De novo chromosome-level genome assembly of Chinese walnut (Juglans cathayensis Dode)

Feng Yan[1], Rui-Min Xi[1], Rui-Xue She[1], Yu-Jie Yan[1], Peng-Peng Chen[1], Ge Yang[1], Meng Dang[1], Ming Yue[1], Keith Woeste[2], and Peng Zhao[3]

[1]Affiliation not available
[2]USDA Forest Service
[3]Northwest University

May 21, 2020

Abstract

Chinese walnut (Juglans cathayensis Dode), is a diploid, woody species native to China. It has 16 chromosomes (2n =2x= 32), as do all members of its genus and most members of its family (Juglandaceae). Although high-quality sequence data and reference genomes are available for several Juglans species, our goal was to produce a de novo, chromosome-level assembly of the Chinese walnut genome to gain insights into the species' evolution and biology. Our assembly was based on Nanopore long reads and chromosome conformation capture (Hi-C) data. The final assembly showed a contig N50 size of 6.49 Mb and a scaffold N50 size of 36.1 Mb. The final genome size of Chinese walnut was estimated to be 548 Mb. The sixteen scaffolds of the assembly anchored 99 % of the Chinese walnut genome. The assembly of the gene space (BUSCO) was 92.0 %. We annotated 29,032 protein coding genes with a mean of 6 exons per gene. We detected 2,993 non-coding RNA in the genome. A phylogenetic analysis based on 552 single-copy orthologs indicated that Chinese walnut close relative to Persian walnut (J. regia). The collinearity analysis showed that two whole genome duplication (WGD) events in J. cathayensis and J. regia from a common ancestor. Comparative genome analysis of J. cathayensis versus J. regia showed that 399 and 1,528 gene families were expanded and contracted respectively in the Chinese walnut genome. This J. cathayensis genome should be a useful resource for study of the evolution, breeding, and genetic variation in walnuts (Juglans).

KEYWORDS

Nanopore sequencing, genome assembly and annotation, *Juglans cathayensis* , Hi-C assembly, phylogeny, genomics

1 INTRODUCTION

Walnut (*Juglans* L.) is the most important and valuable genus in the woody plant family Juglandaceae. Walnuts are grown worldwide for their edible nuts and high-quality wood (Feng et al., 2018; Zhang et al., 2019). Chinese walnut (*J. cathayensis* Dode) is an ecologically important, wind pollinated, endemic species that grows between China's tropical zone (<22°N) and the Qinling Mountains/Huai River line (˜34°N) in hilly mid-elevation areas (Zhang et al., 2015; Bai et al., 2014; Dang et al., 2015). It is a diploid plant with 16 chromosomes (2n = 2x = 32) that belongs to the group of species called Asian butternuts (section *Cardiocaryon*) that also includes Japanese walnut (*J. ailantifolia*) and Manchurian walnut (*J. mandshurica*) (Figure 1; Zhao et al., 2018; Bai et al., 2016; Zhao et al., 2014). Chinese walnut is less valuable as a commodity than its close relative, *Juglans regia* (Persian or common walnut; Han et al., 2016; Dang et al., 2016; Feng et al., 2018), but Chinese walnut has been evaluated in breeding programs for its resistance to lesion nematodes (Trouern-Trend et al., 2020). In addition, Chinese walnut has potential as a medicinal crop (Yu et al., 2011; Li et al., 2013; Li et al., 2008; Sun et al., 2012; Bi et al., 2016).

The population genetics, morphology, and diversity of *J. cathayensis* have been described (Manning et al., 1978; Aradhya et al., 2007; Hu et al., 2017; Zhang et al., 2019), but in general, interest in Chinese walnut is based on its potential as a tertiary germplasm pool for improvement of common walnut (Zhou et al., 2017). Chinese walnut expresses horticultural traits such as cluster bearing habit (6–13 fruits per terminal) that make it attractive to Persian walnut breeders, and disease tolerance/resistance as compared to *J. regia* (Xu et al., 2007; Zhou et al., 2017) that recommend it as a rootstock. Wild populations of Chinese walnut and cultivated orchards of Persian walnut (*J. regia*) grow sympatrically (Dang et al., 2019) but hybridization between these two walnut species is reportedly rare (Shu et al., 2016).

A high-quality genome is an important genetic resource for the improvement of horticultural traits in perennial crops (Dong et al., 2019; Zhang et al., 2020a). The availability of high-throughput sequencing has accelerated the publication of the genomes of walnut (*Juglans*) species and hybrids (*J. regia* × *J. microcarpa*) (Zhu et al. 2019; Martínez-García et al., 2016; Stevens et al., 2018; Bai et al., 2018; Zhang et al., 2020b). A combination of long reads (Nanopore sequencing platform), Illumina and Hi-C auxiliary assembly can be used to produce a high-quality, chromosome-level genome (Choi et al., 2020; Suryamohan et al., 2020; Zhang et al., 2019). Despite its importance for understanding walnut evolution and its utility for breeding, functional gene mining, and disease resistance, genomic resources for *J. cathayensis* are minimal. For these reasons, we undertook the assembly of a chromosome-level, high-quality reference genome assembly for Chinese walnut.

2. MATERIALS AND METHODS

2.1 Chinese walnut sample collection and genomic DNA extraction

We collected leaf samples from a single individual of Chinese walnut (*J. cathayensis*) growing in the Qingling Mountains, Xi'an, Shaanxi, China (altitude: 1489 m, 33°46'58"E, 108°34'06"N). Genomic DNA was obtained using a plant DNA extraction Kit (TIANGEN, Beijing, China). A total of 18 tissues were collected from the same tree described above for RNA sequencing (Figure 1a, b; Table S1).

2.2 Illumina short-read sequencing

Chinese walnut was sequenced on the Illumina HiSeq X Ten platform using 20Kb libraries. The Illumina sequencing raw reads were processed with SOAPnukev1.5.6 to removing adapters or low-quality bases with the parameters is '-n 0.01 -l 20 -q 0.1 -i -Q 2 -G -M 2 -A 0.5 –d'. Finally, 62.87 Gb of clean reads were used to assemble after assessment and error correction (Table S1).

2.3 Nanopore sequencing and assembly

We prepared DNA using Oxford Nanopore Technologies' standard ligation sequencing kit SQK-LSK109DNA. Genomic DNA was size-selected using high-pass mode (> 20 kb) using a BluePippin BLF7510 cassette (Sage Science). After completion of sequencing, the raw nanopore sequencing reads were corrected using the program Canu version 1.5 with the parameters 'minReadLength 3000–min Overlap Length 500' and Smartdenovo with the parameters '-k 17 -c 1' (Koren et al., 2017). A preliminary de novo assembly was constructed using the Nanopore sequence, then we aligned the Illumina reads to the draft genome assemblies using BWA-MEM (Li, 2013).

