Plant Communications
Correspondence

CellPress
Partner Journal

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

Dear Editor,

Jujube (*Ziziphus jujuba* Mill.), commonly called Chinese jujube, is a vital member of the Rhamnaceae family. It is famous for its tolerance to dry, barren, and saline-alkali soils, and its fruit has important nutritional and medicinal value. Recent fundamental research on jujube has involved assembly of draft genome sequences for the fresh-eating cultivar 'Dongzao' (Liu et al., 2014), dry-eating cultivar 'Junzao' (Huang et al., 2016), and wild sour jujube 'Suanzao' (Shen et al., 2021). However, genome evolution studies based on high-quality genome assemblies with large collinear regions have not been performed. Likewise, the spatial architecture of jujube chromosomes and its effect on gene transcription and regulation remain to be explored. Here, we report an updated gapless genome assembly of jujube (*Z. jujuba* Mill. Dongzao) and use it to characterize new features of jujube evolution and spatial chromosome organization.

The jujube genome size was estimated to be 411.6 Mb using 20.5 Gb (50×) clean MGISEQ-2000 paired-end reads (Supplemental Figure 1; Supplemental Table 1). A total of 28.6 Gb (70×) PacBio HiFi circular consensus sequencing (CCS) reads, 50.7 Gb (123×) Oxford Nanopore Technologies (ONT) ultra-long reads, and 43.4 Gb (105×) high-throughput chromosome conformation capture (Hi-C) data were obtained from the whole genome (Supplemental Table 1). We first produced a 417.6-Mb raw assembly from the HiFi reads, then removed contaminants, organelle sequences, and duplicated contigs. We next integrated the ONT and Hi-C data, generating a final assembly of 393 332 932 bp with an N50 (The sequence length of the shortest contig at 50% of the total assembly length) of 32.99 Mb. The assembly consisted of 12 gapless contigs, which we named Chr01–Chr12 in descending order of length (Supplemental information 1.1). BUSCO evaluation revealed 98.5% completeness of the genome (Figure 1A; Supplemental Table 2). Approximately 56.16% of the genome sequences were repetitive, of which 37.57% were transposon elements (Supplemental Table 3). A total of 29 633 protein-coding genes were predicted on the 12 chromosomes, and 27 500 (92.80%) were functionally annotated (Figure 1A and 1B). All telomere sequences, ranging from 4217 to 28 497 bp in length, were identified on the 12 chromosomes (Supplemental Table 4). Centromeres were predicted by considering both the long tandem repeats (Supplemental Table 5) and the Hi-C matrix, in which the highly repetitive centromere region is typically difficult to fully cover with Hi-C reads (Figure A4 of Supplemental information 1.1). The centromeres varied in length and contained diverse monomer sequences across different chromosomes (Supplemental Table 5). In addition, the centromere regions typically lacked genes, and different centromeres were enriched in distinct types of repeats, such as retrotransposons with long terminal repeat (LTR) and without LTR (Supplemental Figure 2).

Using homologous genes from collinear genomic regions, we calculated synonymous substitutions per site (*Ks*) and four-fold synonymous third-codon transversion rates (4DTv) for *Z. jujuba* Mill., *Populus trichocarpa*, and *Prunus persica*. Because the three available jujube genotypes, Dongzao, Junzao, and Suanzao, had similar *Ks* and 4DTv distributions, we will refer to them simply as "jujube" in this paragraph. All three species shared a common peak around *Ks* = 1.5 and 4DTv = 0.5 (peak 2), with a median value of paralogous sequence similarity (MPSS) below 80%. *P. trichocarpa* showed a sharp peak around *Ks* = 0.27 and 4DTv = 0.09 (peak 1), with an MPSS of approximately 90%, which has previously been reported as a recent species-specific duplication (Tuskan et al., 2006). This peak was absent in *P. persica*. We also identified a mini peak in jujube at approximately *Ks* = 0.15 and 4DTv = 0.05, which was more recent than *P. trichocarpa* peak 1 (two peaks for Dongzao, MPSS 90% and 98%) (Figure 1C–1E). The paralogous genes surrounding this mini peak accounted for approximately 20% of those surrounding jujube peak 2 (Supplemental Table 6). Jujube has been speculated to lack species-specific genome-wide duplications (Liu et al., 2014). However, owing to the highly fragmented genome assembly, this conclusion may not reflect the true evolutionary process. The mini peak, although it represents only a small-scale duplication, highlights the different evolutionary processes after the speciation events of jujube and *P. trichocarpa* (*P. persica*). *P. trichocarpa* underwent a new, recent round of genome-wide duplication, whereas *P. persica* showed no signs of new duplication. Jujube evolved more slowly than *P. trichocarpa* but faster than *P. persica*, and we speculate that the mini peak represents an ongoing duplication in jujube. The *Ks* and 4DTv peaks of Dongzao relative to Suanzao (Junzao), representing the speciation event, were both close to the y axis (Figure 1C and 1D). This suggests that the three individuals, whether cultivated or wild, are not fully separated at the species level compared with the separation between *Z. jujuba* and *Ziziphus mauritiana* (Supplemental information 1.2). This finding provides insight into the controversial taxonomic issue of whether jujube and wild jujube belong to the same species (Supplemental information 1.2).

