless assembly of Jujube genome

The jujube genome was completely karyotyped and sequenced in this study with the tissue-cultured seedlings (Figure A1). Using the 50.7 Gb ONT ultra-long pass reads, we first preliminary assembled the genome to 65 clean contigs without microbial and organelle sequences (N50 26.3 Mb). The 65 ONT contigs, with the Hi-C data, were integrated into a 411,141,465 bp assembly with 12 anchored chromosomes and 28 singleton contigs, which account for 95.2% and 4.8% of total sequences, respectively (Figure A2). We then assembled the genome using 28.6 Gb (71×) Pacbio HiFi ccs reads with Hifiasm software and took the primary contigs, representing the complete assembly with long stretches of phased blocks, from the assembly results for further analysis. The assembly results contained a total of 83 raw primary contigs, in which 10 (12.5 Mb) and eight (1.2 Mb) were contaminations from microorganisms and organelles, respectively; after removing them, 65 contigs remained with an N50 size of 30.4 Mb. Out of the 65 contigs, 18 large ones perfectly aligned to the ONT-based assemblies (Figure A3), and after comparison, all potential gaps among the 18 contigs were filled by the ONT assembly to generate 12 telomere-to-telomere (T2T) gapless contigs, representing 12 chromosomes of jujube genome. The 12 chromosomes were further validated by Hi-C data (Figure A4) and the genetic map previously reported (Liu et al. 2014) (Figure A5). All of the remaining 47 unanchored contigs were found to be duplicated repeats, and they were covered by the 12 chromosomes (Table A1). Consequently, these contigs were excluded from the final assembly. The final assembled size is 393,332,932 bp, with an N50 length of 32.99 Mb, including 12 T2T gapless chromosomes.

We validated the completeness of the genome using BUSCO with 98.5% conserved proteins entirely detected. In continuity, 94.37% of NGS reads, 97.76% of HiFi ccs reads and 98.40% of ONT corrected reads were mapped, covering 99.45%, 99.98%, and 99.44% of the genome size, respectively (Table A2). Except for the two ends of each contig, which cannot be mapped by the algorithm, all other genome regions can be continually spanned through a combination of raw sequencing read from HiFi, ONT, and NGS. Finally, the SNPs and Indels from NGS short-reads helped estimate the base accuracy of the genome to be 99.998% (Table A3).

We further measured the genome assembly using the standards recommended by the Earth Biogenome Project (EBP) (https://www.earthbiogenome.org/assembly-standards) (Table A4). In all involved items, except the k-mer completeness, all other items reach the standard of the finished genome. The k-mer completeness were 85.11% for HiFi ccs reads, and this may due to the contamination reads, organelle reads and low-frequency reads that were discarded in the final assembly or not considered when assembling the genome. The related output data has been restored in the online share database at https://figshare.com/s/56c2299b47a5efd8708f.

The quality of this current assembly has substantially improved over the previous assembly based on NGS data (Liu et al. 2014), with a reduction of 410 folds in the number of contigs and an increase of about 6% in BUSCO completeness. The collinearity of the 12 chromosomes was generally consistent between the two versions of the genomes, however, chromosome four of the NGS assembly presented a large inversion error (Figure A6).

1.2. Taxonomic relationship between jujube and wild sour jujube.

It is generally accepted that jujube evolved from wild sour jujube; however, they are not wholly independent in phylogeny, as the semi-wild sour jujube (as a transitional type) is now widely distributed (Liu et al. 2020; Huang et al. 2016). Therefore, the taxonomic relationship between jujube and wild sour jujube regarding whether they belong to the same species has been extensively discussed. Some studies supported that they are two different species with scientific names as *Ziziphus spinosa* Hu (Tang and Eisenbrand 1992; Zhao et al. 2022) and *Ziziphus acidojujuba* Liu et Cheng (Liu et al. 2020); whereas others supported that they are the same species with scientific names as *Ziziphus jujuba* var. spinosa (Wu et al. 2022; Hua et al. 2022). The World Flora Oline attributed the scientific name to *Ziziphus jujuba* var. spinosa (Bunge) Hu ex H.F.Chow, which considered sour jujube as an infraspecific taxon of the species Ziziphus jujuba Mill. (WFO 2023).

The Ks and 4DTv analysis based on the genomic collinear region in pairwise comparison (‘Dongzao’ – ‘Junzao’, ‘Dongzao’ – ‘Suanzao’, and ‘Junzao’ – ‘Suanzao’) revealed that the values of all three peaks representing the speciation events are nearly the same and all close to the y-axis, which represented their close relationship. To make a comparison, we assembled the draft genome of *Ziziphus mauritiana* (The draft Genome has been deposited in the National Genomics Data Center under BioProject PRJCA016173), which is in the same genus as *Ziziphus* but different species from jujube. The peaks of Ks and 4DTv generated by the orthologous genes between *Z. mauritiana* and *Ziziphus jujuba* Mill. ‘Dongzao’ appeared to be more complete and occurred earlier than those of three jujube genotypes (Figure A7), supporting the conception that jujubes and wild jujubes belong to the same species.

Additional Tables for Part 1.

Table A1. Outputs of Purge_dups Software.

Chr04	28,590,891	28,624,008	OVLP	Chr01
Chr06	12,989,326	13,006,670	OVLP	Chr03
Chr07	24,197,410	24,224,351	OVLP	Chr02
Chr08	3,909,118	3,966,478	OVLP	Chr04
Chr08	44,370	230,797	OVLP	Chr07
Chr10	14,050,633	14,095,740	OVLP	Chr01
Chr10	19,590,421	19,615,719	OVLP	Chr09
Chr10	20,711,693	20,729,789	OVLP	Chr04
Chr11	19,320,147	19,338,632	OVLP	Chr06
Chr11	2,592,892	2,782,789	OVLP	Chr10
Chr11	3,304,817	3,379,353	OVLP	Chr06
Contig01	0	1,712,805	REPEAT	Chr07
Contig02	0	1,045,995	REPEAT	Chr07
Contig03	0	742,939	REPEAT	Chr07
Contig04	0	522,229	REPEAT	Chr08
Contig05	0	518,913	REPEAT	Chr07

Contig06	0	345,652	REPEAT	Chr07
Contig07	0	326,366	REPEAT	Chr07
Contig08	0	300,273	REPEAT	Chr07
Contig09	0	266,395	REPEAT	Chr08
Contig10	0	238,774	REPEAT	Chr07
Contig11	0	193,878	REPEAT	Chr07
Contig12	0	181,665	REPEAT	Chr07
Contig13	0	166,833	REPEAT	Chr07
Contig14	0	164,191	REPEAT	Chr07
Contig15	0	162,000	REPEAT	Chr07
Contig16	0	139,146	REPEAT	Chr08
Contig17	0	133,196	REPEAT	Chr07
Contig18	0	132,357	REPEAT	Chr08
Contig19	0	118,076	REPEAT	Chr07
Contig20	0	114,639	REPEAT	Chr07
Contig21	0	113,031	REPEAT	Chr07
Contig22	0	111,589	REPEAT	Chr07
Contig23	0	111,542	REPEAT	Chr08
Contig24	0	110,135	REPEAT	Chr07
Contig25	0	104,018	REPEAT	Chr07
Contig26	0	94,367	REPEAT	Chr07
Contig27	0	91,541	REPEAT	Chr07
Contig28	0	69,405	REPEAT	Chr08
Contig29	0	67,775	REPEAT	Chr07
Contig30	0	62,946	REPEAT	Chr08
Contig31	0	57,407	REPEAT	Chr07
Contig32	0	57,020	REPEAT	Chr07
Contig33	0	56,519	REPEAT	Chr07
Contig34	0	56,462	REPEAT	Chr07
Contig35	0	54,660	REPEAT	Chr07
Contig36	0	50,598	REPEAT	Chr12
Contig37	0	45,514	REPEAT	Chr06
Contig38	0	41,467	REPEAT	Chr07
Contig39	0	40,313	REPEAT	Chr08
Contig40	0	40,312	REPEAT	Chr07
Contig41	0	38,695	REPEAT	Chr08
Contig42	0	38,080	REPEAT	Chr07
Contig43	0	37,710	REPEAT	Chr08
Contig44	0	36,009	REPEAT	Chr08
Contig45	1	30,088	JUNK	
Contig46	0	27,755	REPEAT	Chr08
Contig47	0	21,885	REPEAT	Chr07

Table A2. Reads mapping statistics to contig assembly.

Platform	Total Reads	Map Reads	Map Rate	Covered genome
MGI-SEQ	147,060,360	138,776,269	94.37%	99.45%
HiFi	1,631,748	1,595,120	97.76%	99.98%
ONT	94,304	92,793	98.40%	99.44%

Table A3. Genome accuracy evaluation by NGS reads.

Depth	Hetero SNP	Hetero Indel	Homo SNP	Error rate by Homo SNP(%)	Homo Indel	Error rate by Homo Indel(%)	Error rate by homo variants(%)	Accuracy genome(%)
depth>=1x	2,277,973	326,926	5,100	0.001211	9,650	0.002292	0.003503	99.996497
depth>=5x	2,277,611	326,209	3,694	0.000877	8,002	0.0019	0.002778	99.997222
depth>=10x	2,274,785	321,582	2,578	0.000612	5,616	0.001334	0.001946	99.998054

Table A4. Assembly evalution using the recommended approaches by EBP.

Quality Category	Quality Metric	Value	Standard	Software used
Continuity	Contig (NG50)	32.99 Mb	Finished	In house scripts
	Scaffolds (NG50)	32.99 Mb	Finished	
	Gaps/Gbp	0%	Finished	
Structural accuracy	False duplication	0%	Finished	Purge_Dups and Asset
	Reliable blocks	32.99 Mb	Finished	
	Curation improvements	All conflict resolved	Finished	
Base Accuracy	Base pair QV	64	Finished	Merqury
	k-mer completeness	85.11%	4.5.Q30	
Functional completeness	Genes	98.50%	Finished	BUSCO
	Transcript mappability	100%	Finished	STAR and samtools
Chromosome status	% Assigned	100%	Finished	In house scripts
	Organelles	Complete	Finished	

Additional Figures for Part 1.

