Received: 11 May 2020 | Revised: 15 March 2021 | Accepted: 29 March 2021

DOI: 10.1111/1755-0998.13394

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Improved de novo chromosome-level genome assembly of the vulnerable walnut tree *Juglans mandshurica* reveals gene family evolution and possible genome basis of resistance to lesion nematode

Feng Yan[1] | Rui-Min Xi[1] | Rui-Xue She[1] | Peng-Peng Chen[1] | Yu-Jie Yan[1] | Ge Yang[1] | Meng Dang[1] | Ming Yue[2] | Dong Pei[3] | Keith Woeste[4] | Peng Zhao[1]

[1]Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, College of Life Sciences, Northwest University, Xi'an, China

[2]Xi'an Botanical Garden of Shaanxi Province, Xi'an, China

[3]State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding and Cultivation of the State Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China

[4]Department of Forestry and Natural Resources, USDA Forest Service Hardwood Tree Improvement and Regeneration Center (HTIRC), Purdue University, West Lafayette, IN, USA

Correspondence
Peng Zhao, Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, College of Life Sciences, Northwest University, Xi'an, China.
Email: pengzhao@nwu.edu.cn

Funding information
Natural Science Foundation of Shaanxi Province of China, Grant/Award Number: 2019JM-008; Shaanxi Academy of Science Research Funding Project, Grant/Award Number: 2019K-06; Opening Foundation of Key Laboratory of Resource Biology and Biotechnology in Western China (Northwest University), Ministry of Education, Grant/Award Number: ZSK2018009; National Natural Science Foundation of China, Grant/Award Number: 32070372, 41471038 and 31200500

Abstract

Manchurian walnut (*Juglans mandshurica* Maxim.) is a synonym of *J. cathayensis*, a diploid, vulnerable, temperate deciduous tree valued for its wood and nut. It is also valued as a rootstock for *Juglans regia* because of its reported tolerance of lesion nematode. Reference genomes are available for several *Juglans* species, our goal was to produce a de novo, chromosome-level assembly of the *J. mandshurica* genome. Here, we reported an improved assembly of *J. mandshurica* with a contig N50 size of 6.49 Mb and a scaffold N50 size of 36.1 Mb. The total genome size was 548 Mb encoding 29,032 protein coding genes which were annotated. The collinearity analysis showed that *J. mandshurica* and *J. regia* originated from a common ancestor, with both species undergoing two WGD events. A genomic comparison showed that *J. mandshurica* was missing 1657 genes found in *J. regia*, and *J. mandshurica* includes 2827 genes not found in of the *J. regia* genome. The *J. mandshurica* contained 1440 unique paralogues that were highly enriched for flavonoid biosynthesis, phenylpropanoid biosynthesis, and plant-pathogen interaction. Four gene families related to disease resistance notable contraction (rapidly evolving; *LEA*, *WAK*, *PPR*, and *PR*) in *J. mandshurica* compared to eight species. *JmaPR10* and *JmaPR8* contained three orthologous gene pairs with *J. regia* that were highly expressed in root bark. *JmaPR10* is a strong candidate gene for lesion nematodes resistance in *J. mandshurica*. The *J. mandshurica* genome should be a useful resource for study of the evolution, breeding, and genetic variation in walnuts (*Juglans*).

KEYWORDS
gene family evolution, genome assembly and annotation, lesion nematode, Manchurian walnut, Nanopore sequencing, PR gene family

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). *J. mandshurica* Maxim is an ecologically important, wind pollinated, endemic species that grows in northern and northeastern China, Korea, Japan, and the far eastern section of Russia (Bai et al., 2010, 2014; Hu et al., 2016; Lu, 1982). It is a synonym of *J. cathayensis* Dode, 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*; Bai et al., 2016; Zhao et al., 2014, 2018).

The population genetics, morphology, and diversity of *J. mandshurica* have been described (Aradhya et al., 2007; Dang et al., 2015; Hu et al., 2017; Liu et al., 2020; Manning, 1978; Zhang et al., 2019), but in general, interest in *J. mandshurica* based on its potential as a tertiary germplasm pool for improvement of *J. regia* (Chen et al., 2015; Hu et al., 2016; Ji et al., 2020; Trouern-Trend et al., 2020; Zhou et al., 2017). Wild populations of *J. mandshurica* 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). *J. mandshurica* is less valuable as a commodity than its close relative *J. regia* (Dang et al., 2016; Feng et al., 2018; Han et al., 2016). However, *J. mandshurica* expresses horticultural traits such as cluster bearing habit (6–13 fruits per terminal) that make it attractive to *J. regia* breeders, and disease tolerance/resistance to lesion nematodes (*Pratylenchus vulnus*) that recommend it as a rootstock for *J. regia* (Chen et al., 2015; Hu et al., 2016; Ji et al., 2020; Trouern-Trend et al., 2020; Zhou et al., 2017). *J. mandshurica* is also a potential medicinal crop because of its flavonoids (Bi et al., 2016; Sun et al., 2012; Yu et al., 2011).

A high-quality genome is an important genetic resource for the improvement of horticultural traits in perennial crops (Dong et al., 2019; Zhang, Chan, et al., 2020). The availability of high-throughput sequencing has accelerated the publication of the genomes of walnut (*Juglans*) species and hybrids (*J. regia* × *J. microcarpa*; Bai et al., 2018; Martínez-García et al., 2016; Stevens et al., 2018; Zhang, Zhang, et al., 2020). 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. mandshurica* are minimal. For these reasons, we undertook the assembly of a chromosome-level, high-quality reference genome assembly for *J. mandshurica* as well as the complete annotation of its expressed proteins, structural RNAs, miRNA and repeat regions.

2 | MATERIALS AND METHODS

2.1 | *J. mandshurica* sample collection and genomic DNA extraction

In 2019, we collected leaf samples from a single individual of *J. mandshurica* (wild individual "Tree8C22 N") growing in the Qinling 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).

2.2 | Illumina short-read sequencing

J. mandshurica was sequenced on the Illumina HiSeq X Ten platform using 20 kb 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".

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, and we then aligned the Illumina reads to the draft genome assemblies using BWA-MEM (Li, 2013). Finally, a total of 62.87 Gb of reads from Nanopore sequencing were used to assemble after assessment and error correction (Table S1).

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 & 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 and expression analysis

RNA was extracted from 18 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) collected from individual "Tree8C22N", the same tree described above for DNA sequencing (Figure 1a; Table S2). An RNA-Seq library was produced for each tissue using an NEBNext Ultra RNA Library Prep Kit (NEB). Paired end sequencing was performed on Illumina HiSeq X Ten platform (Illumina). After RNA quantification, we also pooled equivalent amounts of RNA from each of the 18 tissues for full-length transcriptome sequencing. Using the purified mRNA as the starting material, a full-length cDNA library (10–15 kb) was constructed for the PacBio Sequel platform (NEB, USA). Bioanalyzer 2100 software (Panaro et al., 2000) was used to test the library quality.

To estimate the expression levels of *J. mandshurica* genes in different tissues and during various developmental stages, clean transcriptome sequencing reads were aligned to the *J. mandshurica* genome using Bowtie2 (Langmead & Salzberg, 2012). The read number of each transcript was calculated using RSEM (Li & Dewey, 2011). The number of fragments per kb of transcript sequence per million bp sequenced value (FPKM) was estimated to measure the expression of each gene (Trapnell et al., 2010). A total of 15 transcriptome data were used to estimate the expression levels of *J. regia* genes (Martínez-García et al., 2016; Table S2).