2.4 Hi-C assembly of the chromosome-level genome

We constructed a Hi-C library using the Illumina NovaSeq platform. Bowtie2-2.2.5 (Langmead and Salzberg, 2012) was used to align the raw reads to the assembled contigs, and then we filtered low quality reads using a HiC-Pro pipeline (Servant et al., 2015) with the default parameters. The valid reads were used to anchor super-scaffolds with Juicer (Durand et al., 2016) and 3d-dna pipeline (Dudchenko et al., 2017).

2.5 RNA sequencing

RNA was extracted from eighteen tissues (bark from stems, axillary buds, immature female flowers, leaves (not fully expanded), mature leaves, immature male inflorescence, mature male inflorescence, new shoots, leaf

buds, mature female flowers, receptive female flowers, immature fruit, mature fruit, fruit epidermis, kernel, seed coat (testa), root, root bark) (Table S1) sampled from one plant. An RNA-Seq library was produced for each tissue using an NEBNext Ultra RNA Library Prep Kit (NEB, Beverly, MA, USA). Paired end sequencing was performed on Illumina HiSeq X Ten platform (Illumina, USA). After RNA quantification, we also pooled equivalent amounts of RNA from each of the eighteen tissues for full-length transcriptome sequencing. Using the purified mRNA as the starting material, a full-length cDNA library (10-15kb) was constructed for the PacBio Sequel platform (NEB, USA). Bioanalyzer 2100 software (Panaro et al., 2000) was used to test the Library quality.

2.6 Evaluation of assembly quality

The quality of the assembly was evaluated using the mapping rate of the paired-end and long reads to the assembly (Figure S1). We also evaluated the completeness and accuracy of the genome assembly using Bench marking universal single-copy orthologs (BUSCO) version 3.0.2 (Simão et al., 2015). Genome completeness was further evaluated by mapping of transcripts from 18 (Table S1) tissues and organs using GMAP (Wu and Watanabe, 2005).

2.7 Genome annotation

We annotated repeat sequences, gene structure, and noncoding RNA in the Chinese walnut genome (workflow, Figure S2). We used both homology based on prediction and *de novo* prediction to identify transposable elements (TEs). For *de novo* prediction, we constructed a repeat sequence database using RepeatModeler (http://www.repeatmasker.org), and predicted the presence of repeat sequences using RepeatMasker software (Maja et al., 2009) (http://www.repeatmasker.org), LTR-FINDER (Zhao and Hao, 2007) and PILER (Edgar and Myers, 2005) with default parameters. For homology based prediction, we identified transposable elements in the DNA and based on predicted proteins by comparing genomic sequence with the Repbase v21.12 database (Jurka, 2000) using RepeatMasker (Maja et al., 2009) (http://www.repeatmasker.org) and RepeatProteinMask v4.0.7 (Maja et al., 2009). Finally, all transposable elements identified by either method were merged into the final transposon annotations. Transposable elements (TEs) in the assembled Chinese walnut genome were also annotated using Tandem Repeats Finder (TRF) v4.09 (Benson et al., 1999).

To ensure accurate gene structure annotations, we combined homology prediction and *de novo* prediction methods. RNA sequences from eighteen tissues (Table S1) were used to train the software AUGUSTUS with default parameters (Stanke et al., 2006). We predicated gene structure *de novo* based on the statistical characteristics of genomic sequence data (such as frequency of codon, distribution of exon and intron) using SNAP (Johnson et al., 2008). We further predicated gene structure in the protein-coding genes by homology with genes identified in *Arabidopsis thaliana* , *Citrus sinensis, Juglans regia* , *Malus domestica* , *Olea europaea* ,*Oryza sativa* , *Populus euphratica* , *Quercus robur* , and Chinese walnut using Exonerate v2.2.0 (Slater et al., 2005). The final structural annotation of protein-coding genes was performed using a MAKER (Holt et al., 2011) pipeline that integrates AUGUSTUS (Stanke et al., 2006) and results from homologous protein mapping, RNA-seq mapping, and Nanopore mapping.

2.8 Functional annotation of protein-coding genes.

Predicted genes were subjected to functional annotation by performing a BLAST v2.2.3 homologue search against the final gene set (Altschul et al. 1990). BLASP (Altschul et al. 1990) was used to predict gene function through searches against five databases (E-value=$1e^{-5}$), including SwissProt (Boeckmann et al., 2003), TrEMBL(Boeckmann et al., 2003), KEGG (Kanehisa et al., 2000), InterPro (Zdobnov et al., 2001), Nr(Sujana et al., 2005), Swissprot(Bairoch and Apweiler, 2000), KOG (koonin et al., 2004), and GO (Ashburner et al., 2000).

2.9 Prediction of non-coding RNA

We annotated tRNA, rRNA, snRNA, and miRNAs across the assembled genome sequence. Non-coding RNA sequence was predicted using tRNAscan-SE 1.3.1(Lowe et al., 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/) based on the RNA structure. The rRNA sequences in the Chinese walnut genome were predicted using

BLASTN to search for conserved characteristics with related species such as *J. regia* . The miRNA and snRNA in the assembled Chinese walnut genome were identified using INFERNAL software (Nawrocki and Eddy, 2013) against the Rfam 13.0 database (Griffiths-Jones et al., 2005).

2.10 Gene family cluster identification and phylogenetic analysis

Nine species (*Arabidopsis thaliana* , *Citrus sinensis, Juglans regia* , *Malus domestica* , *Olea europaea* ,*Oryza sativa* , *Populus euphratica* , *Quercus robur* , and Chinese walnut) were selected for comparative genome analysis. All-versus-all BLASTP (Altschul et al., 1990) search results (Evalue = $1e^{-5}$) were used for gene family construction using OrthoMCL (Fischer et al., 2011). A maximum likelihood (ML) phylogenetic tree was constructed using RAxML v8.2.12 by conducting 1,000 bootstrap replicates. Species divergence times were estimated using PAML v4.5 software and MCMCtree (Yang, 2007) with the following parameters: 10,000 burn-ins, sample-frequency=2, and sample-number=100,000. We applied fossil calibration points to inform the species divergence time using Timetree (http://www.timetree.org/). CAFE v2.2 (Computational Analysis of gene Family Evolution) (Bie et al., 2006) was used to assess the expansions and contractions of orthologous gene families among all nine plant genomes based on the consensus phylogeny.

