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

Feng Yan¹, Rui-Min Xi¹, Rui-Xue She¹, Yu-Jie Yan¹, Peng-Peng Chen¹, Ge Yang¹, Meng Dang¹, Ming Yue¹, Keith Woeste², and Peng Zhao³

¹Affiliation not available
²USDA Forest Service
³Northwest 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). 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 predicted 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 predicted 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 sequences 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 = $1 \times 10^{-5}$) 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 $1 \times 10^{-5}$) 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 (Koari 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 518,392,333 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.

REFERENCES


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.


**CONFLICT OF INTEREST**

The authors declare that they have no competing interests.
AUTHOR CONTRIBUTIONS

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.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

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

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

<table>
<thead>
<tr>
<th>Type</th>
<th>RepBase TEs Length (bp)</th>
<th>% of genome</th>
<th>RepBase TEs Length (bp)</th>
<th>% of genome</th>
<th>TE Proteins Length (bp)</th>
<th>% of genome</th>
<th>TE Proteins Length (bp)</th>
<th>% of genome</th>
<th>De novo Length (bp)</th>
<th>% of genome</th>
<th>De novo Length (bp)</th>
<th>% of genome</th>
<th>Combined TEs Length (bp)</th>
<th>% of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>15,960,075</td>
<td>2.91</td>
<td>12,157,908</td>
<td>2.22</td>
<td>39,526,595</td>
<td>7.20</td>
<td>49,110,954</td>
<td>8.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINE</td>
<td>16,770,965</td>
<td>3.06</td>
<td>33,789,174</td>
<td>6.16</td>
<td>58,406,142</td>
<td>10.65</td>
<td>67,022,583</td>
<td>12.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SINE</td>
<td>54,001</td>
<td>0.01</td>
<td>0</td>
<td>0.00</td>
<td>6,518</td>
<td>0.00</td>
<td>58,768</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTR</td>
<td>52,516,757</td>
<td>9.58</td>
<td>55,824,326</td>
<td>10.18</td>
<td>223,008,440</td>
<td>40.70</td>
<td>226,061,071</td>
<td>41.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>RepBase TEs</td>
<td>RepBase TEs</td>
<td>TE Proteins</td>
<td>TE Proteins</td>
<td>De novo</td>
<td>De novo</td>
<td>Combined TEs</td>
<td>Combined TEs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>85,301,798</td>
<td>15.33</td>
<td>101,771,408</td>
<td>19.00</td>
<td>320,947,695</td>
<td>59.00</td>
<td>342,253,376</td>
<td>62.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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](image-url)  
**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.