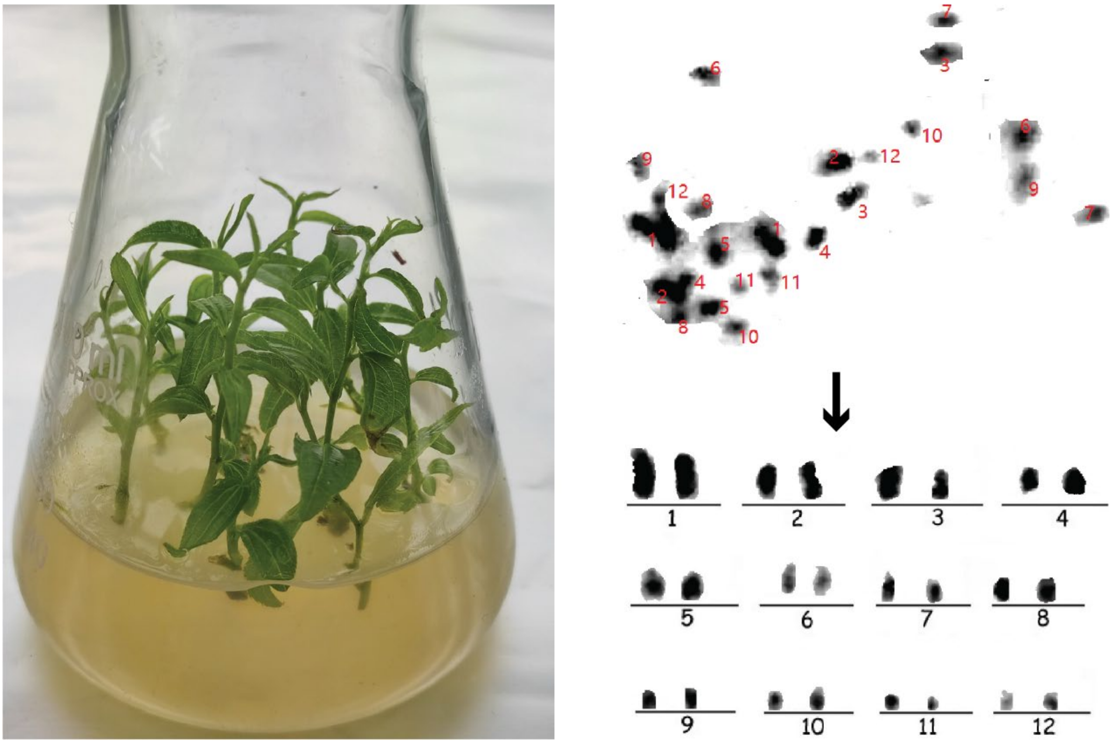


Figure A1. Tissue-cultured sample of ‘Dongzao’ jujube and karyogram for 12 chromosomes.

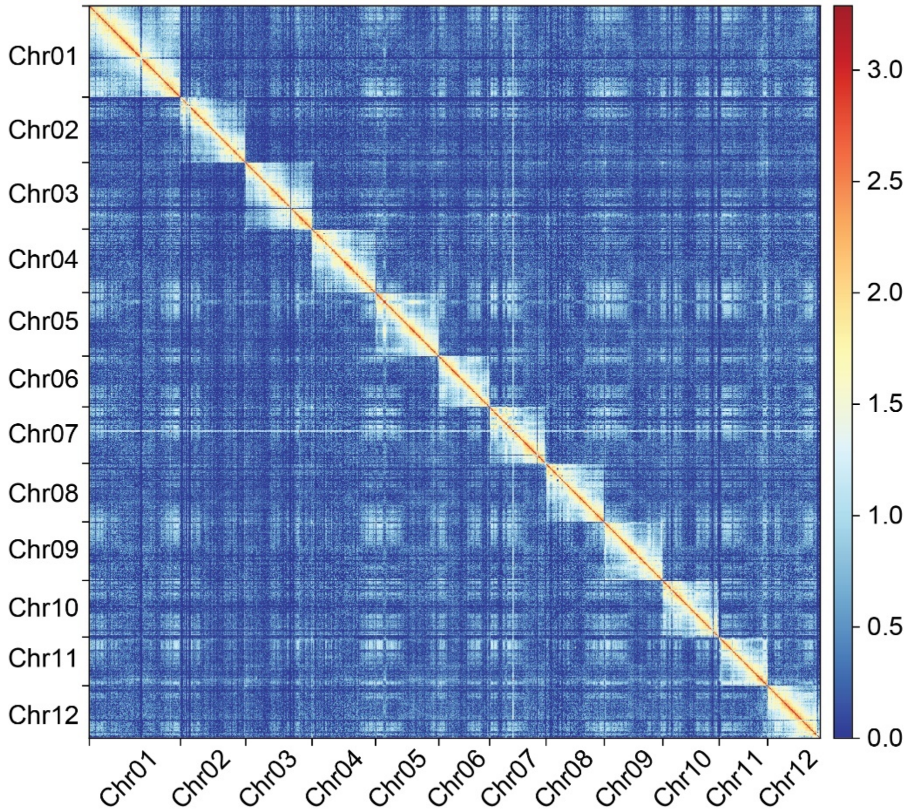


Figure A2. Hi-C chromatin interaction map for the ONT assembly in 100 kb resolution.

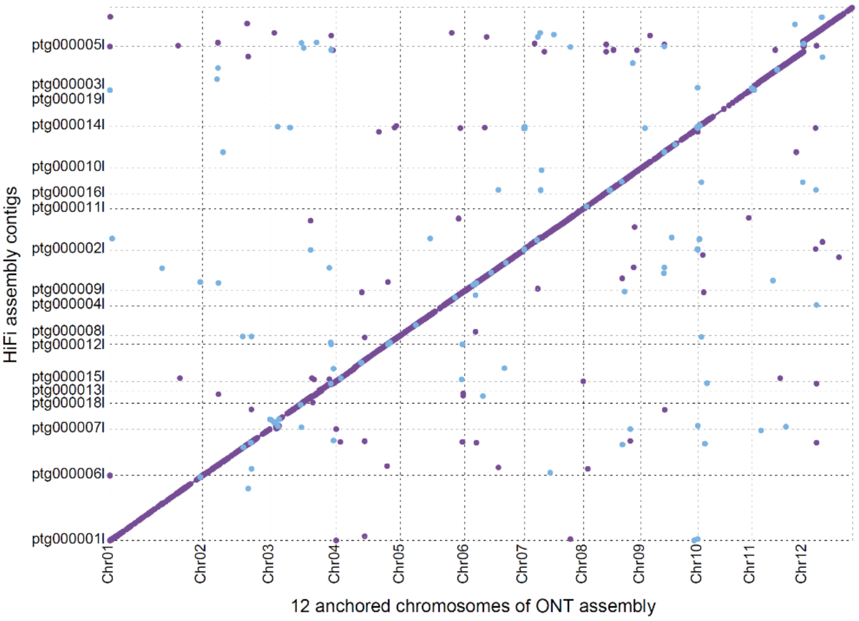


Figure A3. Mummer plot between ONT-based assembly and HiFi-based contigs. The x-axis represents 12 Hi-C-anchored chromosomes of ONT assembly, and the y-axis represents the 18 optimally aligned HiFi contigs.

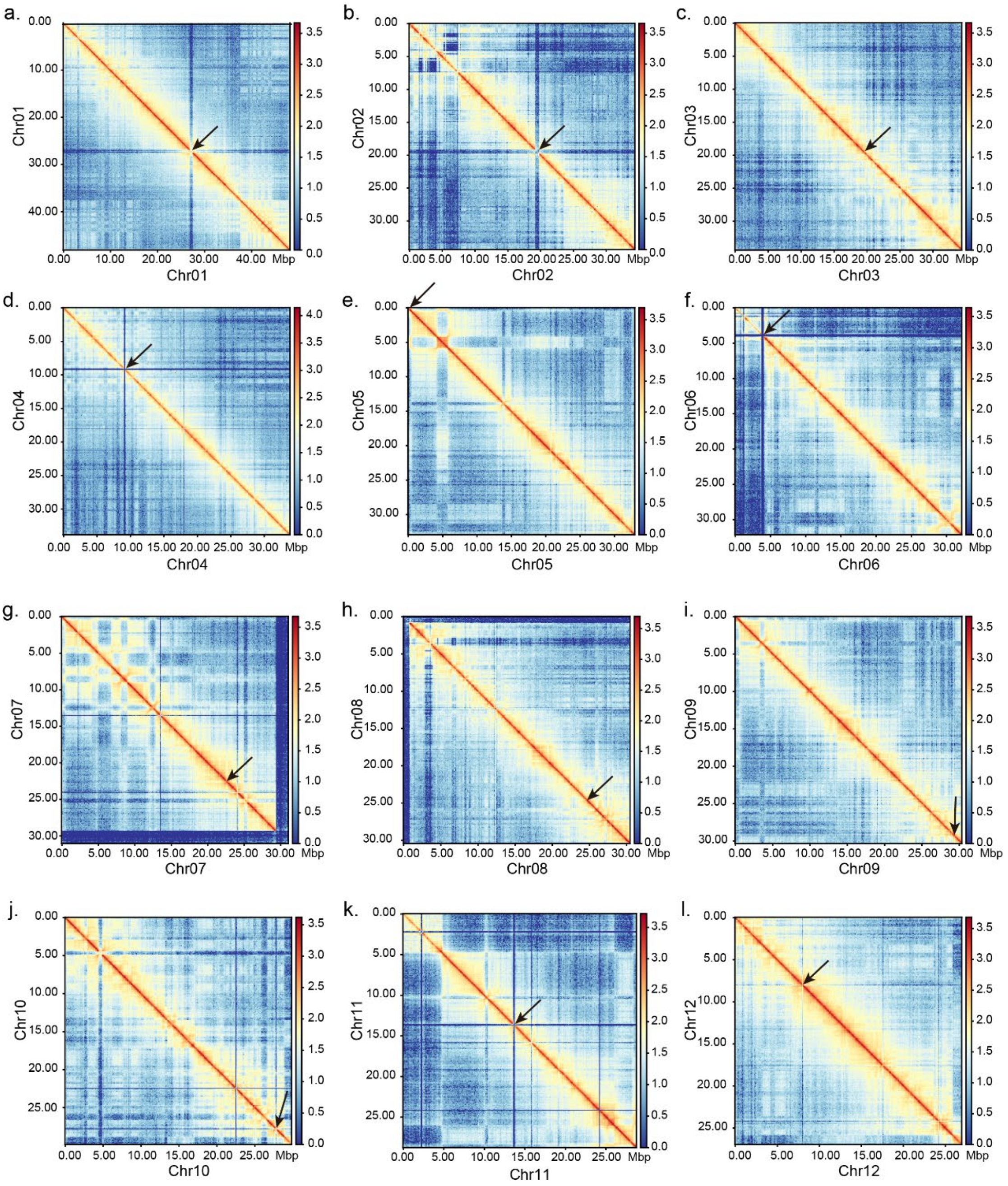


Figure A4. Hi-C intra-chromatin interaction map of the final HiFi assembly in 100 kb resolution. The black arrows represent the putative position of centromeres.