We used the 143.5 Mb of Hi-C reads (Supplemental Table 1), ultimately employing 61 Mb of valid pairs (41.80%) for further analysis (Supplemental Table 7). The interaction signals were

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

A

Information	'Suanzao'	'Junzao'	**'Dongzao'**
Size of Assembly (bp)	405,636,987	362,583,438	**393,332,932**
Ungapped size of assembly (bp)	405,604,715	362,254,853	**393,332,932**
% sequences anchored to chromosomes	93.76%	96.99%	**100.00%**
Number of Contigs	907	269	**12**
Contig N50 (bp)	2,216,458	2,172,811	**32,986,920**
Chromosome number (2n)	2 × 12	2 × 12	**2 × 12**
% BUSCO complete of genome	97.70%	96.90%	**98.50%**
LTR Assembly Index (LAI)	15.87	16.42	**17.89**
Size of repeat sequences (Mb)	226.65	193.52	**220.88**
Number of Intact LTR	2,010	1,804	**2,330**
Number of protein-coding genes	31,115	28,223	**29,633**
% of genes anchored to chromosomes	95.30%	93.42%	**100.00%**
Annotated protein-coding genes	28,863	26,222	**27,500**
% BUSCO of complete genes	93.10%	91.50%	**93.70%**

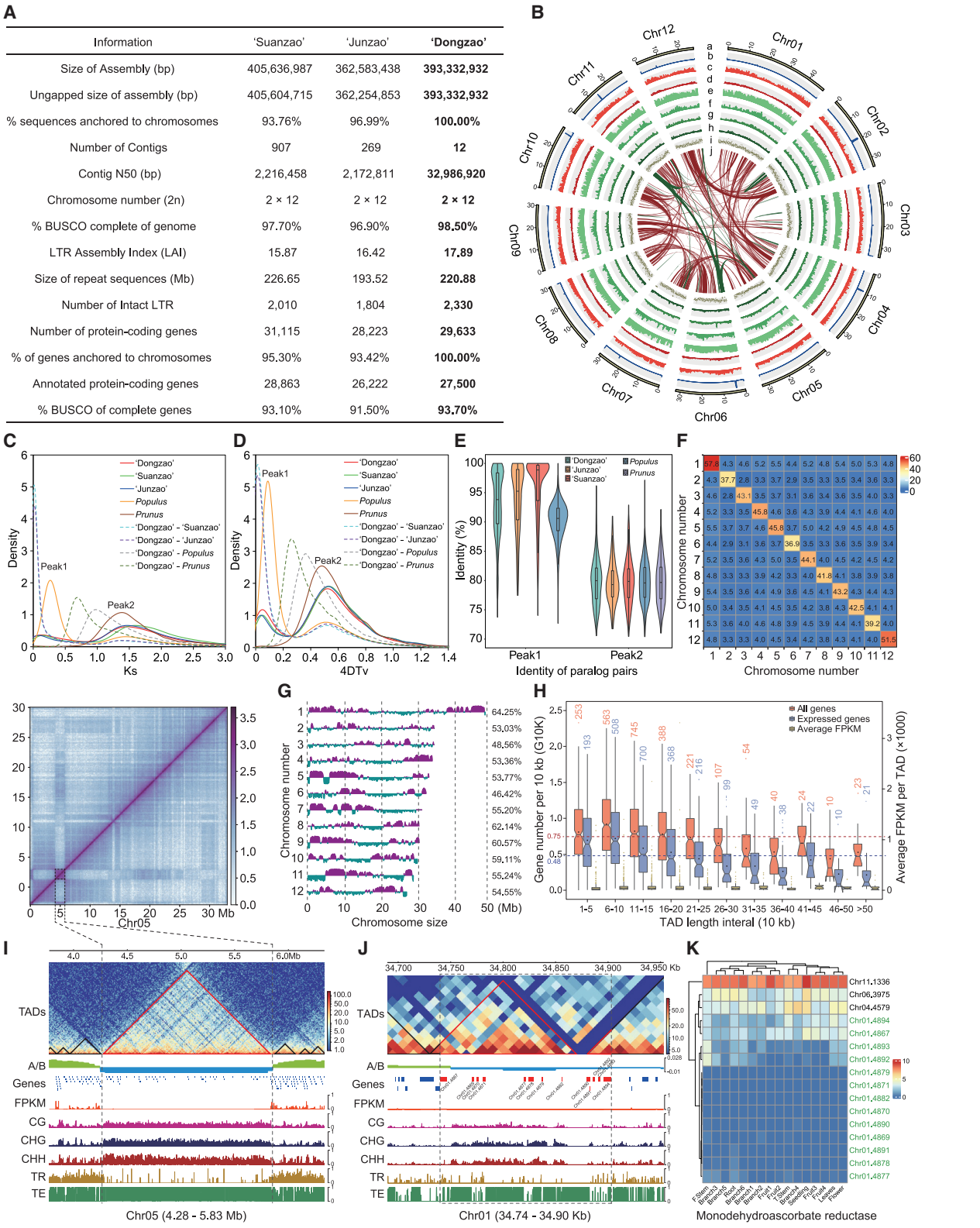

Figure 1. Assembly, annotation, evolutionary insights, and spatial chromosome analysis of the jujube T2T gapless genome.
(A) Jujube genome data. The genomes of Junzao and Suanzao were reannotated using the same method as the updated Dongzao genome.
(B) Circular plot of jujube chromosomes (Chr). **(a)** The twelve chromosomes in proportion to their actual lengths; **(b)** telomeres and putative centromeres; **(c)** genes; **(d)** protein-coding regions; **(e)** total repeats; **(f)** Gypsy repeats; **(g)** Copia repeats; **(h)** intact long-terminal repeats; **(i)** genomic GC content between 30% and 40%; and **(j)** genome-wide collinear blocks, with green and red representing mini and second peaks, respectively.

(legend continued on next page)

much stronger along the main diagonals than in other regions, and the inter-chromosome interaction strength was typically <10% of the intra-chromosome interaction strength (Figure 1F). The A and B compartments accounted for 55.75% and 44.25% of the total genome, respectively, and showed a diverse distribution. The ends and middles of chromosomes 1, 5, 9, and 10 were typically clustered toward the A and B compartments, respectively, whereas other chromosomes exhibited an interleaved distribution of the two compartments (Figure 1G). The chromatin was partitioned into 2428 topologically associating domains (TADs) with a mean length of 149.8 kb, comprising 27 203 genes and representing 92.47% of the genome size (Supplemental Table 8). As reported for the rice genome (Liu et al., 2017), genes along the TAD borders and boundaries were less methylated and showed higher expression than those in TADs (Supplemental Figures 3 and 4).