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

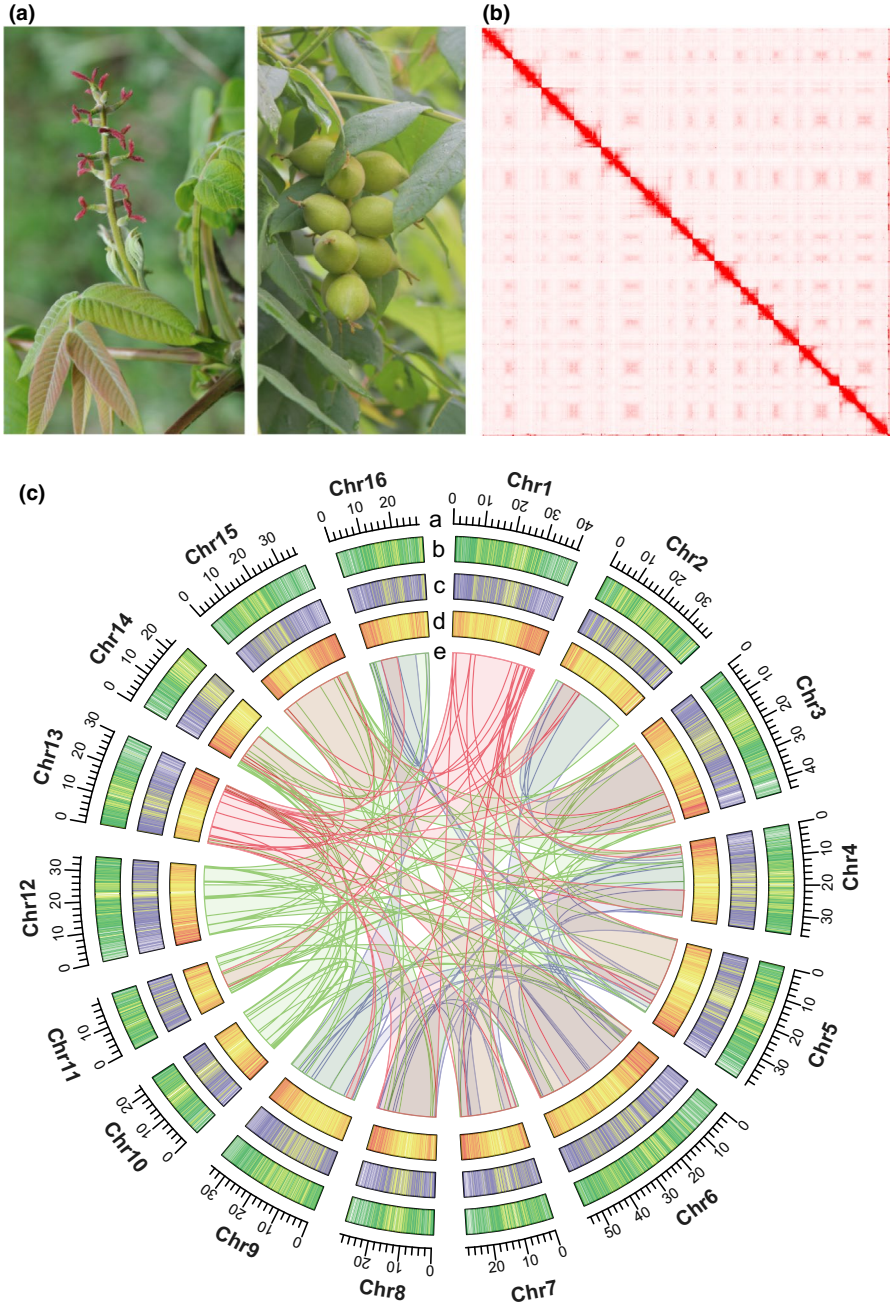

FIGURE 1 The characteristics of morphology and genome of *J. mandshurica*. (a) Morphology of Manchurian walnut: female flowers (left) and fruits (right). (b) Hi-C interaction heat map between 16 chromosomes of the *J. mandshurica* genome. (c) Circos plot of the assembled *J. mandshurica*. Elements are shown in the following scheme (from outer to inner). (a) chromosome number; (b) guanine-cytosine (GC) content; (c) gene density; (d) transcript heat map; (e) syntenic relationships among different chromosomes of *J. mandshurica*

using bench marking universal single-copy orthologues (BUSCO) version 3.0.2 (Simão et al., 2015). Genome completeness was further evaluated by mapping of transcripts from 18 (Table S2) tissues and organs using GMAP (Wu & Watanabe, 2005).

2.7 | Genome annotation

We annotated repeat sequences, gene structure, and non-coding RNA in the *J. mandshurica* genome (workflow, Figure S2). We used both homology-based 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 & Chen, 2009), LTR-FINDER (Zhao & Hao, 2007) and PILER (Edgar & Myers, 2005) with default parameters. For homology-based prediction, we identified transposable elements in the DNA based on predicted proteins by comparing genomic sequence with the REPBASE version 21.12 database (Jurka, 2000) using REPEATMASKER (Maja & Chen, 2009) and REPEATPROTEINMASK version 4.0.7 (Maja & Chen, 2009). Finally, all transposable elements identified by either method were merged into the final transposon annotations. Transposable elements (TEs) in the assembled *J. mandshurica* genome were also annotated using TANDEM REPEATS FINDER (TRF) version 4.09 (Benson, 1999).

To ensure accurate gene structure annotations, we combined homology prediction and de novo prediction methods. RNA sequences from 18 tissues (Table S2) 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* (GCA_000001735.2), *Citrus sinensis* (GCA_000317415.1), *J. regia* (GCA_000001735.2), *Malus domestica* (GCA_002114115.1), *Olea europaea* (GCA_902713445.1), *Oryza sativa* (GCA_014636035.1), *Populus euphratica* (GCA_000495115.1), *Quercus robur* (GCA_000001735.2), and *J. mandshurica* using EXONERATE version 2.2.0 (Slater & Birney, 2005). The final structural annotation of protein-coding genes was performed using a MAKER (Holt & Yandell, 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 version 2.2.3 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 follow databases (*E*-value = $1e^{-5}$), including SWISSPROT (Boeckmann et al., 2003), TREMBL (Boeckmann et al., 2003), KEGG (Kanehisa & Goto, 2000), INTERPRO (Zdobnov & Apweiler, 2001), SWISSPROT (Bairoch & Apweiler, 2000), KOG (Koonin et al., 2004), GO (Ashburner et al., 2000), and KEGG enrichment analysis (Yu et al., 2012).

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 & Eddy, 1997) based on the RNA structure. The rRNA sequences in the *J. mandshurica* genome were predicted using BLASTN to search for conserved characteristics with related species such as *J. regia*. The miRNA and snRNA in the assembled *J. mandshurica* genome were identified using INFERNAL software (Nawrocki & Eddy, 2013) against the RFAM 13.0 database (Griffiths-Jones et al., 2005).

2.10 | The detection of insertions and deletions in *J. mandshurica* versus *J. regia*

Deletions and insertions between the *J. mandshurica* and the *J. regia* assemblies were detected using the Assemblytics suites (Nattestad & Schatz, 2016). Initially, the *J. regia* genome was used as the reference to align the *J. mandshurica* assemblies using the program NUMMER4 (Marais et al., 2018). The delta files were then uploaded onto the online Assemblytics analysis pipeline (Nattestad & Schatz, 2016).

2.11 | Genome duplication and synteny analyses

To estimate the timing of whole-genome duplication events (WGD) in the *J. mandshurica* genome, reciprocal best hit (RBH) gene pairs were identified (E-value is $1e^{-5}$) based on all-versus-all paralogues detected in BLASTP (Altschul et al., 1990). We identified synteny blocks and collinear blocks of gene pairs in the *J. mandshurica* genome using MCScanX with default parameters (Wang et al., 2012). The synonymous substitution rate (Ks) was calculated using the YN model in KAKS _ CALCULATOR version 2.0 (Wang et al., 2010). The Ks distributions of orthologues within *J. mandshurica* and *J. regia*, and between *J. mandshurica* and *J. regia* were used to compare the relative substitution rates in different species by plotting with the GGPLOT2 package (Kaori & Murphy, 2013).