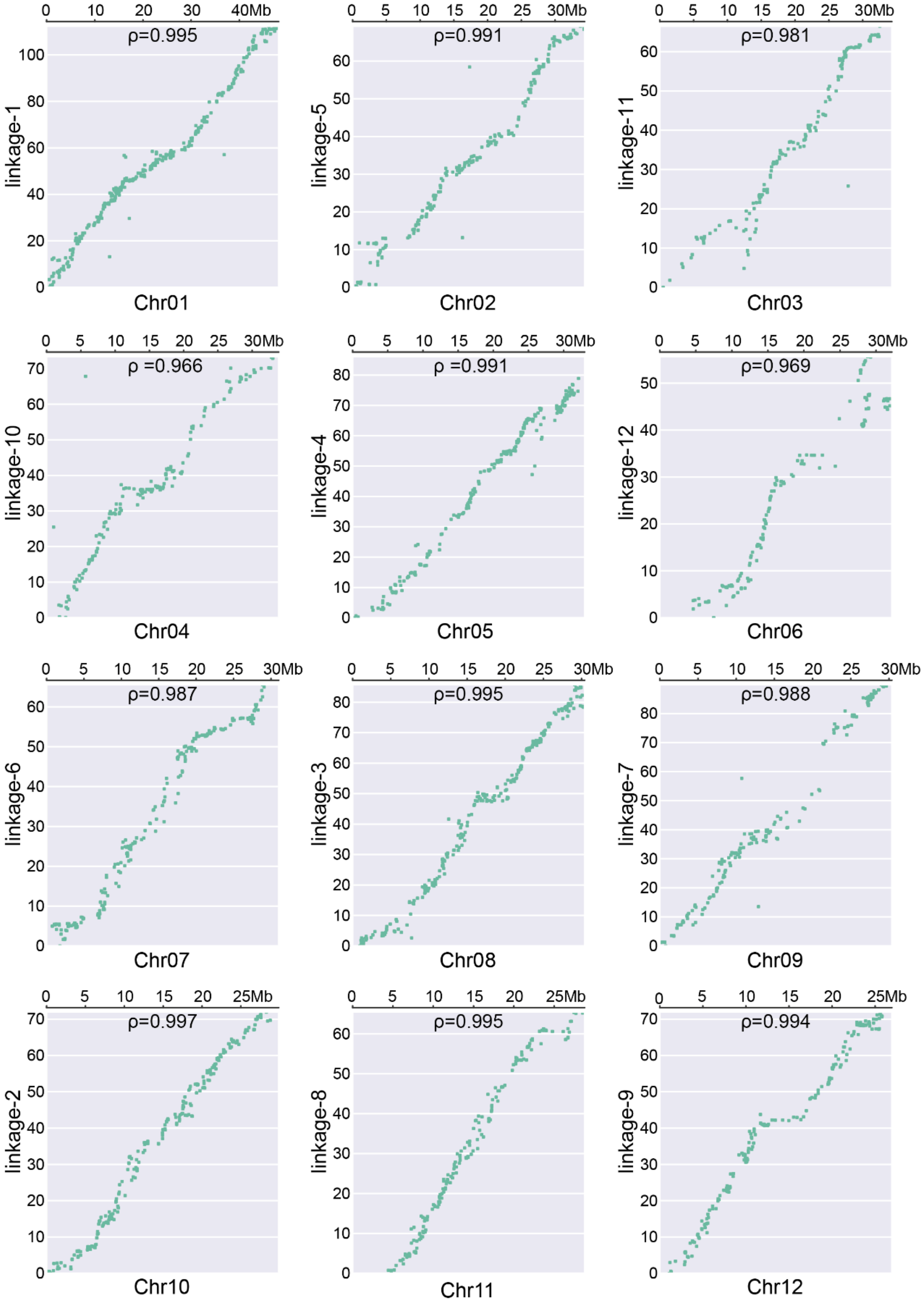


Figure A5. Collinearity between the jujube high-density genetic map and the corresponding chromosome assembly.

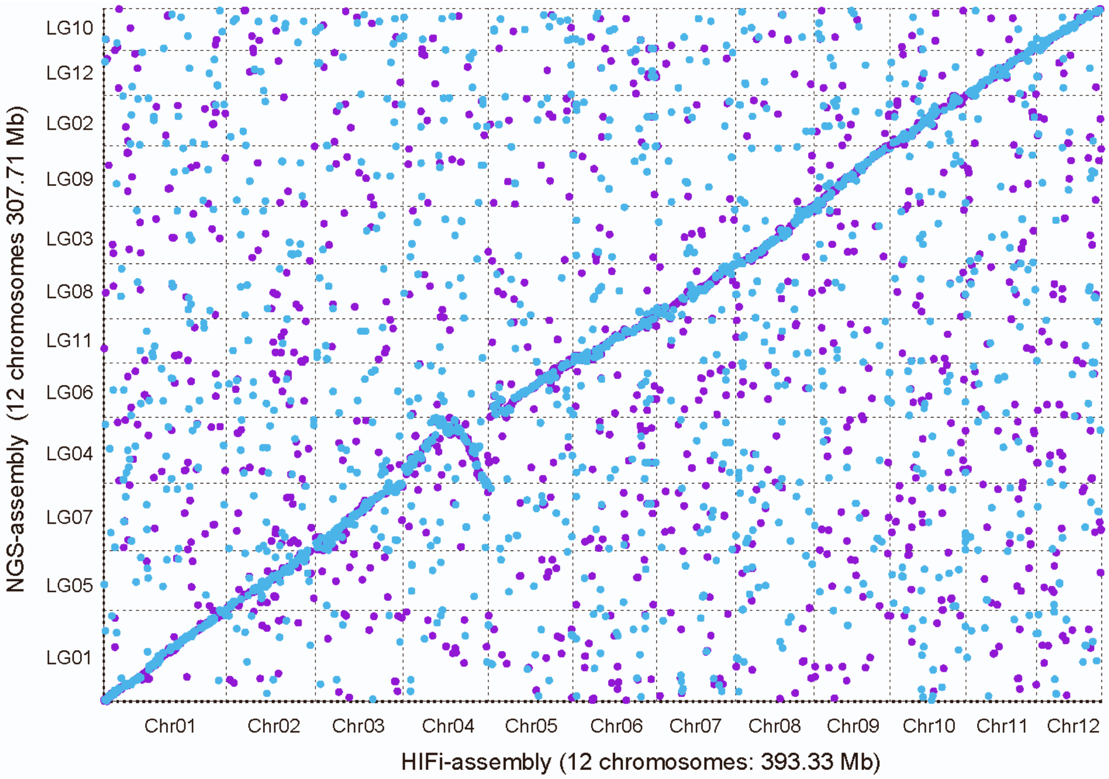


Figure A6. Genome-wide collinear comparison of HiFi assembly and the NGS assembly of *Ziziphus jujuba* Mill. ‘Dongzao’ by using Mummer software. Only the best hits were kept in the plot.

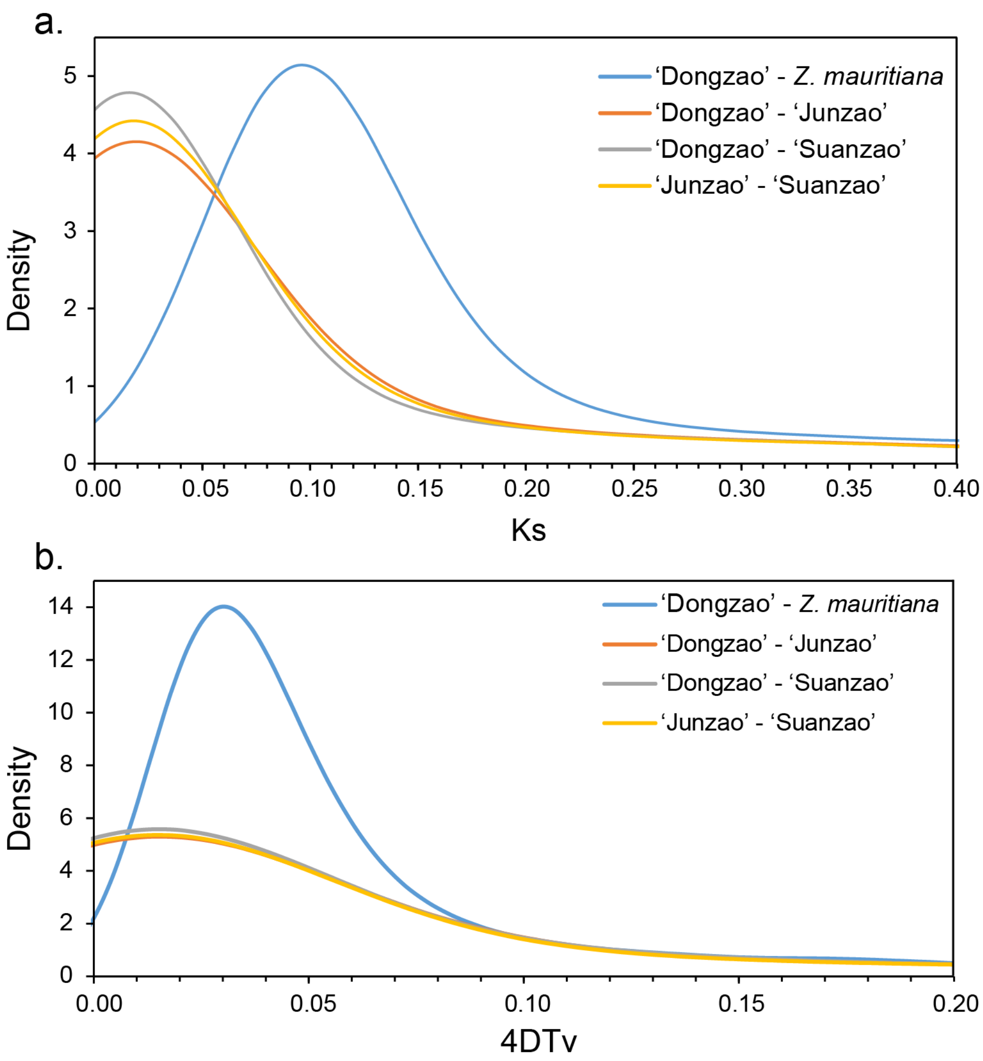


Figure A7. Speciation event based on Ks and 4DTv among three jujube genotypes as well as between 'Dongzao' and *Ziziphus mauritiana*. (a) Ks; (b) 4DTv.

Part 2. Materials and methods

2.1 Sample and karyotype

Young seedlings were obtained from the two-month *Z. jujuba* Mill. 'Dongzao' tissue culture plantlets were cultured at 25℃. To observe and confirm the karyotype of chromosomes, the stem apex of the tissue culture plantlet was pretreated with 0.7 mM colchicine at an average room temperature of 25℃ for 12 h, washed with distilled water, immersed in 0.075 mol/L hypotonic KCl solution at 4°C for 1 h, and then transferred to Carnot fixator for 12 h. Subsequently, the fixed stem apex was thoroughly cleaned with distilled water and stained with carbol fuchsin. Finally, we transferred the dyed and softened materials to a glass slide and observed the karyotype under an oil microscope (Olympus BX51TF).

2.2 Short-read sequencing and quality control

Using the modified cetyltrimethylammonium bromide (CTAB) technique, total genomic DNA was isolated (Murray and Thompson 1980). The DNA purity was evaluated with a NanoDropTM One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), and the DNA integrity was confirmed by agarose gel electrophoresis. The DNA was utilized to generate a paired-end library with an insert size of 200-400 bp on the MGISEQ-2000 technology (BGI, Shenzhen, Guangdong, China). These short reads were created to assess the genome's size and heterozygosity and to correct the Long-reads' preliminary assembly of the genome. Raw readings were filtered by the fastp (v.0.20.0) preprocessor with default parameters to remove low-quality reads, adapters, and poly-N-containing reads before quality control (Chen et al. 2018). The following criteria were applied to discard reads: (1) 10% unidentified nucleotides (N); (2) > 10 nucleotides aligned to the adaptor; (3) the length of bases with Phred quality of 5 in a read longer than >50% of the read length; and (4) with PCR duplicated reads (read 1 and read 2 of two paired-end reads are completely identical). To confirm the absence of contamination, 100,000 random readings were compared to the NCBI nt database.