One noteworthy feature of the jujube TADs was the general decrease in gene count with increasing TAD size for both total and expressed genes. We divided the TADs into 11 length intervals and investigated the number of genes per 10 kb (G10K). The G10K value of TADs was 0.75, consistent with that at the whole-genome level (0.74). However, in TADs >200 kb, G10K values were all below 0.75 and generally decreased with increasing TAD size. However, average gene expression did not appear to differ among length intervals, suggesting that TAD size influenced gene number but not gene expression (Figure 1H).

The largest TAD, located on chromosome 5 between 4.28 and 5.83 Mb, had robust internal interactions with strong signals and weak interactions with other regions; all regions were in the B compartment, surrounded by abundant methylations and transposons (Figure 1I). All 51 predicted genes were either not expressed or were expressed at low levels; alignment to the jujube chloroplast genome revealed that this TAD primarily comprised horizontally transferred chloroplast sequences and degenerated chloroplast genes (Supplemental Figure 5; Supplemental Table 9). This is the first report of chloroplast sequences in the plant nucleus that are structured as a large compact TAD and serves as a foundation for further studies on transferred chloroplast sequences in the nucleus.

One of the most essential traits of jujube fruit is its high vitamin C content (Liu et al., 2014). The monodehydroascorbate reductase (MDHAR) family, which is involved in the vitamin C recycling pathway (Li et al., 2010), has been reported to contain eight specific members in jujube that are not present in other species (Liu et al., 2014). Our assembly brought this number to 13, six of which correspond to the previous eight members (Supplemental Figure 6; Supplemental Table 10). The 13 MDHAR family members were distributed in two TADs with frequent interaction among one another and were tandemly packed over a 163-kb region, uninterrupted by other genes. Chr01.4867 was the only gene in the cluster that was found in the A compartment, and it had the highest expression. All others were in the B compartment and showed extensive methylation, and the majority showed low or no expression (Figure 1J and 1K). The remaining three genes were orthologs shared with other species; one of them, Chr11.1336, was highly expressed in all tissues and is potentially the most important member of the MDHAR family (Figure 1K). These patterns of jujube-specific MDHAR gene expansion and distribution provide new evidence to support further research on the mechanism of vitamin C accumulation in jujube.