2.11 Genome duplication and synteny analyses

To estimate the time of whole-genome duplication events (WGD) in the Chinese walnut genome, reciprocal best hit (RBH) gene pairs were identified (Evalue is $1e^{-5}$) based on all-vs-all paralogs detected in BLASTP (Altschul et al., 1990). We identified synteny blocks and collinear blocks of gene pairs in the Chinese walnut genome using MCScanX with default parameters (Wang et al., 2012). The synonymous substitution rate (Ks) were calculated using the YN model in KaKs_Calculator v2.0 (Wang et al., 2010). The Ks distributions of orthologues within Chinese walnut and Persian walnut, and between Chinese walnut and Persian walnut were used to compare the relative substitution rates in different species by plotting with the ggplot2 package (Kaori and Murphy et al., 2013).

3 RESULTS

3.1 Genome sequencing

We first sequenced a total of ˜47.3 Gb clean reads (equivalent to ˜82x genome coverage) to assemble the Chinese walnut genome based on Illumina HiSeq X-Ten sequencing (Table S2). We then called a total of 62.87 Gb long reads (˜118 x genome coverage) from the Chinese walnut genome using Oxford Nanopore Technology sequencing platform (Table S3). A total of 101 Gb raw data of a chromosome conformation capture (Hi-C) was produced by the Nanopore sequencing platform (˜176 x genome coverage) (Table S4).

3.2 *De novo* genome assembly using Nanopore long reads

After filtering raw reads, the remaining clean reads were assembled into contigs and scaffolds using Illumina data and Nanopore data. A total of 213 scaffolds were generated with N50 size of 7.15 Mb (Table S5). We identified 1,375 complete BUSCOs, including 1,104 duplicated BUSCOs, 71 fragmented BUSCOs, and 1,160 single-copy in the assembled Chinese walnut genome (Table S6). There were 40 genes recognized as missing BUSCOs in the assembled genome (Table S6). Overall, we obtained ˜548 Mb of Chinese walnut genome based on long reads, which is about 94. 8 % of the survey genome (548 Mb) (Table 1).

3.3 Chromosome-level assembly of Hi-C data

A total of 0.54 Gb assembled scaffold sequence was correctly divided into sixteen groups corresponding to the sixteen Chinese walnut chromosomes (Figure 2c; Figure S1). A total of 397 contigs and 189 scaffolds were generated by Hi-C sequencing data; the N50 size of contigs was 6.49 Mb and the N50 size of scaffolds was 36.1 Mb, respectively (Table 1). Hi-C sequence (543 Mb) was mapped and anchored (99 % ;543 Mb/548 Mb) to the assembled 16 chromosomes of the Chinese walnut genome (Table 1).Chromosome numbering for *J. cathayensis* was based on homology to the numbering of *J. regia* chromosomes (Zhang et al., 2020) (Table S7). The lengths of the 16 assembled chromosomes of Chinese walnut ranged from 19,675,958 bp

to 55,052,647 bp with mean length is 33,963,507 bp, while chromosomes of Persian walnut ranged from 20,184,194 bp to 518,39,233 bp with mean length is 33,799,624 bp (Table S7).

3.4 Repeat annotation

We identified a total of 340.4 Mb of repeats (62.1 % of the genome) in the Chinese walnut genome, of which ˜62.42 % were transposable elements (TEs) (Table 1; Table 2). The most abundant repetitive sequences were long terminal repeat retrotransposons (LTR-RTs), which accounted for 41.2 % of the assembled genome (Table 2), followed by LINE (long interspersed nuclear element, 12.22 %), DNA (Class II TEs, 8.96 %), and SINE (short interspersed nuclear element, 0.01 %) (Table 2).

3.5 Gene annotation

A combination of *ab initio* prediction, homology search, and transcript mapping were used to predict the protein-coding genes in the Chinese walnut genome. RNA from eighteen tissues was used to predict gene models (Table S1). Predicted protein-coding genes (27,901) had an average gene length of 5,735 bp, an average coding sequence (CDS) length of 1,226 bp, and an average of 6 exons per gene (Table 1). When we compared Chinese walnut to Arabidopsis based on genome structural features, we found the distribution of CDS lengths exon lengths of*J. cathayensis* was similar to *A. thaliana* ; however, the distribution of mRNA lengths and intron lengths of *J. cathayensis*was unlike *A. thaliana* (Table 1; Figure S3). Among 27,901 predicted genes, 96.1 % could be functionally annotated in at least one of these seven databases (Table S8). There were 2,014 genes annotated in Nr database only, 23 genes annotated in InterPro only, 6 genes annotated in KEGG only, and no gene was annotated in swissProt or COG only (Figure S4). The GC density with an average length of 900 bp and an average GC content of 51.21% (Figure 3b). Gene density throughout the genome was about 11 genes per 100 kb, with 56,553 genes (94.96 %) present on chromosomally anchored contigs (Figure 3c); this was equivalent to 307 transcripts per 1Mb of chromosome (Figure 3d). There are 82 syntenic blocks in the Chinese walnut genome (Figure 3e). The portion of the Chinese walnut genome comprised of non-coding RNA was small; it included miRNA, tRNA, rRNA, and snRNA (Table S9). A total of 581 tRNA (Table S9), 792 small nuclear RNA (snRNA) and 132 microRNA (miRNA) were identified (Table S9).

3.5 Gene family cluster identification and phylogenetic analysis

We compared genomes of Chinese walnut and eight other plants based on 523 single-copy orthologs (Figure 4a). The number of single-copy orthologs in the genome of Chinese walnut was similar to Arabidopsis, the percentage of the Chinese walnut genome occupied by single-copy orthologs was higher than all other species in the comparison except*Q. robur* and *C. sinensis* . (Figure 4a). We identified 125,530 orthologous gene families that consist of 310,273 genes, with 9,906 orthogroups containing proteins from all species (Figure 4b, Table S10). We further compared Chinese walnut with three cultivated woody species: Persian walnut (*J. regia*), apple (*M. domestica*) and olive (*O. europaea*). We found 17.4 % (10,321/59,377) of all gene families existed in all four species, while 5 % (2,972) were specific to walnuts (*Juglans*) (Figure 4a; Figure S5). When the two *Juglans* species were compared, 457 genes were specific to Chinese walnut, and 1,704 gene families were shared in both walnut genomes. We discovered 399 gene families were expanded in Chinese walnut compared to all other species in the phylogenetic tree, and 1,528 were contracted (Figure 4b). As a comparison, Chinese walnut's close relative*Juglans regia* , showed 2,025 expanded gene families (5-fold more than *J. cathayensis*) and 243 contracted gene families (about 1/7 the number in the Chinese walnut genome) (Figure 4b).