In addition, a Hi-C library was constructed and sequenced to facilitate chromosome-level genome assembly. Approximately 2 g of fresh leaves were utilized for library construction, and the technique involved formalin fixation, crosslinking, nuclei suspension, digestion, DNA ligation, end-repair, purification, and quantification, as previously described (Belton et al. 2012). The qualifying library was subsequently sequenced on an MGI-2000 platform. The quality control measures were identical to those described above for paired-end sequencing.

2.3 Genome size and heterozygosity Estimation

KMC software (Kokot et al. 2017) was used to generate the k-mer counts (k=21) from the cleaned short reads, and GenomeScope2 software (Ranallo-Benavidez et al. 2020) subsequently used these k-mers to estimate the genome size and heterozygosity.

2.4 Long reads sequencing by Pacbio HiFi and Oxford Nanopore

To prepare DNA for the long-read sequencing, high molecular weight genomic DNA was extracted using the SDS technique and purified using the QIAGEN® Genomic kit (Cat#13343, QIAGEN). On 1% agarose gels, the level of DNA degradation and contamination was evaluated. Using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), the OD 260/280 and 260/230 ranges were determined to be between 1.8 and 2.0 and 2.0 and 2.2, respectively.

Pacbio HiFi SMRTbell libraries were constructed following the standard protocol using the SMRTbell Express Template Prep Kit 2.0 (PacBio, CA, USA). The lengthy DNA fragments were skillfully sheared down to 15-18 kb using a g-TUBE (Covaris, MA, USA). Single-strand overhangs were cut off and damaged and broken DNA was patched up with the chemicals in the Template Prep Kit. Once the ends were fixed, SMRTbell hairpin adapters were ligated to them, and then the libraries were concentrated and purified using AMPure PB beads (PacBio, CA, USA). BluePippin was utilized to size-select SMRTbell templates more than 15 kb to get large-insert SMRTbell libraries for sequencing (SageScience, MA, USA). The sequencing was performed using a PacBio Sequel II device with Sequencing Primer V2 and a Sequel

II Binding Kit 2.0. For the raw sequencing reads, the min passes = 3 and min RQ = 0.99 default parameters in CCS software (https://github.com/PacificBiosciences/ccs) were utilized to generate high-precision HiFi reads with quality over Q20.

For ONT sequencing, the NEBNext Ultra II End Repair/dA-tailing Kit was used to fix the ends of the lengthy DNA fragments that were size-selected with the BluePippin system (Sage Science, USA) (Catalog number E7546). The fragment size of the library was then measured with a Qubit® 3.0 Fluorometer after a second ligation reaction was performed with an LSK109 kit (Invitrogen, USA). Library sequencing was performed on Nanopore PromethION instruments (ONT: Oxford Nanopore Technologies, UK). The raw information is presented as FAST5 binary signal data. We utilized a high-precision flip-flop model with the guppy basecaller command in the GPU-enabled Guppy program (v3.4.4) to collect the fastq data. Reads with Q scores greater than 7 were considered passed after the raw data in fastq format had been analyzed for base quality.

2.5 Genome assembly and evaluation

The ONT and HiFi sequencing reads were initially utilized to create ONT-based and Hifi-based assemblies, respectively. To generate the ONT-based assembly, the ONT passed reads were *de novo* assembled using NextDenovo (V2.3.1). Subsequently, the assembly was refined with the ONT-passed long reads using Racon (Vaser et al. 2017) and curated using the paired-end short reads with Nextpolish (Hu et al. 2020). Contaminations including microbial and organelle sequences were removed by aligning to the NCBI nt database. The cleaned Hi-C reads were then used to anchor the ONT contigs into chromosomes. First, unique mapped reads were recognized with bowtie2 (v2.3.2) (Langmead and Salzberg 2012), followed by the identification of paired reads with valid interaction using HiC-Pro (v2.8.1) (Servant et al. 2015). These valid pairs of reads were next applied to build the pseudo-chromosome sequences by LACHESIS (Burton et al. 2013), with the following key parameters:

CLUSTER_MIN_RE_SITES=100

CLUSTER_MAX_LINK_DENSITY=2.5

CLUSTER_NONINFORMATIVE_RATIO = 1.4

ORDER_MIN_N_RES_IN_TRUNK=60

ORDER_MIN_N_RES_IN_SHREDS=60

The whole process involves clustering, ordering, and orientation according to the interaction relationship of Hi-C reads, accompany by the manual operation to adjust the position and orientation of discrete contigs based on the chromatin interaction patterns. Finally, all positioned contigs were linked by 100 bp Ns to generate the pseudo-chromosomes.

The HiFi ccs readings were assembled with Hifiasm (v0.16.1-r375) (Cheng et al. 2021) using the default parameters to construct the Hifi-based assembly. Because jujube is typically propagated vegetatively through grafting, and the parents are unavailable, properly assembling the genome into two haplotypes is difficult. So, for the following analysis, we chose Hifiasm's primary assembly, which is a complete assembly with extensive stretches of phased blocks (https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies). The primary contigs were aligned with the NCBI nt database to exclude microbial and organelle contaminations. Using mummer (v4.0.0rc1) (Marcais et al. 2018), the cleaned primary contigs were then directed to the final telomere-to-telomere gapless assembly by comparison to the above ONT assembly.

The completeness of the jujube genome assembly was assessed using the embryophyta_odb10 of BUSCO v4.0.5 (Simao et al. 2015). To evaluate the base accuracy, BWA (Li and Durbin 2010) and minimap2 (Li 2018) were used to respectively align the short paired-end reads and the long HiFi/ ONT reads to the assembled genome, and the results were interpreted by SAMtools (Li et al. 2009) for mapping rate, base accuracy, as well as genome coverage of the short reads.

2.6 RNA sequencing and data analysis

RNA was collected from the same sample as DNA using a plant RNA isolation kit (Tiangen Biotechnology Co.). Following the manufacturer's instructions, sequencing libraries were created using the TruSeq RNA Library Preparation Kit (Illumina, United States). Brief procedures include mRNA purification using oligo poly-T probes, cDNA synthesis, adaptor ligation, size selection and purification, PCR, PCR product purification, and library quality evaluation. Finally, the library was sequenced on an Illumina Novaseq platform to obtain 150 bp paired-end reads.

The raw paired-end RNA-seq reads were first performed for quality control using fastp (Chen et al. 2018). Then the clean reads were mapped to the jujube genome using STAR (v2.7.10) with the default parameters (Dobin and Gingeras 2015b). The result BAM file was used as the input to the RSEM software (Li and Dewey 2011) to calculate the expression level for each transcript using the fragments per kilobase of exon per million mapped reads (FPKM).

2.7 PacBio full-length cDNA sequencing

Total RNA was extracted from the same sample containing DNA by grinding tissue using the CTAB-LiCl technique on dry ice. Agilent 2100 Bioanalyzer (Agilent Technologies) and agarose gel electrophoresis were used to assess the RNA's integrity. Only high-quality RNA (OD260/280: 1.82.2, OD260/230: 2.0, RIN: 8; quantity: >1 g) was utilized to create the sequencing library. Approximately 300 ng of RNA was reverse-transcribed into cDNA and amplified with the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module and Iso-Seq Express Oligo Kit. Using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences), the library was produced by damage repair, end repair, A-tailing, and adapters ligation. Finally, the SMRTbell template was annealed to the sequencing primer, bound to polymerase, and sequenced using Sequel II Binding Kit 2.0 on the PacBio Sequel II platform (Pacific Biosciences).

2.8 Genome annotation

The interspersed repetitions were discovered using *ab initio* and homology-based methods. Briefly, an *ab initio* species-specific repeat library was prepared using RepeatModeler (Price et al. 2005); this library and the Repbase database (http://www.girinst.org/repeatbase) were used as the inputs to RepeatMasker software (Chen 2004) to search for repetitions at the whole genome-level. Subsequently, the entire long terminal repeat retrotransposons (LTR-RTs) were detected by LTR FINDER (Xu and Wang 2007) and ltr_harverst (Ellinghaus et al. 2008), followed by the integration using LTR_retriver (Ou and Jiang 2018), which included the search for false positives, terminal motifs, and transposon protein domains. Finally, the intact LTR insertion time was computed using the formula:

$T=K/2\mu$,

where K is the divergence rate estimated by the identity of LTRs using the baseml model of PAML, and μ is the neutral mutation rate denoting mutations per bp per year, using a value of 7.77×10^{-9} as proposed in peaches (Xie et al. 2016).

The protein-coding genes were predicted by combining protein homology, transcriptome, and *ab initio* approach. The homologous proteins of related species, including *Malus domestica*, *Arabidopsis thaliana*, *P. trichocarpa*, *Prunus persica*, *Prunus armeniaca*, and *Pyrus pyrifolia*, were aligned using GeMoMa (Keilwagen et al. 2019). In transcriptome-based prediction, RNA-seq reads were mapped to the genome with STAR (Dobin and Gingeras 2015a), and the mapping information was passed to string tie (Pertea et al. 2015) to assemble the transcripts. Subsequently, the transcripts and full-length PacBio cDNA were imported to PASA (Haas et al. 2003) to obtain the prediction. *Ab initio* gene prediction was performed by importing the string tie transcripts to Augustus (Stanke et al. 2006) to generate a training set suited for the jujube genome using the default parameters. The final gene prediction was accomplished with EVidenceModeler (EVM) (Haas et al. 2008) by merging the prediction findings of the aforementioned approaches, followed by a comparison to the genome to eliminate genes wholly located in repetitive regions. The transposon genes were further filtered using the TransposonPSI software (https://github.com/NBISweden/TransposonPSI). Finally, the function of proteins was annotated using the Interproscan (v5.57-90.0) (Jones et al. 2014) as well as eggNOG-mapper (v2.1.6) (Cantalapiedra et al. 2021); the GO function and KEGG pathway information were extracted from the former and the latter, respectively.

Homology search or *ab initio* prediction was also used to identify the non-coding RNAs (ncRNAs), including transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and microRNAs (miRNAs). The tRNAscan-SE program was used to identify tRNAs (Lowe and Eddy 1997). MiRNA and other non-coding RNAs were identified by searching the Rfam database with Infernal (http://infernal.janelia.org). The rRNAs and their subunits were predicted using the default parameters of RNAmmer (https://github.com/tseemann/barrnap).

2.9 Identification of telomeres and centromeres

The telomere-specific motif "CCCTAAA" and "TTTAGGG" were used to locate the telomeres. Two approaches were adopted to identify the centromere sequences: (1) Tandem Repeats Finder (TRF) (Benson 1999) was utilized to identify the abundant top repeats, and the results with core repeat unit >50 bp and at least repeated 20 times were retained; (2) Using the method in *Arabidopsis*, wherein the periodic 12-mer in the 1-kb windows was identified from the genome assembly to determine the telomere sequences (Naish et al. 2021).