2.12 | Gene family cluster identification

Nine species (*A. thaliana*, *C. sinensis*, *J. regia*, *M. domestica*, *O. europaea*, *O. sativa*, *P. euphratica*, *Q. robur*, and *J. mandshurica*) were selected for comparative genome analysis. All-versus-all BLASTP (Altschul et al., 1990) search results (E-value = $1e^{-5}$) were used for gene family construction using ORTHOMCL (Fischer et al., 2011). A

maximum likelihood (ML) phylogenetic tree was constructed using RAXML version 8.2.12 (Stamatakis, 2014) by conducting 1000 bootstrap replicates using single-copy orthologues. Species divergence times were estimated using MCMCTREE (Yang, 2007) with the following parameters: 10,000 burnins, 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/). Computational analysis of gene family evolution (CAFE) version 2.2 (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.

2.13 | Genome-wide analysis of evolution and expression profiles of gene family

Based on the results of CAFE and the species' resistance traits, we selected four rapidly evolving gene families, including late embryogenesis abundant protein (*LEA*), wall-associated receptor kinase (*WAK*), PPR repeat (*PPR*), and pathogenesis-related protein (*PR*), and identified their gene family members in nine species (*A. thaliana*, *C. sinensis*, *J. regia*, *M. domestica*, *O. europaea*, *O. sativa*, *P. euphratica*, *Q. robur*, and *J. mandshurica*). The sequence of *LEA* (CCO06495.1), *WAK* (QCE08590.1), *PPR* (ABW04887.1), and *PR* (ABA41593.1) were used as queries in a BLASTP search against nine protein databases to identify candidate orthologues. The BLASTP parameters were *E*-value <1e−5, identity ≥50%, and coverage ≥50% (Altschul et al., 1990). Protein domains in the candidate sequences were determined using PFAM (Finn et al., 2008), only proteins with *LEA*, *WAK*, *PPR*, and *PR* domains were retained.

To detect the *PR10* members in *J. mandshurica* and *J. regia*, we download a total of 17 *PR10* members from NCBI (details see Table S3), and combined with the all *PR* genes in *J. mandshurica*, *J. regia*, and *A. thaliana* to construct a phylogenetic tree with MEGA (Kumar et al., 2008). To search for the presence of potential domains of *PR* genes using the PFAM webserver (El-Gebali et al., 2018). A conserved domain database search was conducted in NCBI (Marchler-Bauer et al., 2016). The exon and intron structures were displayed using the online gene structure display server (Hu et al., 2015). The heatmap was visualized with the TBtools (Chen et al., 2020).

3 | RESULTS

3.1 | Improvement of *J. mandshurica* genome assembly and annotation

To obtain a high-quality genome assembly, we first sequenced a total of ~47.3 Gb clean reads (equivalent to ~82× genome coverage) to assemble the *J. mandshurica* genome based on Illumina HiSeq X-Ten sequencing (Table S4). We then called a total of 62.87 Gb long reads (~118 × genome coverage) from the *J. mandshurica* genome using Oxford Nanopore Technology sequencing platform (Table S1). A total of 101 Gb raw data of a chromosome conformation capture (Hi-C) was produced by the Nanopore sequencing platform (~176 × genome coverage; Table S5).

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 S6). We identified 1375 complete BUSCOs, including 104 duplicated BUSCOs, 71 fragmented BUSCOs, and 1160 single-copy orthologues in the assembled *J. mandshurica* genome (Table S7). There were 40 genes recognized as missing BUSCOs in the assembled genome (Table S7). Overall, we obtained ~548 Mb of *J. mandshurica* genome based on long reads, which is about 94.8% of the survey genome (578.1 Mb; Table 1).

A total of 0.54 Gb assembled scaffold sequence was divided into 16 groups corresponding to the 16 *J. mandshurica* chromosomes (Figures 1b and 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 (Table 1). Hi-C sequence (543 Mb) was mapped and anchored (99%; 543 Mb/548 Mb) to the assembled 16 chromosomes of the *J. mandshurica* genome (Table 1). Chromosome numbering for *J. mandshurica* was based on homology to the numbering of *J. regia* chromosomes (Zhang, Zhang, et al., 2020; Table S8). The lengths of the 16 assembled chromosomes of *J. mandshurica* ranged from 19,675,958 to 55,052,647 bp with a mean length of 33,963,507 bp, while chromosomes of *J. regia* ranged from 20,184,194 to 518,39,233 bp with a mean length of 33,799,624 bp (Table S8).

We identified 340.4 Mb of repeats (62.1% of the genome) in the *J. mandshurica* genome, of which ~62.42% were transposable elements (TEs; Tables 1 and 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).

A combination of ab initio prediction, homology search, and transcript mapping were used to predict the protein-coding genes in the *J. mandshurica* genome. RNA from 18 tissues was used to predict gene models (Table S2). Predicted protein-coding genes (27,901) had an average gene length of 5735 bp, an average coding sequence (CDS) length of 1226 bp, and an average of six exons per gene (Table 1). When we compared *J. mandshurica* to *A. thaliana* based on genome structural features, we found the distribution of CDS lengths (exon lengths) of *J. mandshurica* was similar to *A. thaliana*; however, the distribution of mRNA lengths and intron lengths of *J. mandshurica* was unlike *A. thaliana* (Table 1; Figure S3). Among 27,901 predicted genes, 96.1% could be functionally annotated in at least one of seven databases (Table S9). There were 2014 genes annotated in NR database only, 23 genes annotated in InterPro only, six genes annotated in KEGG only, and no gene was annotated in swissProt or COG only (Figure S4). The average guanine-cytosine (GC) content was 51.21% (Figure 1c). Gene density throughout the genome was about 11 genes per 100 kb, with 56,553 genes (94.96%) present on

chromosomally anchored contigs (Figure 1c); this was equivalent to 307 transcripts per 1 Mb of chromosome (Figure 1c). There are 82 syntenic blocks in the *J. mandshurica* genome (Figure 1c). The portion of the *J. mandshurica* genome comprised of non-coding RNA was small; it included miRNA, tRNA, rRNA, and snRNA (Table S10).

TABLE 1 Statistics for the *Juglans mandshurica* 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%)

A total of 581 tRNA (Table S10), 792 small nuclear RNA (snRNA) and 132 microRNA (miRNA) were identified (Table S10).

3.2 | Genome comparison between *J. mandshurica* and *J. regia*

The genomes of *J. mandshurica* and *J. regia* were compared based on whole-genome duplication events (WGD), collinearity, the chromosomal distribution of repeats, repeat expansion, gene density, and CDS density (Figure 2). CDS density of *J. mandshurica* genome was 480 genes per 100 kb which higher than 438 genes in *J. regia*. Gene density throughout the *J. mandshurica* genome was about 19 genes per 100 kb versus 30 genes per 100 kb in *J. regia*. The repeat contraction per 100 kb in *J. mandshurica* was 4.3 versus 3.9 in *J. regia*, the repeat expansion per 100 kb in *J. mandshurica* was 3.8 versus 4.3 in *J. regia*. These variables summarize some of the structural differences between the two genomes (Figure 2a–b).

We identified a total of 86 synteny blocks and 5614 genes in all blocks that covered 20.1% of *J. mandshurica* genome (Figure S5). The peak of Ks at ~0 for orthologous gene pairs between *J. mandshurica* and *J. regia* genomes reflects recent species differentiation (Figure 2c). The *J. mandshurica* and *J. regia* genomes showed a high degree of synteny on each chromosome, a further sign of the quality of our *J. mandshurica* genome assembly (Figure S5). The comparison of *J. mandshurica* with *J. regia* revealed large-scale inversions on chromosome 7, 13, and 16 (Figure 2d). Gene function annotation results showed that many of the genes in the inversions were related to disease resistance, including members of GDSL-like lipase (*GDSL*), glutathione s-transferase (*GST*), ABC transporter transmembrane (*ABC*), myb DNA-binding domain (*MYB*), leucine rich repeat (*NBS-LRR*), and PPR repeat (*PPR*) gene families (Figure 2d; Table S11).