We constructed a phylogenetic tree of these nine plant species with the monocot rice (*Oryza sativa*) as outgroup. The phylogeny was based on 552 single-copy orthologous genes (Figure 4b). As expected, the two closely related walnut species clustered on a branch with 100 % bootstrap support (Figure 4b). The divergence between Chinese walnut and Persian walnut was estimated to have occurred ˜28 Mya (Figure 4b).

We investigated whether any whole genome duplication (WGD) events have occurred during Chinese walnut

evolution. We identified a total of 86 synteny blocks and 5,614 genes in all blocks that covered 20.1 % of Chinese walnut genome (Figure 3; Table S11). We calculated the density distribution of the Ks values for the paired genes within each syntenic genomic block based on the collinear blocks between the genomes of Chinese walnut and Persian walnut (Figure 5). The peak of Ks was ˜0.25 and ˜1.5 for orthologous gene pairs between the two walnuts, indicating that ancestors of these two walnuts evolved through two ancient WGD events (Figure 5). Because the peak of Ks was ˜0 for orthologous gene pairs between Chinese walnut and Persian walnut, their genomes reflect recent species differentiation (Figure 5).

DISCUSSION

We report the first assembly of a high-quality, chromosome-level genome for Chinese walnut using a combination of Illumina HiSeq X Ten, Nonopore, and Hi-C sequencing platforms. Compared to previously available genome assemblies for this species, the Scaffold N50 value was improved 163 fold (contig N50 size of *J. cathayensis* v2.0 was 23,789,296 bp versus 145,095 bp contig N50 size for *J. cathayensis* v1.0), and the final calculated genome size is smaller (580Mb, v1.0 versus 548Mb, v2.0) (Figure1; Table S12). Through Hi-C, a chromosome-level genome was obtained with a scaffold size of 36Mb (Table S12) and scaffolds resolved into 16 chromosomes, unlike the previously available genome (*J. cathayensis* v1.0; Stevens et al., 2018) (Table S5, Table S12) (DeMaere & Darling, 2019; Zhang et al., 2020c; Chen et al., 2020; Choi et al., 2020). We predicted 29,032 protein-coding genes from the generated assembly. The phylogenetic tree revealed that Chinese walnut and Persian walnut diverged ˜4.6 million years ago. This high-quality chromosome-level genome will be benefit for breeding and genetic variation discovery in walnut (*Juglans*) species.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (41471038 and 31200500), Shaanxi Academy of Science Research Funding Project (2019K-06), Natural Science Foundation of Shaanxi Province of China (2019JM-008), and Opening Foundation of Key Laboratory of Resource Biology and Biotechnology in Western China (Northwest University), Ministry of Education (ZSK2018009). Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the U.S. Dept. of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

REFERENCS

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* , *215* (3), 403–410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, J. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* , *25* (1), 25-29.

Aradhya, M. K., Potter, D., Gao, F., & Simon, C. J. (2007). Molecular phylogeny of *Juglans* (*Juglandaceae*): a biogeographic perspective.*Tree Genetics & Genomes*, *3* (4), 363-378.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nuclc Acids Research* ,*28* (1), 45-48.

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

Bai, W. N., Wang, W. T., & Zhang, D. Y. (2014). Contrasts between the phylogeographic patterns of chloroplast and nuclear DNA highlight a role for pollen mediated gene flow in preventing population divergence in an East Asian temperate tree. *Molecular Phylogenetics and Evolution* ,*81* , 37-48.

Bai, W. N., Wang, W. T., & Zhang, D. Y. (2016). Phylogeographic breaks within Asian butternuts indicate the existence of a phytogeographic divide in East Asia. *New Phytologist* , *209* (4), 1757-1772.

Bai, W. N., Yan, P. C., Zhang, B. W., Woeste, K. E., Lin, K., & Zhang, D. Y. (2018). Demographically idiosyncratic responses to climate change and rapid Pleistocene diversification of the walnut genus

Juglans(Juglandaceae) revealed by whole-genome sequences. *New Phytologist* , *217* (4), 1726-1736.

Bi, D., Zhao, Y., Jiang, R., Wang, Y., Tian, Y., Chen, X., ... & She, G. (2016). Phytochemistry, bioactivity and potential impact on health of*Juglans* : the original plant of walnut. *Natural Product Communications* , *11* (6), 870-879.

Bie, T. D., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics* , *22* (10), 1269-1271.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., ... & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* , *31* (1), 365-370.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., ... & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* , 10(1), 421.

Chen, Y. X., Chen, Y. S., Shi, C. M., Huang, Z. B., Zhang, Y., Li, S. K., ... & Chen, Q. (2018). SOAPnuke: A mapreduce acceleration supported software for integrated quality control and preprocessing of High-Throughput sequencing data. *GigaScience* , *7* (1), gix120.

Chen, Y., Ma, T., Zhang, L. S., Kang, M. H., Zhang, Z. Y., Zheng, Z. Y., ... & Yang, Y. Z. (2020). Genomic analyses of a "living fossil": the endangered dove-tree. *Molecular Ecology Resources, 10* , 1-14

Choi, J. Y., Lye, Z. N., Groen, S. C., Dai, X., Rughani, P., Zaaijer, S., ... & Purugganan, M. D. (2020). Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biology* , *21* (1), 21.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., ... & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments.*Cell Systems* , *3* (1), 95-98.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... & Aiden, E. L. (2017). De novo assembly of the*Aedes aegypti* genome using Hic yields chromosome-length scaffolds. *Science* , *356* (6333), 92-95.

Dang, M., Liu, Z.L., Chen, X., Zhang, T., Zhou, H. J., Hu, Y. H., Woeste, K., & Zhao P. (2015). Identification, development, and application of 12 polymorphic EST-SSR markers for an endemic Chinese walnut (*Juglans cathayensis* L.) using next-generation sequencing technology. *Biochemical Systematics and Ecology, 60* , 74-80.

Dang, M., Zhang, T., Hu, Y. H., Zhou, H. J., Woeste, K., & Zhao P. (2016). *De novo* assembly and characterization of bud, leaf and flowers transcriptome from *Juglans regia* L. for the identification and characterization of new EST-SSRs. *Forests* ,*7* (10), 247.

