Completion of the Chloroplast Genomes of Five Chinese Juglans and Their Contribution to Chloroplast Phylogeny

Yiheng Hu 1, Keith E. Woeste 2 and 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 United States Department of Agriculture Forest Service Hardwood Tree Improvement and Regeneration Center, Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA

Juglans L. (walnuts and butternuts) is an economically and ecologically important genus in the family Juglandaceae. All Juglans are important nut and timber trees. Juglans regia (Common walnut), J. sigillata (Iron walnut), J. cathayensis (Chinese walnut), J. hopeiensis (Ma walnut), and J. mandshurica (Manchurian walnut) are native to or naturalized in China. A strongly supported phylogeny of these five species is not available due to a lack of informative molecular markers. We compared complete chloroplast genomes and determined the phylogenetic relationships among the five Chinese Juglans using Illumina sequencing. The plastid genomes ranged from 159,714 to 160,367 bp encoding 128 functional genes, including 88 protein-coding genes and 40 tRNA genes each. A complete map of the variability across the genomes of the five Juglans species was produced that included single nucleotide variants, indels (insertions and deletions), and large structural variants, as well as differences in simple sequence repeats (SSR) and repeat sequences. Molecular phylogeny strongly supported division of the five walnut species into two previously recognized sections (Juglans/Dioscaryon and Cardiocaryon) with a 100% bootstrap (BS) value using the complete cp genomes, protein coding sequences (CDS), and the introns and spacers (IGS) data. The availability of these genomes will provide genetic information for identifying species and hybrids, taxonomy, phylogeny, and evolution in Juglans, and also provide insight into utilization of Juglans plants.

Keywords: persian walnut, ma walnut, iron walnut, chinese walnut, manchurian walnut, phylogeny, China, butternut

INTRODUCTION

The estimate of phylogenetic relationships plays a key role in understanding evolution and has been an essential component of evolutionary biology. In plants, much effort in reconstructing the Tree of Life has focused on the relationships of major clades, and significant advances have been made above the order or family levels (The Angiosperm Phylogeny Group III, 2009; Soltis et al., 2011). Until recently, progress in inferring phylogenetic relationships at lower taxonomic levels and among recently diverged species has been less encouraging, especially for species-rich,
morphologically diverse lineages (Waterway et al., 2009). In the past few years, however, important advances have been made in multispecies coalescent approaches for resolving genome-level relationships among closely related species using next generation sequencing to resolve incomplete lineage sorting and interlineage hybridization (Huang et al., 2014; Carbonell-Caballero et al., 2015; Danielli et al., 2016).

Walnuts and butternuts (Juglans) are known for their edible nuts and high-quality wood (Manning, 1978; Aradhya et al., 2007). The genus Juglans includes about 21 species distributed in Asia, southern Europe, North America, Central America, western South America, and the West Indies (Manning, 1978; Stanford et al., 2000; Aradhya et al., 2007). Species of Juglans are diploid, with a karyotype of 2n = 2x = 32 (Woodworth, 1930; Komanich, 1982). *J. regia* (common walnut), *J. sigillata* (iron walnut), *J. cathayensis* (Chinese walnut), *J. hopeiensis* (Ma walnut), and *J. mandsurica* (Manchurian walnut) grow in China (Manning, 1978; Fjellstrom and Parfitt, 1995; Aradhya et al., 2007). *Juglans* is taxonomically and phylogenetically challenging. Classical taxonomy divides the genus into four sections (sect. Dioscaryon, sect. Cardiocaryon, sect. Trachycaryon, and sect. Rhysoacaryon) mainly based on species’ geographical distribution, leaf, flower, and fruit morphology (Dode, 1909; Manning, 1978). Molecular evidence, however, including sequence data from the internal transcribed spacer (ITS), five chloroplast DNA spacer sequences (*atpB-rbcL*, *psbA-trnH*, *trnS-trnFM*, *trnT-trnF*, and *trnV-16S rRNA*), a hyper-variable *matK*, and restriction fragment length polymorphisms (RFLPs), has been interpreted as supporting three or four sections (Fjellstrom and Parfitt, 1995; Stanford et al., 2000; Aradhya et al., 2007). *Juglans* is taxonomically and phylogenetically challenging. Classical taxonomy divides the genus into four sections (sect. Dioscaryon, sect. Cardiocaryon, sect. Trachycaryon, and sect. Rhysoacaryon) mainly based on species’ geographical distribution, leaf, flower, and fruit morphology (Dode, 1909; Manning, 1978). Molecular evidence, however, including sequence data from the internal transcribed spacer (ITS), five chloroplast DNA spacer sequences (*atpB-rbcL*, *psbA-trnH*, *trnS-trnFM*, *trnT-trnF*, and *trnV-16S rRNA*), a hyper-variable *matK*, and restriction fragment length polymorphisms (RFLPs), has been interpreted as supporting three or four sections (Fjellstrom and Parfitt, 1995; Stanford et al., 2000; Aradhya et al., 2007).