We characterized the insertions and deletions (InDels) in the genome of *J. mandshurica* compared to *J. regia* (Figure 3), which totalled 28.1 Mb (4.8%) in *J. mandshurica* (Figure 3a; Table S12). Chromosome 1 of *J. mandshurica* contained the most deletions, and chromosome 15 was the lowest density of deletion events, chromosome 9 was unusually enriched for insertions (Figure 3a–b). The most common

TABLE 2 Genomic footprint of transposable elements in the genome of *Juglans mandshurica*

	RepBase TEs		TE proteins		De novo		Combined TEs	
Type	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
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
Total	85,301,798	15.33	101,771,408	19.00	320,947,695	59.00	342,253,376	62.42

Abbreviations: DNA, Class II TEs; LINE, long 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; SINE, short interspersed nuclear element; TEs, transposable elements.

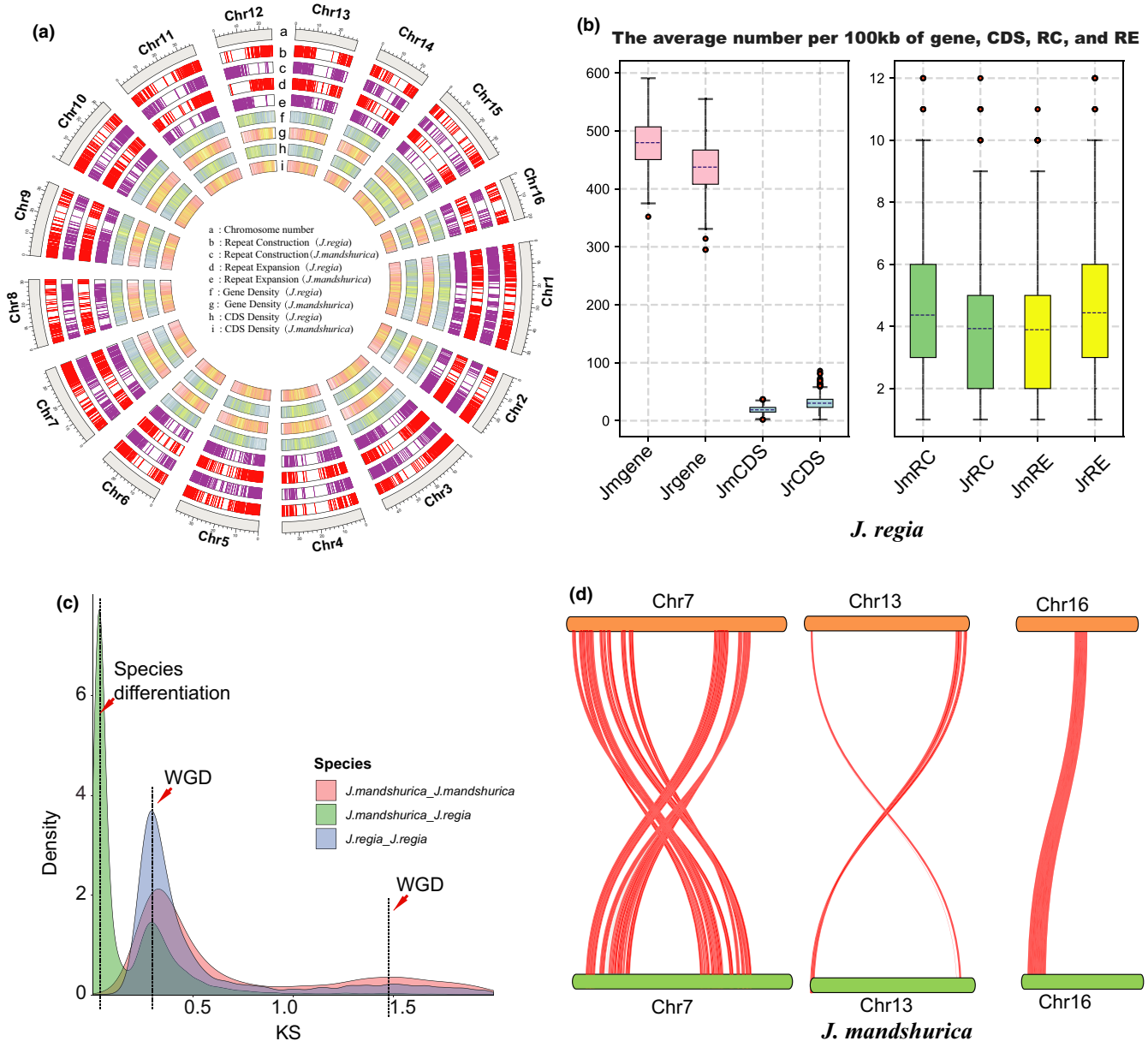

FIGURE 2 The comparative analysis of genome characteristics between *J. mandshurica* and *J. regia*. (a) The Circos plot of variation of *J. regia* and *J. mandshurica*. (b) Comparative of genome per 100,000 bp window for each characterization, including CDS (coding sequence) density, gene density, and repeat construction, repeat expansion within whole genome. The pink box indicates that CDS, the blue box indicates that gene, the green box indicates that RC (repeat construction), the orange box indicates that RE (repeat expansion). The RC and RE result from assemblytics pipline between *J. mandshurica* and *J. regia*. (c) The whole-genome duplication (WGD) events of *J. mandshurica* and *J. regia*. Distribution of synonymous substitution rate (KS) for syntenic genes from *J. mandshurica* and *J. regia*. Two WGD events were indicated by the peaks. (d) The inversion events between *J. mandshurica* and *J. regia*. The red lines represent inversion events

InDels sizes ranged from 100 to 500 bp. InDels >5 kb were about 4.7% of the total (2766/58,585), but InDels >10 kb in size were rare across the *J. mandshurica* genomes (Figure 3c; Table S12). The result revealed that more than 28% InDels was 100–500 bp-sized InDels in genomes (Figure 3c; Table S12). Rare, large InDels in *J. mandshurica* were associated with gain or loss of genes (Table S13–S14). A total of 1657 gene deletions and 2827 gene insertions were identified in *J. mandshurica* compared to *J. regia* (Figure 3; Tables S13–S14). The deleted genes were functionally enriched for biosynthesis of amino acids (Figure S6A), whereas the inserted genes were related to phenylpropanoid biosynthesis (Figure S6B). It is possible that enrichment in phenylpropanoid biosynthesis contributes to pest and disease resistance traits of *J. mandshurica* (Figures 3 and S6).