Dang, M., Yue, M., Zhang, M., Zhao, G. F., & Zhao, P. (2019). Gene introgression among closely related species in sympatric populations: a case study of three walnut (*Juglans*) species. *Forests* ,*10* (11), 965.

Dong, X. G., Wang, Z., Tian, L. M., Zhang, Y., Qi, D., Huo, H. L., ... & Cao, Y. F. (2019). De novo assembly of a wild pear (*Pyrus betuleafolia*) genome. *Plant Biotechnology Journal* ,*18* (2), 581-595.

DeMaere, M. Z., & Darling, A. E. (2019). bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes.*Genome Biology* , *20* (1), 46.

Edgar, R.C., & Myers, E.W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* , *21* , 152.

Feng, X. J., Zhou, H.J., Zulfiqar, S., Luo, X., Hu, Y. H., Feng, L., ... & Zhao, P. (2018). The phytogeographic history of common walnut in China. *Frontiers in Plant Science* , 9, 1399.

Fischer, S., Brunk, B. P., Chen F., Gao X., Harb O., Iodice, J. B., ... & Stoeckert, C. J. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Current Protocols in Bioinformatics, 6(6), 12-19.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* , *33* (1), 121-124.

Gross, S. S., Do, C. B., Sirota, M., & Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informantde novogene prediction. *Genome Biology* , *8* (12), R267.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* , *9* (1), R7.

Han, H., Woeste, K., Hu Y., Dang, M., Zhang, T., Gao, X. X., ... & Zhao, P. (2016). Genetic diversity and population structure of common walnut (*Juglans regia*) in China based on EST-SSRs and the nuclear gene phenylalanine ammonia lyase (*PAL*). *Tree Genetics & Genomes* , *12* (6), 111–122.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.*BMC Bioinformatics* , *12* (1), 491.

Hu, Y. H., Woeste, K., & Zhao, P. (2017). Completion of the chloroplast genomes of five Chinese *Juglans* and their contribution to chloroplast phylogeny. *Frontiers in Plant Science* , 7, 1955.

Raval, S., Gowda, S. B., Singh, D. D., & Chandra, N. R. (2005). A database analysis of jacalin-like lectins: sequence-structure-function relationships. *Glycobiology* ,*14* (12), 1247-1263.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., & Nizzari, M. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* , *24* (24), 2938-2939.

Jurka, J. (2000). Repbase Update: A database and an electronic journal of repetitive elements. *Trends in Genetics* , *16* (9), 418-420.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* , *28* (1), 27-30.

Kaori, I., & Murphy, D. (2013). Application of ggplot2 to pharmacometric graphics. *CPT: Pharmacometrics and Systems Pharmacology* , 2(10), e79.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., ... & Natale, D. A. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* , *5* (2), R7.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* , *27* (5), 722-736.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* , *9* (4), 357.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics* , *1303* , 3997.

Li, J., Sun, J. X., Yu, H. Y., Chen, Z. Y., Zhao, X. Y., & Ruan, H. L. (2013). Diarylheptanoids from the root bark of *Juglans cathayensis. Chinese Chemical Letters* , *24* (6), 521-523.

Li, Y. X., Ruan, H. L., Zhou, X. F., Zhang, Y. H., Pi, H. F., & Wu, J. Z. (2008). Cytotoxic Diarylheptanoids from Pericarps of *Juglans cathayensis* Dode. *Chemical Research in Chinese Universities* ,*24* (4), 427-429.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Aacids Research* , *25* (5), 955-964.

Maja T. G, & Chen N. S. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* , 4(4), 10.

Manning, W. E. (1978). The classification within the Juglandaceae.*Annals of the Missouri Botanical Garden* , *65* (4), 1058-1087.

Martinez-Garcia, P. J., Crepeau, M. W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K. A., ... & Neale, D. B. (2016). The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *The Plant Journal* , 87, 507–532.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics, 29* (22), 2933-2935.

Panaro, N. J.,Yuen, P. K., Sakazume, T., Fortina, P., Kricka, L. J., & Wilding, P. (2000). Evaluation of DNA fragment sizing and quantification by the Agilent 2100 Bioanalyzer. *Clinical Chemistry* , *46* (11), 1851-1853.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., ... & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* , *16* (1), 259.

Simao, F. A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* ,*31* (19), 3210-3212.

Shu, Z., Zhang, X., Yu, D., Xue, S., & Wang, H. (2016). Natural hybridization between Persian walnut and Chinese walnut revealed by simple sequence repeat markers. *Journal of the American Society for Horticultural Science* , *141* (2), 146-150.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* , 6(1), 31-40.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* , *34* , W435-W439.

Stevens, K. A., Woeste, K., Chakraborty, S., Crepeau, M. W., Leslie, C. A., Martinez-Garcia, P. J., ... & Kluepfel, D. (2018). Genomic variation among and within six *Juglans* species. *G3-Genes Genomes Genetics* , *8* (7), 2153-2165.

Sun, J. X., Zhao, X. Y., Fu, X. F., Yu, H. Y., Li, X., Li, S. M., ...& Ruan, H. L. (2012). ChemInform abstract:three new naphthalenyl glycosides from the root Bark of *Juglans cathayensis* . *Chemical & pharmaceutical bulletin* , *60* (6), 785-789.

Suryamohan, K., Krishnankutty, S. P., Guillory, J., Jevit, M., Schroder, M. S., Wu, M., ...& Seshagiri, S. (2020). The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nature Genetics* , *52* , 1-12.

Trouern-Trend, A., Falk, T., Zaman, S., Caballero, M., Neale, D. B., Langley, C. H., ... & Wegrzyn, J. L. (2020). Comparative genomics of six *Juglans* species reveals disease-associated gene family contractions. *The Plant Journal* , doi: 10.1111/tpj.14630

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* , *9* (11).

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics* , *8* (1),

77-80.

Wang, Y. P., Tang, H. B., Debarry, J. D., Tan, X., Li, J. P., Wang, X. Y., ... & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* , *40* (7), e49.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* ,*21* (9), 1859.

Xu, Y., Zhang, M. Y., Gao, L., & Liu, J. F. (2007) Development of new lines from walnut interspaces hybrids. *Shandong Agriculture Science* , *3* , 25–27.

Yang, Z. H. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* , *24* (8), 1586-1591.

Yu, H. Y., Li, X., Meng, F. Y., Pi, H. F., Zhang, P., & Ruan, H. L. (2011). Naphthoquinones from the root barks of *Juglans cathayensis* Dode. *Journal of Asian Natural Products Research* ,*13* (7), 581-587.