ACCESSION NUMBERS

The genome assembly and related raw sequencing data are deposited at the National Genomics Data Center (NGDC) under BioProject PRJCA016173. The annotation files for protein-coding genes, along with other related files for genome assembly validation, have been stored in the figshare, an online data repository, at https://figshare.com/s/56c2299b47a5efd8708f.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xplc.2023.100662.

FUNDING

This work was supported by the general program of the Natural Science Foundation of Hebei Province, China (C2022204030); the general program of the National Natural Science Foundation of China (32171817); special research projects for the new talent of Hebei Agricultural University, Hebei Province, China (YJ2020025); the China Agricultural Research System (CARS-30-2-07); and grants from the Hebei Province Key R&D Program (21326304D).

AUTHOR CONTRIBUTIONS

M.Y. and M.L. conceived the project; M.Y., L.D., S.H., J.Z., L.H., P.L., and Z.Z. prepared the samples and performed the experiments; M.Y., L.H., S.Z., and B.L. performed the bioinformatics analysis; M.Y. wrote the paper; and M.L. revised the paper. All authors reviewed and approved the paper.

ACKNOWLEDGMENTS

No conflict of interest is declared.

Received: May 3, 2023
Revised: July 17, 2023
Accepted: July 20, 2023
Published: July 23, 2023

(C and D) Distributions of synonymous substitutions per site (*Ks*) **(C)** and four-fold synonymous third-codon transversion rate (4DTv) **(D)**. The solid and dashed lines represent duplication and speciation events, respectively.
(E) Identity percentages of paralogs.
(F) Quantification of interactions between pairs of chromosomes.
(G) A/B compartments of the 12 chromosomes. The upward purple and downward blue regions represent the A and B compartments, respectively; density values range from −0.03 to 0.03. The percentage of the A compartment for each chromosome is indicated on the right-hand side of the graph.
(H) Gene number and expression of different TAD length intervals. Black dot in each boxplot denotes the average value.
(I) TAD between 4.28 and 5.83 Mb of chromosome 5 (between dashed lines). A/B, A/B compartments; CG, CHG, and CHH, different methylation types; TR, tandem repeats; TE, transposon elements.
(J) Jujube-specific 163-kb MDHAR region (between dashed lines).
(K) MDHAR gene expression heatmap. There were 16 benchmarked tissues, 15 reported by Liu et al. (2014) and one "seedling" from this work.

Meng Yang[1,4,], Lu Han[1,4], Shufeng Zhang[1], Li Dai[1,2], Bin Li[1], Shoukun Han[1], Jin Zhao[3], Ping Liu[1,2], Zhihui Zhao[1,2] and Mengjun Liu[1,2,*]*

[1]College of Horticulture, Hebei Agricultural University, Baoding, Hebei 071001, China
[2]Research Center of Chinese Jujube, Hebei Agricultural University, Baoding, Hebei 071001, China
[3]College of Life Sciences, Hebei Agricultural University, Baoding, Hebei 071001, China
[4]These authors contributed equally to this article.
***Correspondence: Meng Yang (yangm@hebau.edu.cn), Mengjun Liu (kjliu@hebau.edu.cn)**
https://doi.org/10.1016/j.xplc.2023.100662

REFERENCES

Huang, J., Zhang, C., Zhao, X., Fei, Z., Wan, K., Zhang, Z., Pang, X., Yin, X., Bai, Y., Sun, X., et al. (2016). The jujube genome provides insights into genome evolution and the domestication of sweetness/acidity taste in fruit trees. PLoS Genet. **12**:e1006433.

Li, M., Ma, F., Liang, D., Li, J., and Wang, Y. (2010). Ascorbate biosynthesis during early fruit development is the main reason for its accumulation in kiwi. PLoS One **5**:e14281.

Liu, M.J., Zhao, J., Cai, Q.L., Liu, G.C., Wang, J.R., Zhao, Z.H., Liu, P., Dai, L., Yan, G., Wang, W.J., et al. (2014). The complex jujube genome provides insights into fruit tree biology. Nat. Commun. **5**:5315.

Liu, C., Cheng, Y.J., Wang, J.W., and Weigel, D. (2017). Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. Nat. Plants **3**:742–748.

Shen, L.Y., Luo, H., Wang, X.L., Wang, X.M., Qiu, X.J., Liu, H., Zhou, S.S., Jia, K.H., Nie, S., Bao, Y.T., et al. (2021). Chromosome-scale genome assembly for Chinese sour jujube and insights into its genome evolution and domestication signature. Front. Plant Sci. **12**:773090.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**:1596–1604.