3.3 | Unique paralogues function and gene family evolution of *J. mandshurica*

We searched for single-copy orthologues in the genome of *J. mandshurica* as compared to eight other genomes (i.e., *A. thaliana*,

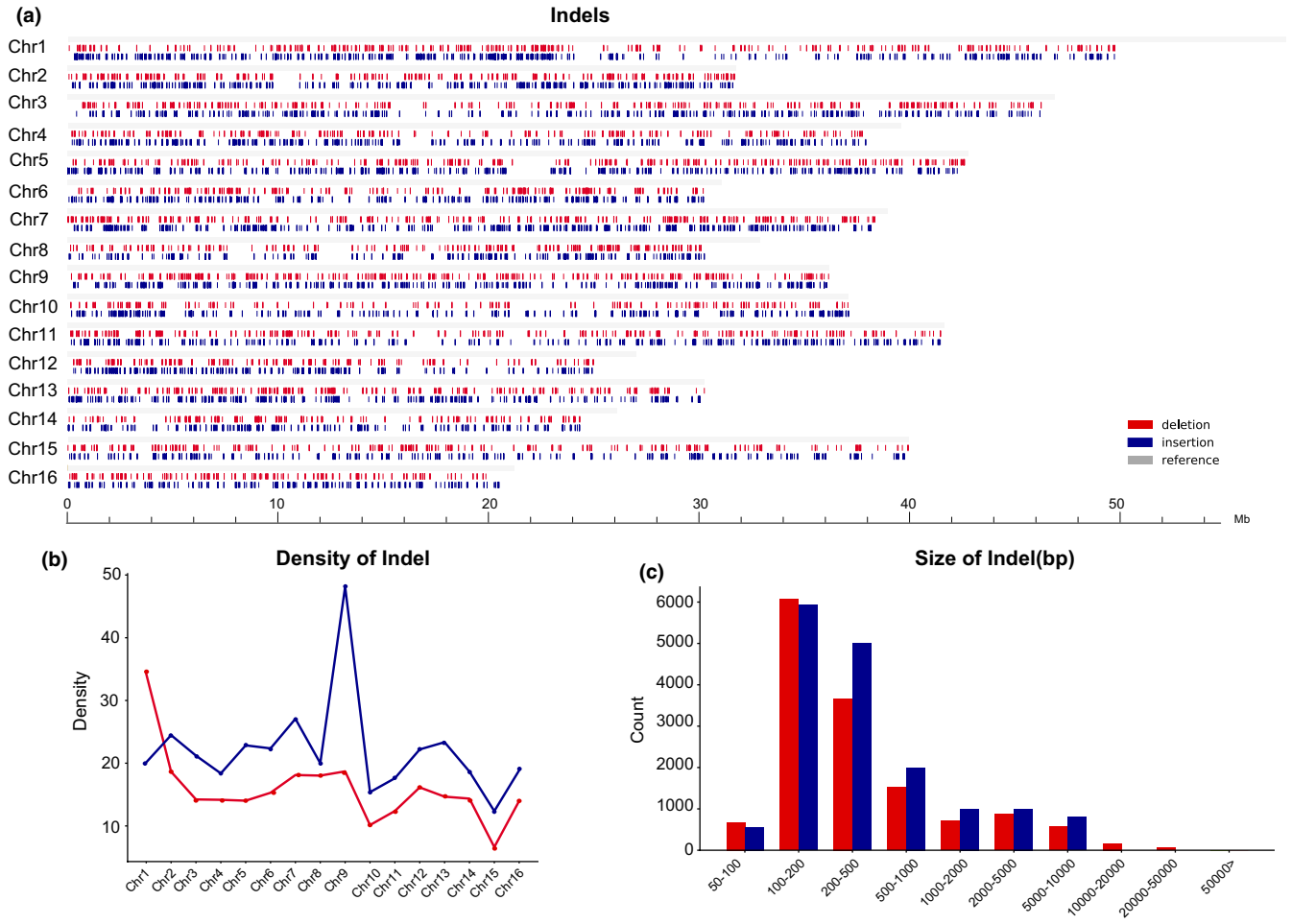

FIGURE 3 Deletion and insertion events across the *J. mandshurica* genome and *J. regia* assemblies. (a) Chromosome-wide distribution of deletion and insertion variation for each *J. mandshurica* genome, relative to the *J. regia* coordinates. (b) The insertion density of each chromosome. (c) Distribution of the deletion and insertion sizes compared to *J. regia*

C. sinensis, *J. regia*, *M. domestica*, *O. europaea*, *O. sativa*, *P. euphratica*, *Q. robur*). This comparison was intended to identify orthologues that may contribute to the distinctiveness of *J. mandshurica* as a species. Within the group of nine species, we identified 125,530 orthologous gene families that consisted of 310,273 genes (Figures 4a and S7; Table S15). The percentage of the *J. mandshurica* genome occupied by single-copy orthologues was higher than all other species in the comparison except *Q. robur* and *C. sinensis* (Figure S7). We found 17.4% (10,321/59,377) of all gene orthologues were common to *J. mandshurica*, *J. regia*, *M. domestica*, and *O. europaea* (Figure S7), and 1704 (5%) orthogroups were specific to the two *Juglans* species (*J. mandshurica* and *J. regia*; Figure S7). KEGG functional analysis of 1440 unique paralogues of *J. mandshurica* were four pathways, flavonoid biosynthesis, phenylpropanoid biosynthesis, plant–pathogen interaction, and fatty acid degradation, which could have a role in pest or disease resistance in *J. mandshurica* (Figure 4a–b; Tables S16–S17). Unique paralogues for *J. regia* (selected by humans for nut production) were functionally involved in cutin, suberin, and wax biosynthesis and fatty acid metabolism (Figure 4a–b; Tables S16–S17). Unique paralogues for fatty acid metabolism were enriched in *M. domestica* and fatty acid biosynthesis in *O. europaea*, but others for fatty acid degradation were found in *J. mandshurica* (Figures 4a–b and S8; Tables S16–S17).

The expansion or contraction of gene families has a profound role in adaptive evolution in plants. Compared with nine representative species, 399 gene families were expanded, 1528 were contracted, and 58 were rapidly evolving gene families (+9/−49) in the *J. mandshurica* genome (Figure 4c; Table S18). The genome of *J. regia* contained expanded in 2025 gene families, contracted in 243, and 57 were rapidly evolving gene families (+50/−7; Figure 4c; Table S18). In a comparison of the two walnut genomes, we found that gene families associated with pathogen resistance, including wall-associated receptor kinase (*WAK*; Trouern-Trend et al., 2020), late embryogenesis abundant protein (*LEA*; Gao et al., 2020), pathogenesis-related protein (*PR*; Ozyigit et al., 2017; Soh et al., 2012; Zhao et al., 2015), and PPR repeat (*PPR*; Liu et al., 2016) were significantly contracted (rapidly evolving; family-wide *p*-value ≤ .01) in *J. mandshurica* (Figure 4d; Table S18). The nine plant species genomes we studied in detail were highly divergent in terms of the amount of expansion or contraction in these four gene families; even the closely related species *J. regia* and *J. mandshurica* were markedly different in terms of levels of expansion. For example, *WAK* gene family members expanded in *J. regia* whereas they contracted in