Zdobnov, E., & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* , *17* (9), 847-848.

Zhao, P., Zhou, H.J., Potter, D., Hu, Y.H., Feng, X.J., Dang, M., ... & Woeste, K. (2018) Population genetics, phylogenomics and hybrid speciation of *Juglans* in China determined from whole chloroplast genomes, transcriptomes, and genotyping-by-sequencing (GBS). *Molecular Phylogenetics and Evolution* , *126* , 250-265.

Zhao, P., Zhao, G. F., Zhang, S. X., Zhou, H. J., Hu, Y. H., & Woeste, K. (2014). RAPD derived markers for separating Manchurian walnut (*Juglans mandshurica*) and Japanese walnut (*J. ailantifolia*) from close congeners. *Journal of Systematics and Evolution* , *52* (1), 101–111.

Zhao, X., & Hao, W. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* , 35, W265-W268.

Zhang, B. W., Xu, L. L., Li, N., Yan, P. C., Jiang, X. H., Woeste, K. E., ... & Bai, W. N. (2019). Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Molecular Biology and Evolution* ,*36* (11), 2451-2461.

Zhang, W., Jiao, Z., Shang, T., & Yang, Y. (2015). Demography and spectrum analysis of *Juglans cathayensis* populations at different altitudes in the west Tianshan valley in Xinjiang, China. *The Journal of Applied Ecology* , *26* (4), 1091–1098.

Zhang, Z., Chen, Y., Zhang, J., Ma, X., Li, Y., Li, M., ... & Ma, T. (2020a). Improved genome assembly provides new insights into genome evolution in a desert poplar (*Populus euphratica*).*Molecular Ecology Resources* , 1-14, https://doi.org/10.1111/1755-0998.13142

Zhang, J., Zhang, W., Ji, F., Qiu, J., Song, X., Bu, D., ... & Chang, Y. (2020b). A high-quality walnut genome assembly reveals extensive gene expression divergences after whole-genome duplication. *Plant Biotechnology Journal* , 1-3. https://doi.org/10.1111/pbi.13350

Zhang, T., Ren, X. Y., Zhang, Z., Ming, Y., Yang, Z., Hu, J. B., ...& Sun, Z. Q. (2020c). Long-read sequencing and de novo assembly of the *Luffa cylindrica* (L.) Roem. genome. *Molecular Ecology Resources* , *20* (2), 511-519.

Zhou, Z., Han, M., Hou, M., Deng, X., Tian, R., Min, S., ...& Zhang, J. (2017). Comparative study of the leaf transcriptomes and ionoms of Juglans regia and its wild relative species *Juglans cathayensis* . *Acta Physiologiae Plantarum* , *39* (10), 224.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHORCONTRIBUTIONS

P.Z conceived and designed the study. F.Y., and P.Z. collected the samples. P.Z. took the morphology picture of Chinese walnut. F.Y., P.P.C., R.M.X., R.X.S., Y.J.Y., and G.Y. and performed the experiments. F.Y., R.X.S., and P.Z. analyzed and interpreted the assembly and annotations. F.Y., P.P.C., M.D., G.Y., and M.Y. supported the software. F.Y. and P. Z. performed the comparative genome analysis. F.Y., and P.Z. performed the whole genome duplication analysis. F.Y., and P.Z. wrote the draft manuscript and then P.Z. and K.W. edited and revised the English writing of this manuscript. All authors contributed and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The whole genome sequence data including Illumina short reads, Nanopore long reads, Hi-C interaction reads, and transcriptome data have been deposited in the NCBI, under accession numbers: PRJNA273527 and PRJNA319865.

Tables:

TABLE 1 Statistics for the Chinese walnut genome assembly and annotation

Characteristics	Statistics
Length of genome (bp)	548,463,652
Contig N50 length (bp)	6,490,758
Scaffold N50 length (bp)	36,084,664
Contig N90 length (bp)	1,434,691
Scaffold N90 length (bp)	23,789,296
Anchored rate (%)	0.99
GC Content (%)	38.51
Raw base (bp)	101,117,316,600
Protein-coding gene number	29,032
Average of mRNA length (bp)	5,734.98
Average of CDS length (bp)	1,226.35
Average of exon number	6.06
Average of exon length (bp)	244.07
Average of intron length (bp)	840.57
Exon number	175,961
Intron number	146,984
Intron length (bp)	123,551,119
Tandem repeats finder	18,999,643 (3.46 %)
Repeat masker	84,059,561 (15.33 %)
Protein mask	101,620,383 (18.53 %)
De novo	332,557,997 (60.65 %)
Total	340,401,005 (62.08 %)

TABLE 2. Genomic footprint of transposable elements in the genome of Chinese walnut

Type	RepBase TEs	RepBase TEs	TE Proteins	TE Proteins	*De novo*	*De novo*	Combined TEs	Com
	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of
DNA	15,960,075	2.91	12,157,908	2.22	39,526,595	7.20	49,110,954	8.96
LINE	16,770,965	3.06	33,789,174	6.16	58,406,142	10.65	67,022,583	12.22
SINE	54,001	0.01	0	0.00	6,518	0.00	58,768	0.01
LTR	52,516,757	9.58	55,824,326	10.18	223,008,440	40.70	226,061,071	41.23

Type	RepBase TEs	RepBase TEs	TE Proteins	TE Proteins	*De novo*	*De novo*	Combined TEs	Com
Total	85,301,798	15.33	101,771,408	19.00	320,947,695	59.00	342,253,376	62.42

Notes : TEs (transposable elements), DNA (Class II TEs), LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), LTR (long terminal repeats), RepBase TEs, TE Proteins, and*De novo* indicated three methods for detecting genomic footprint of transposable elements (details see materials and methods). Combined TEs indicates results based on combined methods of RepBase TEs, TE Proteins, and *De novo* .