2.10 Whole-genome bisulfite sequencing

Total genomic and control unmethylated lambda DNA were combined to a volume of 80 l using 1× TE buffer and fragmented to 300 bp. The dA-tailed fragment was ligated with methylated adaptors following end repair and dA-tailing. The ligated DNA was then treated with bisulfite and amplified using the uracil-binding pocket of KAPA HiFi DNA Polymerase. Finally, the library was quantified and sequenced as paired-end 150-bp reads on an Illumina Hiseq X10 sequencer (Illumina, Inc.). Raw sequencing data were curated by removing adaptor-polluted reads, low-quality reads, and reads with over 10% Ns. Subsequently, clean reads were mapped to the jujube genome using Bismark (V0.23.1) (Krueger and Andrews 2011), and only uniquely mapped reads were retained. Methylated cytosines were identified based on the binomial test followed by Benjamini–Hochberg false discovery rate correction.

2.11 Comparative genomics analysis

Genome collinear region was identified using the MCScanX (Wang et al. 2012) with e-value 1e-05, and nonsynonymous and synonymous substitution rates (Ka and Ks) of collinearity genes were computed using KaKs_Calculator with the NG module (Zhang et al. 2006). The fourfold synonymous third-codon transversion rates (4DTv) were calculated by using calculate_4DTV_correction.pl (https://github.com/JinfengChen/Scripts/blob/master/FFgenome/03. evolution/distance_kaks_4dtv/bin/calculate_4DTV_correction.pl). The sequence similarity of paralogs was obtained from the pairwise alignment of paralogs using BLASTN.

2.12 3D chromosomes interaction analysis

Singleton-, multi-mapped-, and duplicated- reads were removed through HIC-Pro (v3.0.0) (Servant et al. 2015) and uniquely mapped reads were retained to generate the interaction matrix. Three bin sizes of 100 kb, 50 kb, and 10 kb from HiC-Pro were utilized to generate the contact matrix files for matrix plotting, A/B compartments, and TADs analysis. The Hi-C contact map was generated using HiCexplorer (v3.7.2) (Wolff et al. 2020) utilizing the 100 kb bin matrix file. To quantify the interactions between pairs of chromosomes (Figure 1f), the number (N) in each square lattice is N=1000*S/L, where S is the sum of all contacted reads pairs between two chromosomes, and L is the sum of the length of two chromosomes. Principal component analysis was performed on a 50 Kb matrix file, and the positive and negative values of the first eigenvector were used to define the A and B compartments, respectively, using Cworld (V0.0.1) (https://github.com/dekkerlab/cworld-dekker). TopDom (V0.0.2) (Shin et al. 2016) was applied to identify the TADs using the 10-Kb matrix file.

Supplemental References

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58 (3):268-276

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27 (2):573-580

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 31 (12):1119-1125

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* 38 (12):5825-5829

Chen N (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4:Unit 4 10

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17):i884-i890

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18 (2):170-175

Dobin A, Gingeras TR (2015a) Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* 51:11 14 11-11 14 19

Dobin A, Gingeras TR (2015b) Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics* 51 (1):11.14. 11-11.14. 19

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the

Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31 (19):5654-5666

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9 (1):R7

Hu J, Fan J, Sun Z, Liu S (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36 (7):2253-2255

Hua Y, Xu X-x, Guo S, Xie H, Yan H, Ma X-f, Niu Y, Duan J-A (2022) Wild Jujube (Ziziphus jujuba var. spinosa): A Review of Its Phytonutrients, Health Benefits, Metabolism, and Applications. *Journal of Agricultural and Food Chemistry* 70 (26):7871-7886

Huang J, Zhang C, Zhao X, Fei Z, Wan K, Zhang Z, Pang X, Yin X, Bai Y, Sun X, Gao L, Li R, Zhang J, Li X (2016) The Jujube Genome Provides Insights into Genome Evolution and the Domestication of Sweetness/Acidity Taste in Fruit Trees. *PLoS genetics* 12 (12):e1006433

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30 (9):1236-1240

Keilwagen J, Hartung F, Grau J (2019) GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods Mol Biol* 1962:161-177

Kokot M, Długosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33 (17):2759-2761

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27 (11):1571-1572

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9 (4):357-359

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12 (1):1-16

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18):3094-3100

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26 (5):589-595

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25 (16):2078-2079

Liu M, Wang J, Wang L, Liu P, Zhao J, Zhao Z, Yao S, Stanica F, Liu Z, Wang L, Ao C, Dai L, Li X, Zhao X, Jia C (2020) The historical and current research progress on jujube-a superfruit for the future. *Hortic Res* 7:119

Liu MJ, Zhao J, Cai QL, Liu GC, Wang JR, Zhao ZH, Liu P, Dai L, Yan G, Wang WJ, Li XS, Chen Y, Sun YD, Liu ZG, Lin MJ, Xiao J, Chen YY, Li XF, Wu B, Ma Y, Jian JB, Yang W, Yuan Z, Sun XC, Wei YL, Yu LL, Zhang C, Liao SG, He RJ, Guang XM, Wang Z, Zhang YY, Luo LH (2014) The complex jujube genome provides insights into fruit tree biology. *Nat Commun* 5:5315

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25 (5):955-964

Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* 14 (1):e1005944

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* 8 (19):4321-4325

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova T, Jamge B, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K, Smith LM, Ton J, Kakutani T, Martienssen RA, Schneeberger K, Lysak MA, Berger F, Bousios A, Michael TP, Schatz MC, Henderson IR (2021) The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* 374 (6569):eabi7489

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* 176 (2):1410-1422

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33 (3):290-295

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1:i351-358

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11 (1):1432

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 16:259

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ (2016) TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research* 44 (7):e70-e70

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31 (19):3210-3212

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34 (Web Server issue):W435-439

Tang W, Eisenbrand G (1992) Ziziphus jujuba Mill. and Z. spinosa Hu. In: Tang W, Eisenbrand G (eds) Chinese Drugs of Plant Origin: Chemistry, Pharmacology, and Use in Traditional and Modern Medicine. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1017-1024. doi:10.1007/978-3-642-73739-8_124

Vaser R, Sovic I, Nagarajan N, Sikic M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27 (5):737-746

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* 40 (7):e49-e49

WFO (2023) Ziziphus jujuba var. spinosa (Bunge) Hu ex H.F.Chow. *Published on the Internet; http://wwwworldfloraonlineorg/taxon/wfo-0000742370*

Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, Grüning BA (2020) Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Research* 48 (W1):W177-W184

Wu M, Gu X, Zhang Z, Si M, Zhang Y, Tian W, Ma D (2022) The effects of climate change on the quality of Ziziphus jujuba var. Spinosa in China. *Ecological Indicators* 139:108934

Xie Z, Wang L, Wang L, Wang Z, Lu Z, Tian D, Yang S, Hurst LD (2016) Mutation rate analysis via parent-progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proc Biol Sci* 283 (1841)

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35 (Web Server issue):W265-268

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4 (4):259-263

Zhao Q, Mi ZY, Lu C, Zhang XF, Chen LJ, Wang SQ, Niu JF, Wang ZZ (2022) Predicting potential distribution of Ziziphus spinosa (Bunge) H.H. Hu ex F.H. Chen in China under climate change scenarios. *Ecol Evol* 12 (2):e8629

Part 3. Supplemental Tables and Figures

Supplemental Tables

Supplemental Table 1. Statistics of reads information.

MGISEQ-2000 reads	total reads	147,060,360
	total bases	22,059,054,000 bp
	clean reads	146,937,342
	clean bases	20,549,591,610 bp
	Q20 rate	97.27%
	Q30 rate	92.77%
	GC content	35.95%
HiFi reads	subreads number	28,180,197
	subreads bases	428,936,355,668 bp
	ccs reads Num	1,631,748
	ccs bases	28,621,088,987 bp
	ccs reads N50	18,732 bp
	ccs reads mean length	17,540 bp
	ccs longest read	49,987 bp
	ccs rate	6.67%
ONT passed reads (Q>7)	Total reads	1,503,196
	Total bases	50,689,713,213 bp
	reads N50 length	52,708 bp
	reads mean length	33,721 bp
	maximum read length	703,323
Hi-C reads	Number of raw read pairs	144,695,164
	Number of raw bases (bp)	43,408,549,000
	Number of clean read paris	143,508,269
	Number of clean bases (bp)	40,114,785,000
	Clean reads rate (%)	99.18
	Clean bases rate (%)	92.41

Supplemental Table 2. BUSCO evaluation results for genome and genes.

Classification	Genome	Protein-coding genes
Complete BUSCOs (C)	1590 (98.5%)	1512 (93.7%)
Complete and single-copy BUSCOs (S)	1569 (97.2%)	1491 (92.4%)
Complete and duplicated BUSCOs (D)	21 (1.3%)	21 (1.3%)
Fragmented BUSCOs (F)	14 (0.9%)	27 (1.7%)
Missing BUSCOs (M)	10 (0.6%)	74 (4.6%)
Total BUSCO groups searched	1614 (100%)	1614 (100%)

Supplemental Table 3. Repeat information of ‘Dongzao’ genome.

Class	Super family	Family	Number of family members	Length of sequence (bp)	Ratio in the genome (%)	Average length (bp)
Class I			**153,352**	**114,349,303**	**29.07**	**745.67**
	LINE		**16,049**	**4,298,506**	**1.09**	**267.84**
		L1	7,140	3,118,613	0.79	436.78
		L2	3,126	393,751	0.10	125.96
		CR1	659	157,028	0.04	238.28
		Penelope	1,953	160,212	0.04	82.03
		RTE	600	138,985	0.04	231.64
		Tad1	314	71,472	0.02	227.62
		Other	2,257	258,445	0.07	114.51
	LTR		**135,896**	**109,668,261**	**27.88**	**807.00**
		Gypsy	74,326	67,362,501	17.13	906.31
		Copia	40,271	35,511,237	9.03	881.81
		Cassandra	8,327	3,184,342	0.81	382.41
		Caulimovirus	1,771	1,380,860	0.35	779.71
		Pao	1,987	406,117	0.10	204.39
		ERV1	1,472	61,477	0.02	41.76
		Ngaro	555	92,537	0.02	166.73
		Other	984	43,706	0.01	44.42
		Unknown	6,203	1,625,484	0.41	262.05
	SINE		**1,407**	**382,536**	**0.10**	**271.88**
		tRNA-Deu	597	52,716	0.01	88.30
		Other	406	23,964	0.01	59.02
		Uknown	404	305,856	0.08	757.07
Class II			**97,376**	**33,440,673**	**8.50**	**343.42**
	DNA		**86,175**	**29,403,433**	**7.48**	**341.21**
		MULE-MuDR	14,422	8,985,025	2.28	623.01
		CMC-EnSpm	22,713	8,348,069	2.12	367.55
		hAT	18,645	6,562,279	1.67	351.96
		PIF-Harbinger	7,581	2,202,724	0.56	290.56
		Maverick	1,628	744,416	0.19	457.26
		Zisupton	3,720	729,522	0.19	196.11
		TcMar	2,221	557,404	0.14	250.97
		Crypton-V	1,250	112,003	0.03	89.60
		CMC-Transib	802	57,431	0.01	71.61
		Other	8,927	411,949	0.10	46.15
		Unknown	4,266	692,611	0.18	162.36
	Rolling cirlces		**11,201**	**4,037,240**	**1.03**	**360.44**
		Helitron	11,190	4,036,790	1.03	360.75
		Other	11	450	0.00	40.91
Total TEs			**250,728**	**147,789,976**	**37.57**	**589.44**
Unknown			207,507	58,742,266	14.93	283.09

Low_complexity	40,918	1,927,480	0.49	47.11
Satellite	648	84,893	0.02	131.01
Simple_repeat	253,683	9,209,302	2.34	36.30
Small RNAs	3,374	3,123,189	0.79	925.66
Total Repeats	**756,858**	**220,877,106**	**56.16**	**291.83**

Supplemental Table 4. Telomere information of the genome.