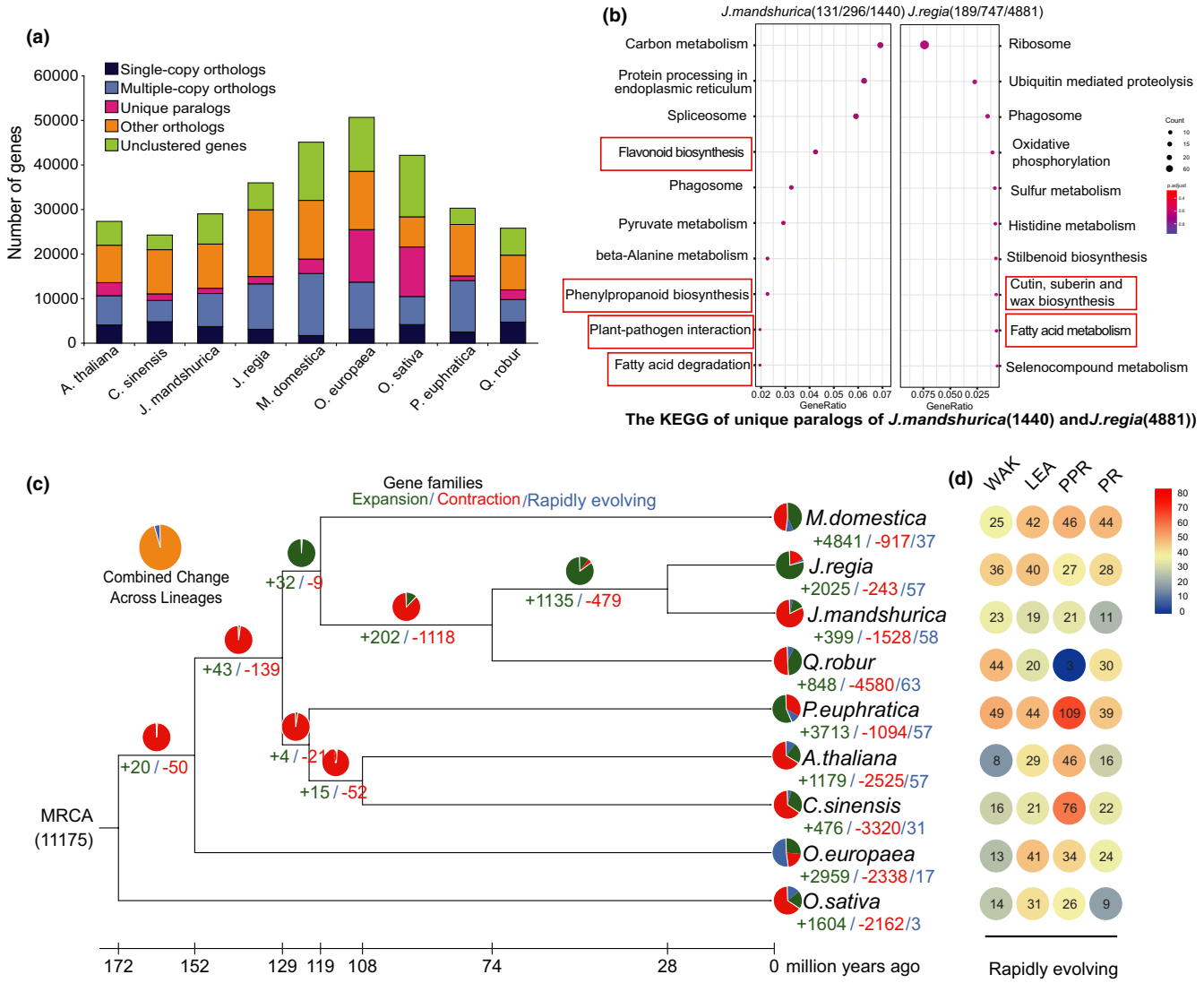

FIGURE 4 The *J. mandshurica* genome evolution. (a) The proportion of various gene classes among nine species, including single-copy orthologues, multicopy genes, unique paralogues, other paralogues, and unclustered genes. (b) The KEGG enrichment analysis of unique paralogues of *J. mandshurica* and *J. regia*. (c) Expansion, contraction, and rapidly evolving within 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), loss (red), and rapidly evolving (blue) 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. "+" indicates that gene families expanded, "−" indicates that gene families contracted. (d) The number of *WAK*, *LEA*, *PR*, and *PPR* of rapid evolution gene family in *J. mandshurica* genome. The inner number represents the gene family members in this related species

J. mandshurica (36/23; Figure 4d; Table S19). As reported, the *WAK* gene family was also contracted in *J. hindsii* (Trouern-Trend et al., 2020) but expanded in *Q. robur* (Figure 4d; Table S19). *LEA* expanded in *J. regia* but contracted in *J. mandshurica* (40/19), for *PPR* the difference was 1.2-fold (27/21), and for *PR* genes it was 2.5-fold (28/11; Figure 4d; Table S19).

3.4 | *JmaPR10* may involve in *J. mandshurica* lesion nematodes resistance

To detect the genome basis of lesion nematodes resistance in *J. mandshurica* we focused on *PR* gene subfamily 10 members; this subfamily was reported to be involved in response to lesion nematodes (Ozyigit et al., 2017; Soh et al., 2012; Zhao et al., 2015). We identified 11 *PR* genes in *J. mandshurica*, 28 in *J. regia*, and 15 in *A. thaliana* (Table S19). The phylogenetic tree showed that all *PR* genes were divided into two groups (Figure 5a). The syntenic analysis showed that a total of 20 orthologous gene pairs between *J. regia* and *J. mandshurica* (Figure 5b). Of these, both *JmaPR10* and *JmaPR8* contained three orthologous gene pairs, and *JmaPR2* contained two orthologous gene pairs compared with *J. regia PR* genes (Figure 5b). The 20 *PR* homologues found in *J. regia* and *J. mandshurica* were derived from the WGD event, and 19 were amplified via tandem duplication (Figure 5b; Table S20). The *PR* proteins exhibited high