List of Figures:

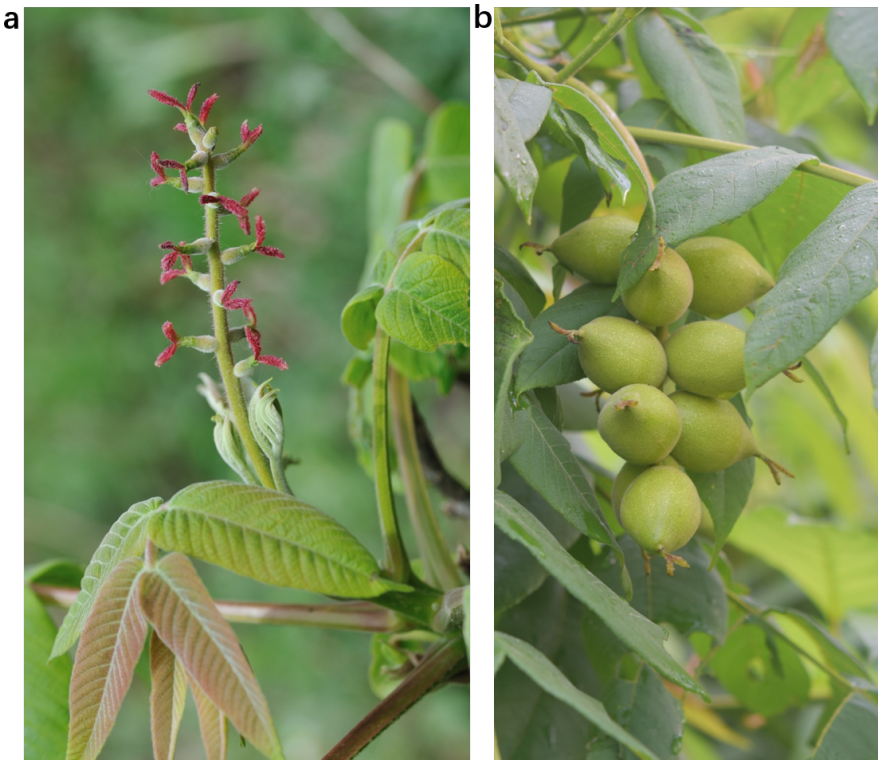

FIGURE 1 . Morphology of *Juglans cathayensis* . (a) female flowers (b) fruits.

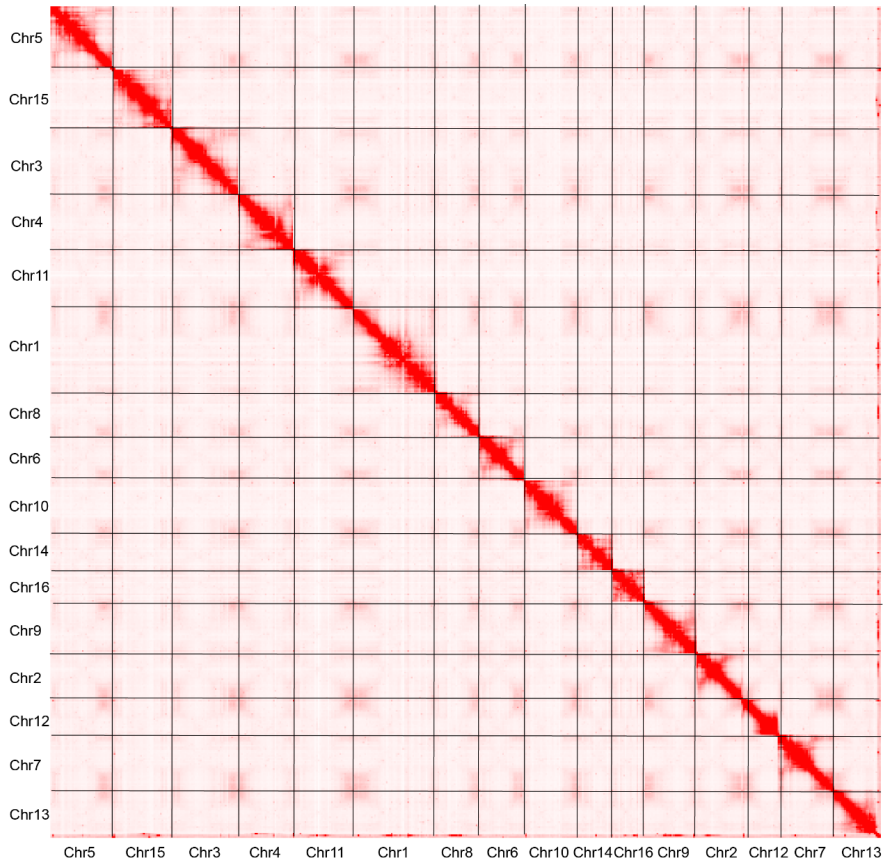

FIGURE 2 . Hi-C interaction heat map between 16 chromosomes of the Chinese walnut genome. X and Y axis indicate that the number of chromosomes.

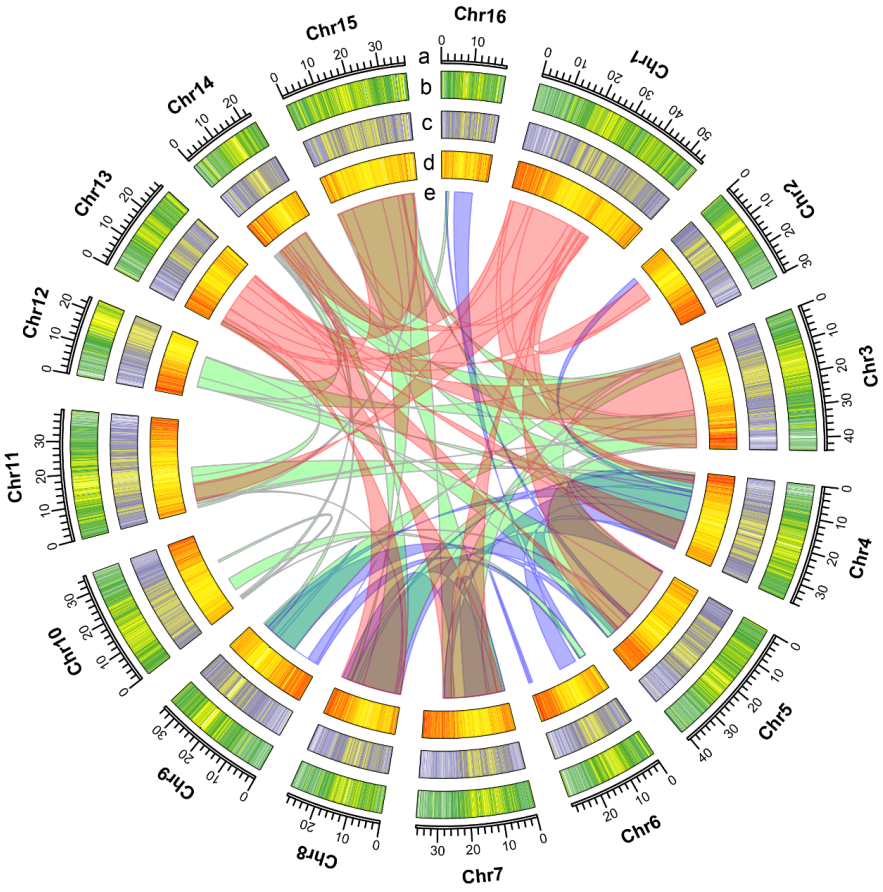

FIGURE 3 Circos plot of the assembled Chinese walnut. Elements are shown in the following scheme (from outer to inner). a Chromosome number; b GC content; c gene density; d transcript heat map; e syntenic relationships among different chromosomes of Chinese walnut.