Chromosome	Left start	Left end	Left length	Left motif	Right start	Right end	Right length	Right motif
Chr01	1	19,630	19,630	CCCTAAA	48159139	48,169,259	10,121	TTTAGGG
Chr02	1	7,982	7,982	CCCTAAA	34364548	34,383,186	18,639	TTTAGGG
Chr03	1	15,603	15,603	CCCTAAA	34303592	34,311,561	7,970	TTTAGGG
Chr04	1	4,217	4,217	CCCTAAA	33864166	33,874,294	10,129	TTTAGGG
Chr05	1	22,400	22,400	CCCTAAA	32969164	32,986,920	17,757	TTTAGGG
Chr06	1	21,972	21,972	CCCTAAA	32188140	32,197,620	9,481	TTTAGGG
Chr07	1	19,321	19,321	CCCTAAA	31014812	31,029,268	14,457	TTTAGGG
Chr08	1	7,232	7,232	CCCTAAA	30406786	30,413,063	6,278	TTTAGGG
Chr09	1	14,481	14,481	CCCTAAA	30355812	30,376,005	20,194	TTTAGGG
Chr10	1	7,001	7,001	CCCTAAA	29851313	29,861,259	9,947	TTTAGGG
Chr11	1	20,425	20,425	CCCTAAA	28775052	28,782,057	7,006	TTTAGGG
Chr12	1	28,497	28,497	CCCTAAA	26943421	26,948,440	5,020	TTTAGGG

Supplementary Table 5. Position and monomers of centromeres in each chromosome.

Chromosome	Start	End	Monomers
Chr01	26,830,000	27,420,000	AGGCCAAATGACTTATATGTATTGATACAGCAAAAATTGGTTAATATAGTGTTAGGCGACGCATTATTTAAACAATGCGTCACCAAATACATGAACAGGCGACGCATTCTTCAACAAATGCATCGCCTGTTCTTTTTTTTTTTCAATTTTTTTTTAAACATTTAAAAAAAATAAAATCAAAATAGGCCATAACAAGAATTGAACCCAGGACCTCCTACACTCTCAAGAAATCACCACACCACC
Chr02	19,184,195	19,831,362	ACTTCGGAGTTCTGATGGGATCCGGTGCATTAGTGCTGGTATGATCGCACCCGACATGGTGATGCTAAAGAATGGATATAAGGAAAAGAAGCAGCGCAGCAGGTCCGCATGCGTTCGGCGCGATCCGGGCAGCGGCATCGACGCACCGGCCCACCGAGCGAGTTCCCTCGGTCGGGCCGGGAGAAAAGGGAAACTCCGAGGGTGAAAGGCGCGGGGAAAGAGAGGAAAAAAAAAAAAAAAAGGGGTGCAACACGAGGACTTCCCAGGAGGTCACCCATCCTAGTACTACTCTCGCCCAAGCACGCTTA
Chr03	19,565,198	19,584,621	AAAAAAAAAAAAAACAAAAAAACTCTCTTTATATTGTTTAAATATCTTCACTGGTTATTTCATGCTGAAGAAAATCAAATTGTTAAAAATGAAGGTTTAAAAAAATATATATATATAGGGTATTAGGCGACGCATAATCATGATTATGCGTCGCCTAAACTAAATT
Chr04	9,031,791	9,365,128	ATTAAAAAAAAAAAAAAAAACTCTCATTTATCTTGTTTAAATATTTTCACTGGTTATTTGATGCTGGAAGAAAATCAAATTGTTAAAAAGAAGGTTTTAAAAAAAATAAATAGGGTATTAGGCGACGCATAATACTGATTATGCGTCGCCTAAACTAA
Chr05	160,000	220,000	ATTTGTTTTTTCGTTGGAGGCCAGGGGTTTGGTGGGATTTTGCTATTGGAGTTGCTATTTGATTGCTAATTTTTGTGACTGGAGGTGAGTTTGGTATTCAGTTAACAGAAAAAAGAAATTAGCTAAATTTGTTTTCAAGCCCTTGAGGTCAACAGTTGGTCTATTTTAATTGAAGTTTGGTACGATTGAGATTTCTTTACTTTTATAAAGCTTGTTATTCTTGCTTGAATGCAAAAAAAAATAAAAATAAAAAAAAATAAAAAAAAATAAAAGAAAAAGAAAAAGAAGAAGGGTTTGCGGAATTTTAAAAAAAAAAAAAAACAGCAC
Chr06	3,769,625	4,171,787	ATTTCTTTCAACTTTAAATAACAAGTGAAAATGTTATAGCTACTTCACAATGTTTTTTTTTTTTGTTTTTGTCAATTGAAATATTTTAGAAAATACGTTATTGATATATAGCGACGCATAAATACCATTATGCGTCGCCTATTATTGTGAAATTTCTTTTTTTTTTTGTTTTTTTTCTAAGTCTAAAACGGTTT
Chr07	22,563,313	22,782,010	TTTTTTTTTTTTTTAAAACCTTCCTTTTTAACAATTTGATTTTCTTCCAGCATCAAATAACCAGTGAAAATATTTAAACAAGATAAAGAGAGTTTTCTTTTTTTTTTAATTATAGTAATAGGCGACGCATAAGCAGTACTATGCGTCGCCTAATACCCAATTT
Chr08	24,604,388	24,623,002	CTGGAAAAATTACCGACGGTTTCGTCGGGAATTCCCGAGGGTGTACAGTCCATCACATTTTACCAATTTTTTGGGCCCCACAAGGCCCCTAAGGTGTGTTGAATGCAAAAACATGCAAAATAGGGATTCTATAATTGTACTAAGGAGAGAAAAGATCAAAATTTAATAAAAGTGTATCAAATTTTGCAAAAAATAGGATGACTGTTCACCCTCGGTAATTCCCGAGGAAAACGTCGGTAACCACCAAATCCCCT

Chr09	28,671,183	28,719,051	TGTTACGGAATTGGGAGAAAAACAAATAGCAACGGAAACAAAGAAAGCGAAGAACAAACACAAATTAACGTGGAAACCCTTGATGGGAAAAACCACGGGCAGGGAGAACAAATCCAATATCGAAAGATTGGTACAAAAGGTGAGCCTGACTGCGCGATACCTTCTAACCCTAATTACAGCCGAAAACTAAATATATATAGTACAGAAGAAACCCTAAAATTGAACAGACGGTGTACCTTCAACCATAGAAAGGGTGTATAGACGGTGCTTCGCAATCAGTTTCGTTGTGAGATTCTTTGTGATGTACAAAGATTCCAACTTGTCCCATAAGTCCTTTGTTGTCTTTTGATCAAGGACCTCCCTCAAAACTTCATTGGACAAGCATAGCTGAATGGTGGATAGGGCACGCTCATCAACCTCGGTTTTCTTTTCGGCATCCCACGAAGACGGCATCTGCTCCTTGCCAACCAACGCCTTGTGTAAACCGTTCTGTGTCAAAATCGCCTTCATCTTGACTTGCCACATGGAGAAACTCATGTTGCGCTCAAATTTCTCAATGTCGAACTTTGTTGCCATGATCGAAAGAAAAAAAATTCCCAAATAGATCGATGCGGCTCTGATACCACT
Chr10	27,753,093	27,910,515	GTTGCTTAAATATTTTCACGGATGATTTGATGCTGAAAGAAAATCAGATAGATTAAGAGAAGATTACACACACAAAAAAAAAAAAAATTTTATTTGCTAACAAGCGACGCATAAGTAGTTTTATGCGTCGCCTCTTACCATGATCTAAAAAAAAAATGTTAATTTAATAA
Chr11	13,632,283	13,868,269	AAAAAAAAAAAACTCTTTTTATTGTTTAAATATTTTCACTGGTTATTTGATGCTGCAAGAAAATCAAATTGTTAAAAAGCAAGGTTTTAAAAAAAAAAAAAAAAAAAATAGGGTATTAGGCGACGCATAGTACTGATTATGCGTCGCCTATTACTATAATTAA
Chr12	8,013,393	8,039,969	AAAAAAAAAAAAAAAAGCACTGTGATGCTGCTTAAACATTTTCACTTGTTATTTAAAGTTGAAAGAAAATCAACTGTTTAATACTTAGAAAAAACACAAAAAAAAATTATTACACTAATAGGCGACGCATAATAGTGTGTCTGCGTCGCCTGATATCAATAATATGTCTTCCAAAAAATTTTTTG

Supplemental Table 6. Collinear genes in paralogs and orthologs.

Species name	Collinear genes	Percentage	Paralog pairs in peak 1 (Ks<0.7 or 4DTv<0.25)	Paralog pairs in the peak 2 (0.7<Ks<3.0 or 0.25<4DTv<1.2)	No. of genes in Peak 1	No. of genes in Peak 2	No. of genes in shared by Peak 1 and 2
Dongzao	4,542	15.33	596	2,144	745	3,808	11
Junzao	4,249	15.06	399	1,707	660	3,101	25
Suanzao	4,752	15.27	411	2,109	794	3,730	87
Populus	21,530	52.03	10,097	4,842	20,058	5,903	4,431
Prunus	4,107	16.41	--	2,064	--	3,660	0

Supplemental Table 7. Statistics of Hi-C reads mapping to the genome.

Mapping information	
Unmapped Paired-end Reads	4,113,168
Unmapped Paired-end Reads Rate (%)	2.86
Paired-end Reads with Singleton	20,909,998
Paired-end Reads with Singleton Rate (%)	14.57
Unique Mapped Paired-end Reads	68,145,208
Unique Mapped Ratio (%)	47.49
Classification of unique mapped reads	
Dangling End Paired-end Reads	6,916,309
Religation Paired-end Reads	1,209,022
Self-Circle Paired-end Reads	30,297
Dumped Paired-end Reads	1,748
Valid Paired-end Reads	59,987,832
Vaild reads of unique mapping reads (%)	88.03
Vaild reads of clean reads (%)	41.80

Supplemental Table 8. The TAD information.

Class	Number	Max len (bp)	Min len (bp)	Median len (bp)	Mean len (bp)	Size (Mb)	Percentage	Genes
Domain	2,428	1,550,000	10,000	130,000	149,805	363.73	92.47%	27,203
Boundary	573	100,000	20,000	30,000	33,962	19.46	4.95%	2,145
Gap	241	1,659,268	10,000	16,005	43,542	10.15	2.58%	285

Supplemental Table 9. Gene expression and function of 51 predicted genes in the TAD located in Chr05:4.28-5.83 Mb.