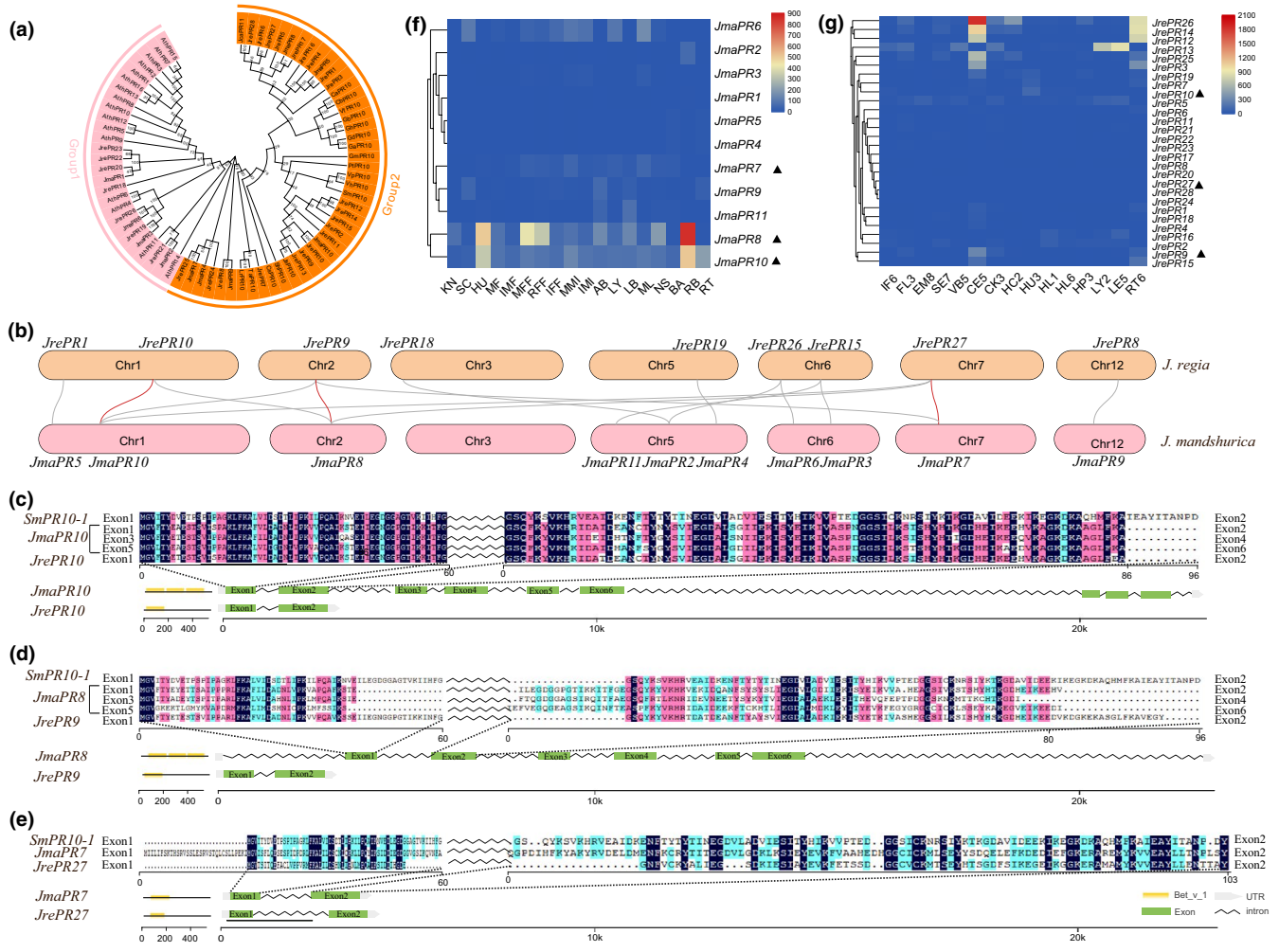

FIGURE 5 Characterization and evolution of *PR* gene family. (a) The phylogenetic tree of *PR10* subfamily gene from NCBI, and *PR* gene family of *J. mandshurica*, *J. regia*, and *A. thaliana*. (b) Synteny analysis of *PR* genes between *J. regia* and *J. cathayensis*. Grey lines in the background indicate the collinear blocks within *J. regia* and *J. mandshurica* genomes, while the red lines highlight the six genes (*JrePR10*, *JrePR9*, *JrePR27*, *JmaPR10*, *JmaPR8*, and *JmaPR7*). (c–e) The domains, gene structure and protein sequences of *PR* genes. Grey lines indicate nonconserved domains and orange box represent bet_v_1 conserved domain is displayed proportionally in each protein. Green boxes indicate exon, and the broken line represents the intron. Grey box represents UTR. (f–g) Expression profiles of the *J. mandshurica* and *J. regia PR* gene family among different tissues, including 18 tissues in *J. mandshurica* and 16 tissues in *J. regia* (abbreviations for tissues are described in Table S2; Martínez-García et al., 2016). The species names with the prefixes "*Jma*", "*Jre*," and "*Sm*" indicate *J. mandshurica*, *J. regia*, and *Salvia miltiorrhiza*, respectively

conservation based on multiple sequence alignments (Figure S9). The domain structural analyses of *PR* genes showed that most *PR* genes possess one PR domain (bet_v_1); however, *JmaPR10* and *JmaPR8* possesses three PR domains and *JmaPR2* possesses two PR domains, possibly derived from domain duplication (Figures 5c–e and S10A–C; Sun et al., 2019). The gene structure analyses showed that whereas most *PR* genes contain two exons, *JmaPR10* contains nine exons, *JmaPR8* contains six exons, and *JmaPR2* contains four exons (Figures 5c–e and S10A-C). There were highly similar protein sequences between exon pairs in *JmaPR10* versus *JrePR10*; and the similarity extended to exon 1 and exon 2 of *SmPR10*. Exon 1 and exon 2 of *PR10*, which *J. mandshurica* shares with *J. regia*, appear to be triplicated in *J. mandshurica* (*JmaPR10* [exon 1 and exon 2 gave rise to exon 3 and exon 4, and exon 5 and exon 6]; Figure S10D). Analysis of the transcriptomes showed that *JmaPR10*, *JrePR10*, *JmaPR8*, *JrePR9*, *JrePR14*, *JrePR15*, and *JrePR2* were more expressed in roots compared with other tissues and organs, and *JmaPR10* and *JmaPR8* also showed higher expression in root bark compared the other *PR* genes in *J. mandshurica* (Figure 5f–g; Table S20). Taken together, these results show that *JmaPR10* will be a good candidate gene for analysis of lesion nematode resistance in *J. mandshurica* (Figures 5 and S9–S10; Table S20; Chen et al., 2015; Ji et al., 2020; Ozyigit et al., 2017; Trouern-Trend et al., 2020).

4 | DISCUSSION

We report the first assembly of a high-quality, chromosome-level genome for *J. mandshurica* 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 248-fold (scaffold N50 size of *J. mandshurica* of this study was 36,084,664 bp vs. 145,095 bp scaffold N50 size for *J. mandshurica* (Stevens et al., 2018)), and the final calculated genome size (548 Mb) is smaller (580 Mb; Stevens et al., 2018; Figures 1–2; Table S21). Through Hi-C, a chromosome-level genome was obtained with a scaffold size of 36 Mb (Table S21) and scaffolds resolved into 16 chromosomes, unlike the previously available genome (*J. mandshurica*; Stevens et al., 2018; Tables S5 and S21; Chen et al., 2020; Choi et al., 2020; DeMaere & Darling, 2019; Zhang, Ren, et al., 2020). We predicted 29,032 protein-coding genes from the generated assembly (Figure 1; Table S6).

This study improves our ability to compare the genome of *J. mandshurica* with that of *J. regia* (Stevens et al., 2018; Zhang, Zhang, et al., 2020) by improving the accuracy of descriptions of genome characteristics, genome synteny, WGD, and deletion and insertion events (Figures 2–3; Table S11–S14). Our assembled genome permitted identification of the locations of deletions and insertions (Figure 3). These Indels were enriched in genes associated with the biosynthesis of amino acids, and phenylpropanoid biosynthesis; they may affect the regulation of these important metabolic pathways in *J. mandshurica* and *J. regia* (Figures 3 and S6).

J. mandshurica is recommend as a rootstock for *J. regia* to confer disease tolerance/resistance (Chen et al., 2015; Hu et al., 2016; Ji et al., 2020; Trouern-Trend et al., 2020; Zhou et al., 2017). The high-quality genome sequence we report here will improve our ability to identify signatures of genome evolution and the genetic basis of important traits. The *J. mandshurica* unique paralogues were enriched in three-disease tolerance/resistance pathways, including flavonoid biosynthesis, phenylpropanoid biosynthesis, plant–pathogen interaction (Figure 4). We also observed notable contraction in the size of gene families of resistance genes, including *WAK* (Trouern-Trend et al., 2020), *LEA* (Gao et al., 2020), *PR* (Ozyigit et al., 2017; Soh et al., 2012; Zhao et al., 2015), and *PPR* (Liu et al., 2016). These three pathways and four notable contractions may provide insight into the resistance phenotypes of *J. mandshurica* that make it a valuable rootstock (Figure 4). Furthermore, in the current study, we described the structure of *JmaPR10*, a member of the *PR* gene family which may be important for the reported resistance of *J. mandshurica* resistance to lesion nematodes, it is consistent with the previous studies (Chen et al., 2015; Ji et al., 2020; Ozyigit et al., 2017; Trouern-Trend et al., 2020). Therefore, our results constitute an important basis for improving the understanding of the genome basis of the resistance traits in *J. mandshurica* (Figure 5).

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (32070372, 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. Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

AUTHOR CONTRIBUTIONS

Peng Zhao conceived and designed the study. Feng Yan, and Peng Zhao collected the samples. Peng Zhao took the morphology picture of Manchurian walnut. Feng Yan, Peng-Peng Chen, Rui-Min Xi, Rui-Xue She, Yu-Jie Yan, Ge Yang and Meng Dang performed the experiments. Feng Yan, Rui-Xue She, and Peng Zhao analysed and interpreted the assembly and annotations. Feng Yan, Peng-Peng Chen, Meng Dang, Ge Yang, Dong Pei and Ming Yue supported the software. Feng Yan and Peng Zhao performed the comparative genome analysis. Feng Yan, and Peng Zhao performed the whole genome duplication analysis. Feng Yan, and Peng Zhao wrote the draft manuscript and then Peng Zhao 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, transcriptome data, and genome file have been deposited in the NCBI, under accession numbers: PRJNA674421.

ORCID

Peng Zhao https://orcid.org/0000-0003-3033-6982

REFERENCES

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.

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.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Bai, W. N., Liao, W. J., & Zhang, D. Y. (2010). Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytologist*, *188*(3), 892–901.