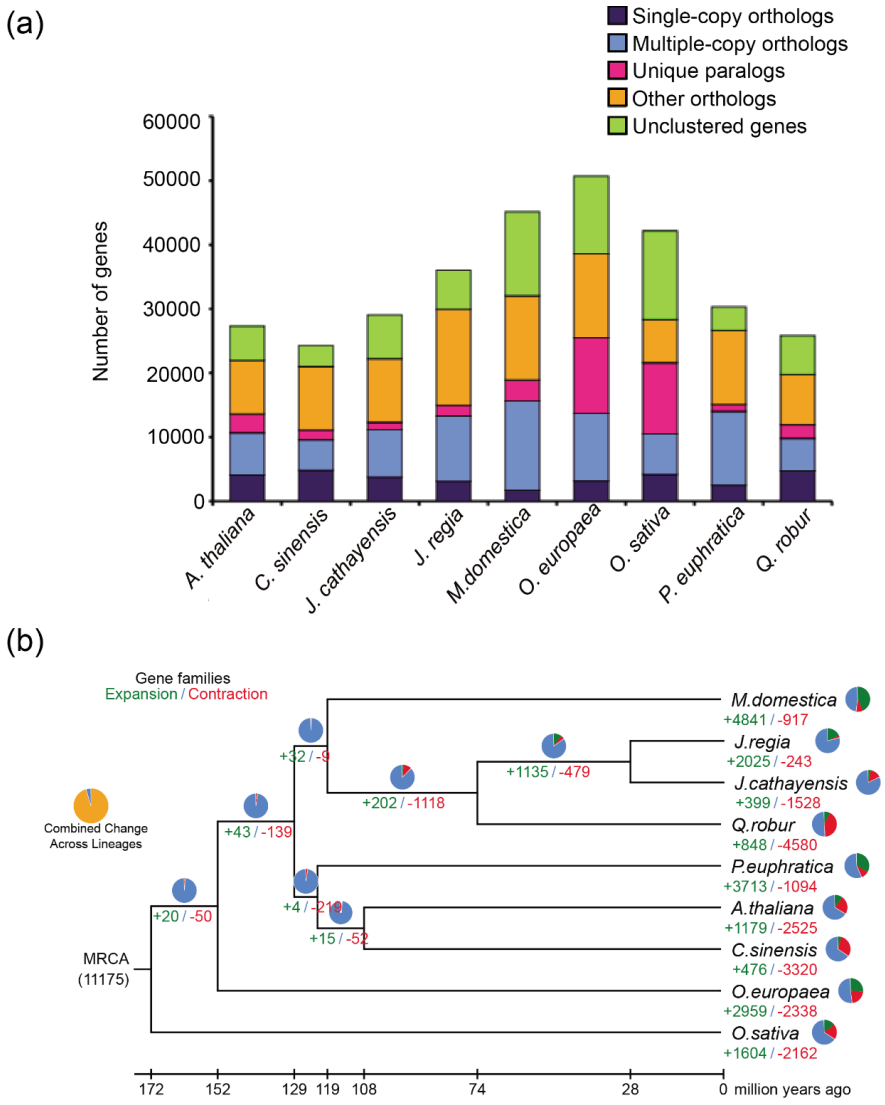

FIGURE 4. Gene family evolution. (a) The proportion of various gene classes among nine species. (b) Expansion and contraction of gene families in nine species. A phylogenetic tree was constructed based on 523 single-copy orthologous genes using *O. sativa* as the outgroup. Pie diagrams on each branch of the tree represent the proportion of genes undergoing gain (green) or loss (red) events, the numbers near the nodes represent number of gene families expanded or contracted. The scale on the X axis shows the estimated divergence time for nodes.

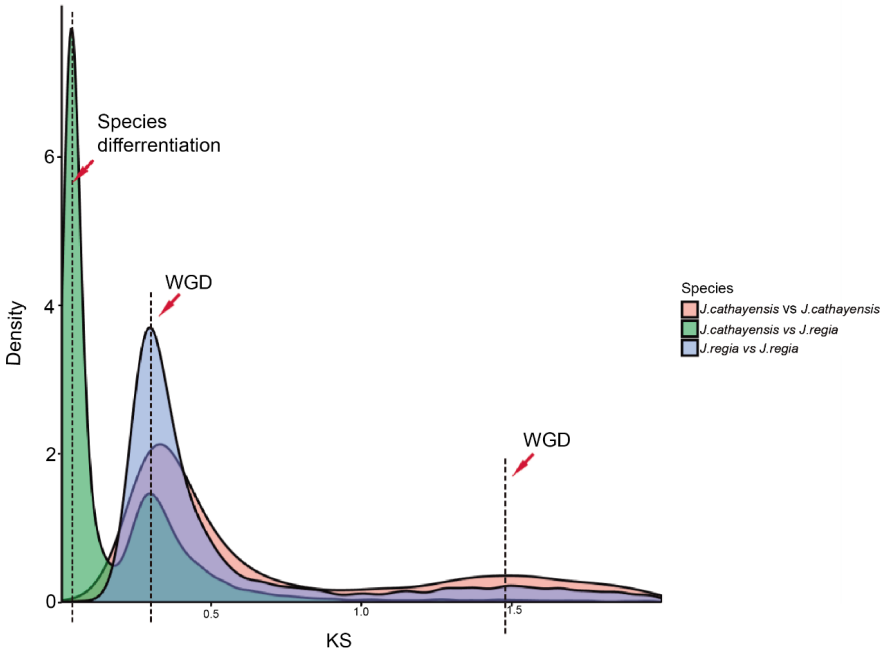

FIGURE 5. The WGD events of Chinese walnut and Persian walnut. Distribution of synonymous substitution rate (KS) for syntenic genes from *J. cathayensis* and *J. regia* . Two whole-genome duplication (WGD) events were indicated by the peaks.