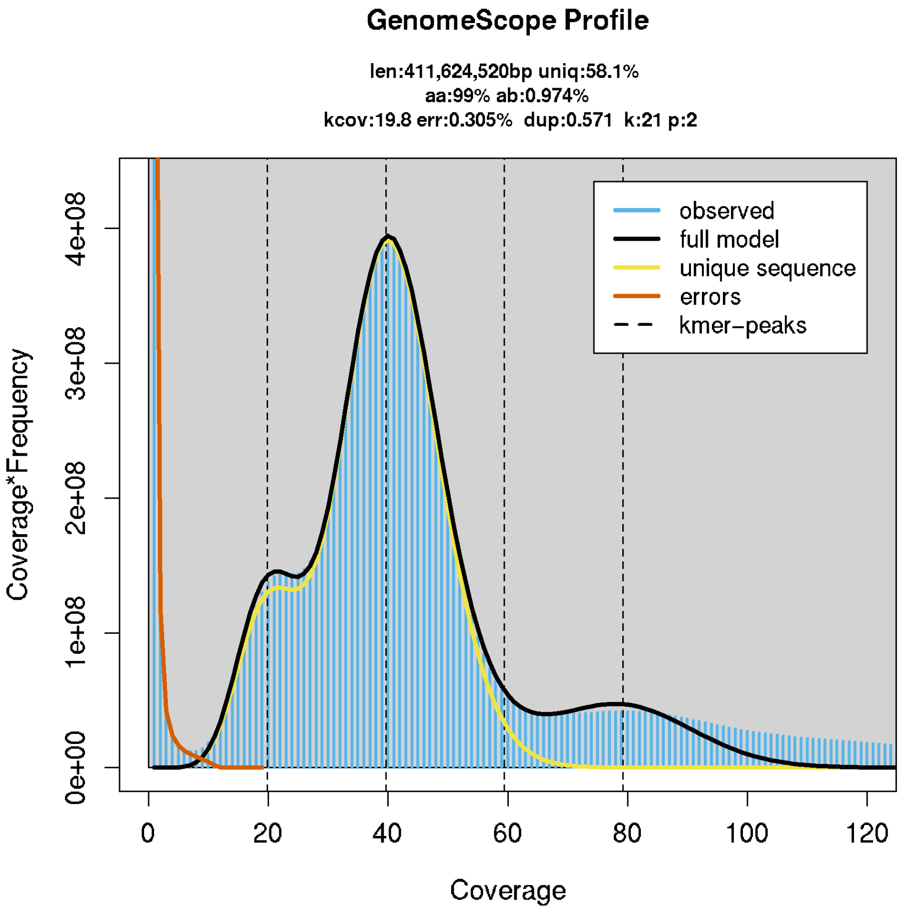
Gene name	FPKM	Function or chloroplast gene name	Gene name	FPKM	Function or chloroplast gene name
Chr05.694	0	Antisense to 16S rRNA	Chr05.771	0	NA
Chr05.695	0	ycf2	Chr05.773	0	psaA
Chr05.698	0	psbC	Chr05.797	0	transposition, RNA-mediated
Chr05.699	0	ycf3	Chr05.822	0	transposition, RNA-mediated
Chr05.702	0	atpB	Chr05.824	0	rbcL
Chr05.703	3.7	rbcL	Chr05.838	0	rps11
Chr05.709	0	atpA	Chr05.846	0	rbcL
Chr05.710	0	transposition, RNA-mediated	Chr05.848	0	rpoC2
Chr05.712	0	ycf2	Chr05.852	0	psaB
Chr05.713	0	psbA	Chr05.853	0	rpoA
Chr05.714	0	atpA	Chr05.861	0	ycf68
Chr05.715	0	psbB	Chr05.868	0	psbB
Chr05.727	0	ndhK	Chr05.871	0	rpoA
Chr05.737	0	Antisense to 23S rRNA	Chr05.872	0	ycf2
Chr05.743	0	ndhB	Chr05.882	0	petA
Chr05.744	0	ycf2	Chr05.883	0	rps12
Chr05.745	0	TMV resistance protein N-like	Chr05.901	0	ycf2
Chr05.751	0	Photosystem II protein	Chr05.921	0	NA
Chr05.753	0	psaB	Chr05.936	0	ycf68
Chr05.756	0	psbB	Chr05.938	0	rpoC1
Ch05.760	0	NA	Chr05.940	0	psbC
Chr05.762	0	rpl2	Chr05.941	0	psaB
Chr05.765	0	rpoC2	Chr05.944	0	rbcL
Chr05.766	0	psbD	Chr05.949	0.14	psbD
Chr05.767	0	psaA	Chr05.952	0	NA
Chr05.768	0	Cytochrome f	--	--	--

Supplemental Table 10. The corresponding relationship of MDHAR genes with those in Liu et al. 2014.

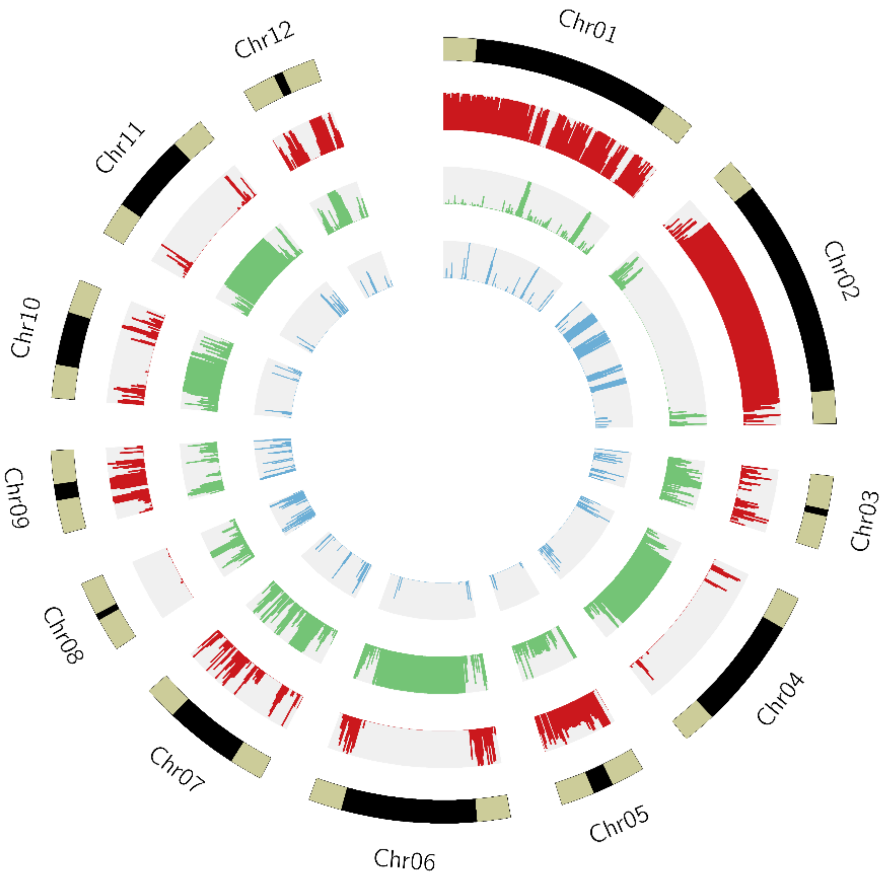
MDHAR gene ID in this study	Average FPKM in 16 tissues	The Highest FPKM in 16 tissues	The tissue with highest FPKM	Corresponding genes in (Liu et al. 2014)	Scaffolds of NGS genes in (Liu et al. 2014)	Phylogeny group in (Liu et al. 2014)
Chr01.4867	7.05	18.71	Seedling	NA	NA	
Chr01.4869	0.04	0.28	Stem	CCG016762.1, CCG016763.1	scaffold402	V
Chr01.4870	0.00	0.06	Flower	CCG016764.1	scaffold402	V
Chr01.4871	0.00	0.00	NA	NA	NA	
Chr01.4877	0.24	0.87	Root	CCG016765.1	scaffold402	V
Chr01.4878	0.04	0.48	Root	CCG016766.1	scaffold402	V
Chr01.4879	0.00	0.00	NA	CCG016767.1	scaffold402	V
Chr01.4882	0.00	0.00	NA	NA	NA	
Chr01.4890	0.03	0.17	Branch	NA	NA	
Chr01.4891	0.02	0.31	Root	NA	NA	
Chr01.4892	1.82	14.91	Stem	NA	NA	
Chr01.4893	1.78	12.15	Stem	NA	NA	
Chr01.4894	4.87	13.98	Leaves	CCG022250.1, CCG022251.1	scaffold627	V
Chr11.1336	307.99	662.00	Seedling	CCG023124.3	scaffold671	III
Chr06.3975	15.84	54.95	Seedling	CCG013319.1	scaffold285	II
Chr04.4579	20.49	78.24	Branch	CCG001931.1	scaffold113	I

Note: Shaded part is the tandem expanded MDHAR genes in jujube.

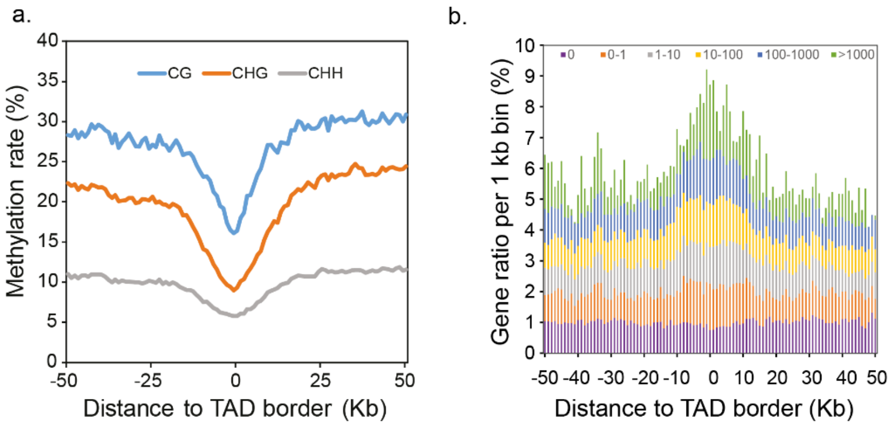
Supplemental Figures



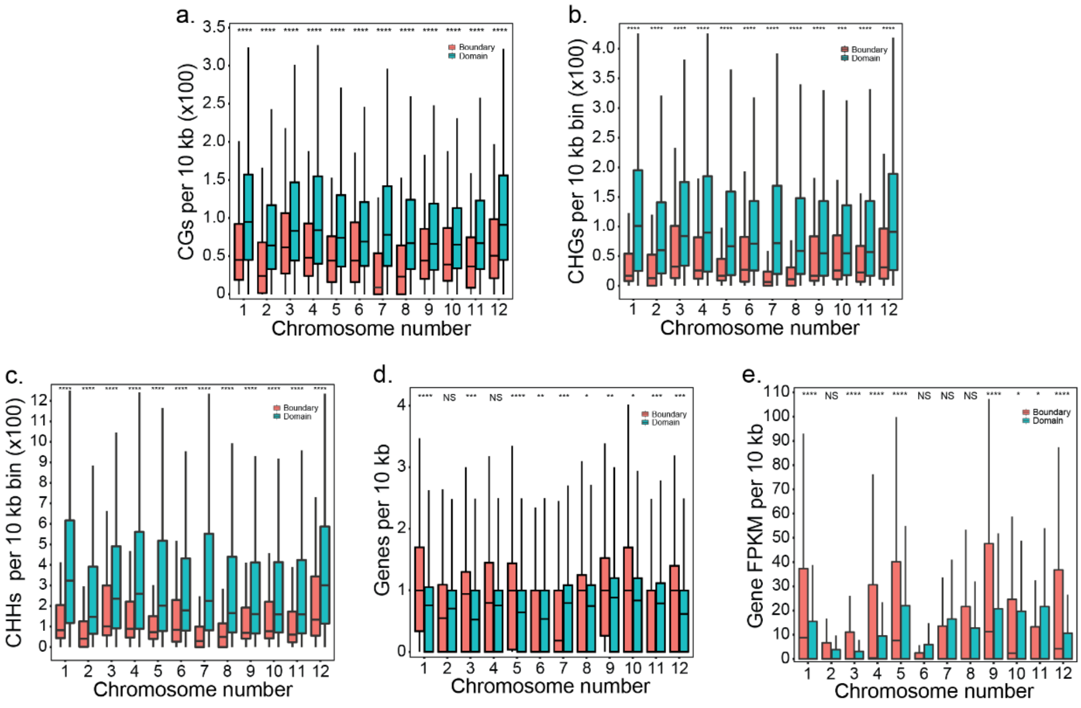
Supplemental Figure 1. Genome size estimation results using GenomeScope2 based on MGISEQ-2000 clean reads.