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.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic 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.

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), 869–880.

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.

Chen, C., Chen, H., Zhang, Y. I., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant*, *13*(8), 1194–1202.

Chen, G., Pi, X. M., & Yu, C. Y. (2015). A new naphthalenone isolated from the green walnut husks of *Juglans mandshurica* maxim. *Natural Product Research*, *29*(2), 174–179.

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

Dang, M., Liu, Z. L., Chen, X., Zhang, T., Zhou, H. J., Hu, Y. H., & 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., Yue, M., Zhang, M., Zhao, G., & Zhao, P. (2019). Gene introgression among closely related species in sympatric populations: A case study of three walnut (*Juglans*) species. *Forests*, *10*(11), 965.

Dang, M., Zhang, T., Hu, Y., Zhou, H., Woeste, K. E., & 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.

DeMaere, M. Z., & Darling, A. E. (2019). Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology*, *20*(1), 46.

Dong, X., Wang, Z., Tian, L., Zhang, Y., Qi, D., Huo, H., Xu, J., Li, Z., Liao, R., Shi, M., Wahocho, S. A., Liu, C., Zhang, S., Tian, Z., & Cao, Y. (2019). De novo assembly of a wild pear (*Pyrus betuleafolia*) genome. *Plant Biotechnology Journal*, *18*(2), 581–595.

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

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.

Edgar, R. C., & Myers, E. W. (2005). PILER: Identification and classification of genomic repeats. *Bioinformatics*, *21*(1), 152–158.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., & Smart, A. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432.

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

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., & Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, *36*, D281–D288.

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), 1–19.

Gao, T., Mo, Y., Huang, H., Yu, J., Wang, Y., & Wang, W. (2020). Heterologous expression of *Camellia sinensis* late embryogenesis abundant protein gene 1 (*CsLEA1*) confers cold stress tolerance in *Escherichia coli* and yeast. *Horticultural Plant Journal*, *7*(1), 89–96.

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.

Han, H., Woeste, K. E., Hu, Y., Dang, M., Zhang, T., Gao, X.-X., Zhou, H., Feng, X., Zhao, G., & 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, B., Jin, J., Guo, A. Y., Zhang, H., Luo, J., & Gao, G. (2015). GSDS2.0: An upgraded gene feature visualization server. *Bioinformatics*, *31*(8), 1296–1297.

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*, *6*(7), 1955.

Hu, Z., Zhang, T., Gao, X.-X., Wang, Y., Zhang, Q., Zhou, H.-J., Zhao, G.-F., Wang, M.-L., Woeste, K. E., & Zhao, P. (2016). De novo assembly and characterization of the leaf, bud, and fruit transcriptome from the vulnerable tree *Juglans mandshurica* for the development of 20 new microsatellite markers using Illumina sequencing. *Molecular Genetics & Genomics*, *291*(2), 849–862.

Ji, L. I., Zhang, Y., Yang, Y., Yang, L., Yang, N. A., & Zhang, D. (2020). Long-term effects of mixed planting on arbuscular mycorrhizal fungal communities in the roots and soils of *Juglans mandshurica* plantations. *BMC Microbiology*, *20*(1), 304.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & de Bakker, P. I. W. (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., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Rogozin, I. B., Smirnov, S., Sorokin, A. V., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & 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.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, *9*(4), 299–306.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323.

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

Liu, B., Zhao, D., Zhang, P., Liu, F., Jia, M., & Liang, J. (2020). Seedling evaluation of six walnut rootstock species originated in China based on principal component analysis and cluster analysis. *Scientia Horticulturae*, *265*, 109–212.

Liu, J. M., Zhao, J. Y., Lu, P. P., Chen, M., Guo, C. G., Xu, Z. S., & Ma, Y. Z. (2016). The E-Subgroup pentatricopeptide repeat protein family in *Arabidopsis thaliana* and confirmation of the responsiveness *PPR96* to abiotic stresses. *Frontiers in Plant Science*, *5*(7), 1825.

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

Lu, A. M. (1982). On the geographical dispersal of Juglandaceae. *Acta Phytotaxonomica Sinica*, *20*, 257–274.

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.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., … Bryant, S. H. (2016). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203.

Martínez-García, 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*(5), 507–532.

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics*, *32*(19), 3021–3023.

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

Ozyigit, I. I., Vatansever, R., & Filiz, E. (2017). Comparative analyses of pathogenesis-related protein-10 (*PR10*) in plants. *Indian Journal of Biotechnology*, *16*, 325–333.

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., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*(1), 259.

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

Simão, 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.

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

Soh, H. C., Park, A. R., Park, S., Back, K., Yoon, J. B., Park, H. G., & Kim, Y. S. (2012). Comparative analysis of pathogenesis-related protein 10 (*PR10*) genes between fungal resistant and susceptible peppers. *European Journal of Plant Pathology*, *132*(1), 37–48.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

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., Martínez-García, 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). Three new naphthalenyl glycosides from the root bark of *Juglans cathayensis*. *Chemical & Pharmaceutical Bulletin*, *60*(6), 785–789.

Sun, Y., Wu, Z., Wang, Y., & Yang, J. (2019). Identification of phytocyanin gene family in legume plants and their involvement in nodulation of *Medicago truncatula*. *Plant and Cell Physiology*, *60*(4), 900–915.

Suryamohan, K., Krishnankutty, S. P., Guillory, J., Jevit, M., Schröder, M. S., Wu, M., Kuriakose, B., Mathew, O. K., Perumal, R. C., Koludarov, I., Goldstein, L. D., Senger, K., Dixon, M. D., Velayutham, D., Vargas, D., Chaudhuri, S., Muraleedharan, M., Goel, R., Chen, Y.-J., … Seshagiri, S. (2020). The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nature Genetics*, *52*(1), 106–117.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., Baren, M., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

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*, *102*(2), 410–423.

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., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & 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), 1859–1875.

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

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287.

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.

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

Zhang, J., Zhang, W., Ji, F., Qiu, J., Song, X., Bu, D., Pan, G., Ma, Q., Chen, J., Huang, R., Chang, Y., & Pei, D. (2020). A high-quality walnut genome assembly reveals extensive gene expression divergences after whole-genome duplication. *Plant Biotechnology Journal*, *18*(9), 1848–1850.

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

Zhang, Z., Chen, Y., Zhang, J., Ma, X., Li, Y., Li, M., Wang, D., Kang, M., Wu, H., Yang, Y., Olson, M. S., DiFazio, S. P., Wan, D., Liu, J., & Ma, T. (2020). Improved genome assembly provides new insights into genome evolution in a desert poplar (*Populus euphratica*). *Molecular Ecology Resources*, *20*(3), 1–14.

Zhao, L., Ma, L. G., Wang, Z. S., Chen, M. L., Shen, Y., & Huang, L. Q. (2015). Molecular cloning and characterization of a pathogenesis-related protein *SmPR10-1* from *Salvia miltiorrhiza*. *Acta Physiologiae Plantarum*, *37*(10), 205.

Zhao, P., Zhao, G. F., Zhang, S. X., Zhou, H. J., Hu, Y. H., & Woeste, K. E. (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, P., Zhou, H.-J., Potter, D., Hu, Y.-H., Feng, X.-J., Dang, M., Feng, L. I., Zulfiqar, S., Liu, W.-Z., Zhao, G.-F., & 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, X., & Hao, W. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*, W265–W268.

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.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Yan F, Xi R-M, She R-X, et al. Improved de novo chromosome-level genome assembly of the vulnerable walnut tree *Juglans mandshurica* reveals gene family evolution and possible genome basis of resistance to lesion nematode. *Mol Ecol Resour.* 2021;21:2063–2076. https://doi.org/10.1111/1755-0998.13394