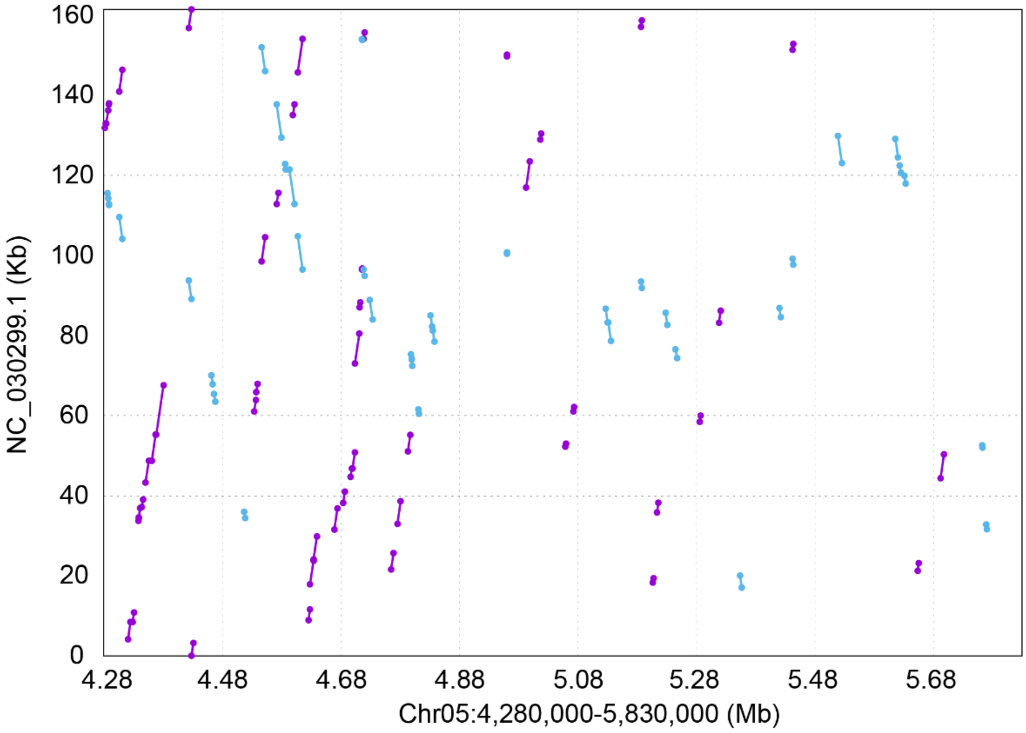
Supplemental Figure 2. The centromere sequences (in black) and their 100 kb flanking regions (in cyan) are displayed in the outer circle. The chromosome position for each centromere can be found in Supplemental Table 5. The LTR-transposons (red), non-LTR repeats (green), and protein gene transcripts (blue) are represented, respectively, by the second, third, and fourth circles (from outer to inner).



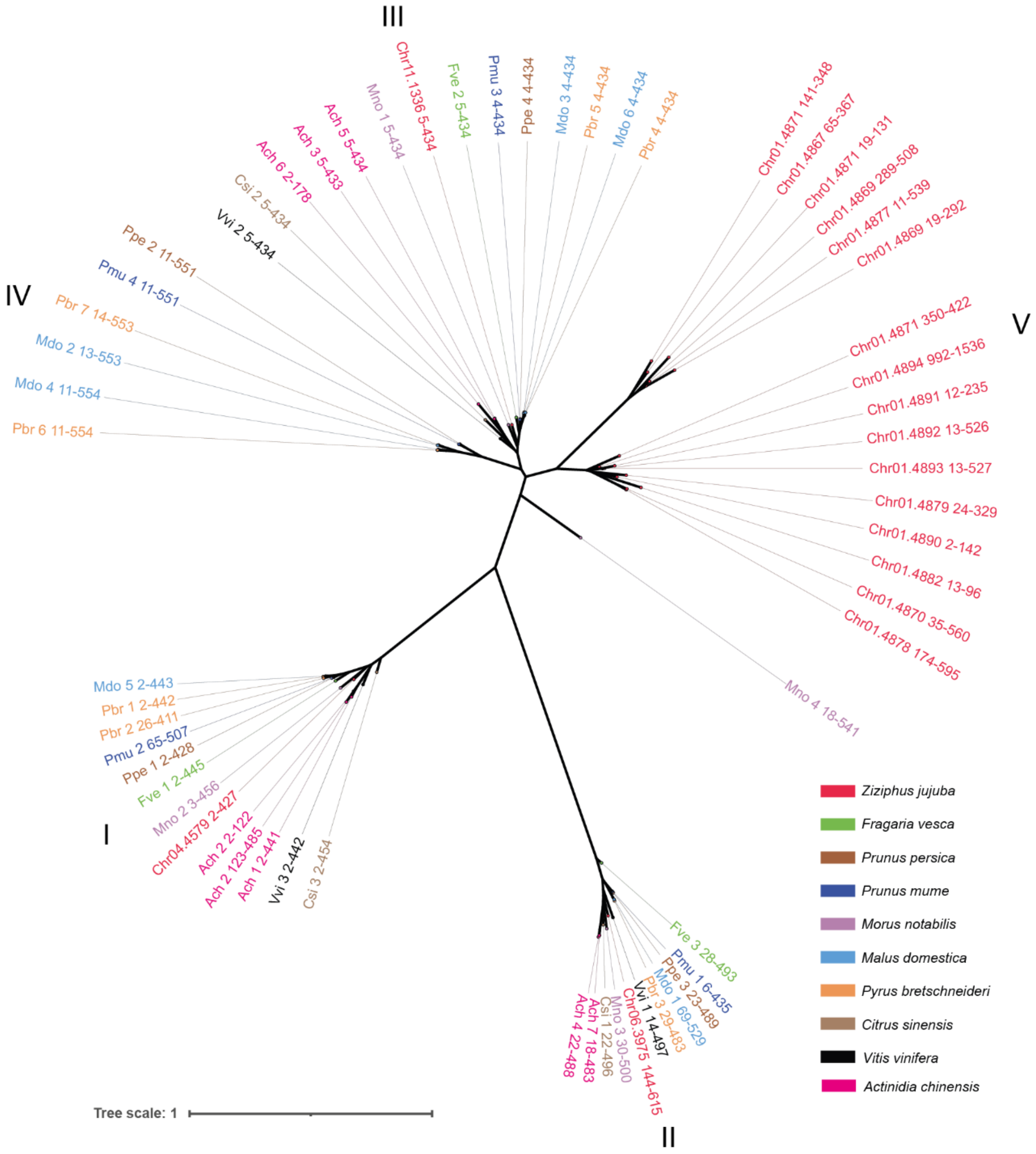
Supplemental Figure 3. Methylation and gene expression at TAD borders. (a) The methylation at the TAD boundary and the flanking 50 kb regions. (b) Histogram of the gene expression alternation at the TAD boundary and the flanking 50 kb regions.



Supplemental Figure 4. Boxplot of methylations and genes features in a 10 kb bin distributed within TAD and in TAD boundary. (a-e) average number of methylated CGs, CHGs and CHHs, number of genes, and gene FPKMs in a 50 kb bin distributed in TAD and in TAD boundary. Asterisks indicate statistically significant differences from the Wilcoxon rank sum test (* $p<0.05$, ** $p<0.01$, *** $p < 0.001$, **** $p < 0.0001$, NS, not significant).



Supplemental Figure 5. Horizontal transfer of chloroplast fragments to the nuclear genome occurred on the largest TAD located in chromosome five between 2.8 Mb and 5.83 Mb. The jujube chloroplast genome (NC_030299.1) was obtained from NCBI. In the total 1.55 Mb region of this TAD, 227 Kb are aligned to the chloroplast genome with an average identity of 93.22%, covering 140 Kb (86.83%) of the chloroplast genome.



Supplemental Figure 6. An updated phylogenetic tree of the MDHAR gene family compared to (Liu et al., 2014) in jujube and nine related species. Group V is the jujube-specific MDHARs containing 13 members. The tree was built with the IQTREE software using the core domain region of protein sequences, and the start and end positions of the region for each gene are indicated after the gene name. The abbreviation name for each species and the corresponding accession number of NCBI are listed as: Csi 1 (XP 006476500), Csi 2 (XP 006481820), Csi 3 (XP 006470310), Fve 1 (XP 004304631), Fve 2 (XP 004303012), Fve 3 (XP 011463921), Mdo 1 (XP 017181664), Mdo 2 (XP 008366501), Mdo 3 (XP 008391762), Mdo 4 (XP 008370471), Mdo 5 (XP 008341454), Mdo 6 (XP 028946174), Mno 1 (XP 024032990), Mno 2 (XP 024029543), Mno 3 (XP 010089047), Mno 4 (XP 010089361), Pbr 1 (XP 009377205), Pbr 2 (XP 048433812), Pbr 3 (XP 048446112), Pbr 4 (XP 009374749), Pbr 5 (XP 009366639), Pbr 6 (XP 018499810), Pbr 7 (XP 009334596), Pmu 1 (XP 016651062), Pmu 2 (XP 008230586), Pmu 3 (XP 008241272), Pmu 4 (XP 008226785), Ppe 1 (XP 007215303), Ppe 2 (XP 020417215), Ppe 3 (XP 007209972), Ppe 4 (XP 007202072), Vvi 1 (XP 010658330), Vvi 2 (XP 010653731), Vvi 3 (XP 002277200), Ach 1 (PSR87572), Ach 2 (PSS30380), Ach 3 (PSS15949), Ach 4 (PSS09385), Ach 5 (PSS01382), Ach 6 (PSR98500), Ach 7 (PSR91476).