Chinese Juglans species are divided into two sections (sect. Dioscaryon and sect. Cardiocaryon). Common walnut (*J. regia*) and Iron walnut (*J. sigillata*) belong to sect. Dioscaryon, and the other three species (*J. cathayensis*, *J. hopeiensis*, and *J. mandsurica*) belong to sect. Cardiocaryon (Dode, 1909; Fjellstrom and Parfitt, 1995; Stanford et al., 2000; Aradhya et al., 2007). Common walnut (*J. regia*) is native to the mountainous regions of central Asia (Pollegioni et al., 2015), while Iron walnut (*J. sigillata*) is indigenous to China, and distributed mainly in southwestern China (Wang et al., 2015). Chinese walnut (*J. cathayensis*) is widely distributed in southern China (Bai et al., 2014; Dang et al., 2015), while *J. mandsurica* is mainly distributed in northern China, northeast China, and the Korean Peninsula (Wang et al., 2016). *J. hopeiensis* is narrowly distributed in northern China in the hilly, mid-elevation area between Hebei province, Beijing, and Tianjin (Hu et al., 2015). A strongly supported phylogeny of these five species is not available due to a lack of informative molecular markers (Fjellstrom and Parfitt, 1995; Stanford et al., 2000; Aradhya et al., 2007). Studies of gene flow and introgression have concluded *J. regia* and *J. sigillata* are particularly closely related, and some have questioned whether they are distinct (Wang et al., 2008, 2015). Aradhya et al. (2007) used ITS, RFLP, and cpDNA sequence data to suggest *J. regia* and *J. sigillata* are distinct species. *J. cathayensis* and *J. mandsurica* were combined into one species in Flora of China (English version) (Lu et al., 1999), which does not consider *J. hopeiensis* (Kuang and Lu, 1979; Aradhya et al., 2004, 2007) a valid taxon. In addition, some previous phylogenetic studies of Juglans omitted *J. hopeiensis* and *J. sigillata* (Fjellstrom and Parfitt, 1995; Stanford et al., 2000; Aradhya et al., 2007). Thus, the phylogeny and systematics of the five Chinese walnut (*Juglans*) species is uncertain.

In this study, we combined de novo and reference-guided assembly of five Chinese walnut (*Juglans*) species’ whole chloroplast genomes (Cpgs). This is the first comprehensive Cpg analysis of multiple *Juglans* species. Our aims were: (1) to investigate global structural patterns of whole chloroplast genome of five *Juglans* species including genome structure, gene order, and gene content; (2) to examine variations of simple sequence repeats (SSRs) and large repeat sequence in the whole Cpgs of *Juglans*; (3) to identify divergence hotspots as regions potentially under selection pressure; and (4) to construct a chloroplast phylogeny for the five Chinese *Juglans* species using their whole cp DNA sequences, protein coding sequences, and the introns and spacers.

**MATERIALS AND METHODS**

**Taxon Sampling, Plant Material, and Deposition of Voucher**

Fresh leaves of four *Juglans* species were collected from different mountains in China, including a *J. mandsurica* tree growing in the Xialongmen National Forest Park, a *J. sigillata* tree from Lijiang, Yunan, a *J. hopeiensis* tree growing Laishui, Beijing, and a *J. cathayensis* tree growing in the Qingling Mountains (Table 1). The leaves were dried in silica gel and stored at −4 °C. The leaves of *J. regia* were collected fresh from a tree growing the orchard of Northwest University, Shannxi, China. Voucher specimens of each of the sampled trees were deposited at the herbarium of Northwest University, Xi’an, China. All the DNA samples were stored at Evolutionary Botany Lab, Northwest University, Xi’an, China. High-quality genomic DNA was extracted using a modified CTAB method (Zhao and Woeste, 2011). The DNA concentration was quantified using a NanoDrop spectrophotometer (Thermo Scientific, Carlsbad, CA, USA). The final DNA concentration >30 ng µL⁻¹ were chosen for further Illumina sequencing. We sequenced the complete chloroplast genome of *J. regia* with the Illumina MiSeq sequencing platform (Sangon Biotech, Shanghai, China). We assembled the chloroplast genomes using SPAdes v3.6.2 (Bankevich et al., 2012) (http://bioinf.spbau.ru/spades) and annotated them with CpgAVAS (http://www.biomedcentral.com/1471-2164/13/715) (Liu et al., 2012a; Hu et al., 2016). We sequenced the complete Cpg of four *Juglans* species using Illumina HiSeq 2500 sequencing technology via a combination of de novo and reference-guided assembly based on the Cpg of *J. regia* (Hu et al., 2016, NCBI Accession number: KT963008). A paired-end (PE) library with 350-bp insert size was constructed using the Illumina PE DNA library kit according to the manufacturer’s instructions and sequenced using an Illumina Hiseq2500 by Novogene (http://www.novogene.com, China).
### TABLE 1 | Summary statistics for assembly of five *Juglans* species chloroplast genomes.

<table>
<thead>
<tr>
<th>Genome features</th>
<th><em>Juglans regia</em></th>
<th><em>Juglans sigillata</em></th>
<th><em>Juglans hopeiensis</em></th>
<th><em>Juglans cathayensis</em></th>
<th><em>Juglans mandshurica</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>160367</td>
<td>160350</td>
<td>159714</td>
<td>159730</td>
<td>159729</td>
</tr>
<tr>
<td>LSC length (bp)</td>
<td>89872</td>
<td>89872</td>
<td>89316</td>
<td>89333</td>
<td>89331</td>
</tr>
<tr>
<td>SSC length (bp)</td>
<td>18423</td>
<td>18406</td>
<td>18352</td>
<td>18351</td>
<td>18351</td>
</tr>
<tr>
<td>IR length (bp)</td>
<td>26036</td>
<td>26036</td>
<td>26023</td>
<td>26023</td>
<td>26023</td>
</tr>
<tr>
<td>Coding (bp)</td>
<td>80475</td>
<td>80475</td>
<td>80202</td>
<td>80110</td>
<td>80344</td>
</tr>
<tr>
<td>Noncoding (bp)</td>
<td>79892</td>
<td>79875</td>
<td>79512</td>
<td>79620</td>
<td>79385</td>
</tr>
<tr>
<td>Number of genes</td>
<td>129</td>
<td>129</td>
<td>129</td>
<td>129</td>
<td>129</td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>rRNA genes</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Number of genes duplicated in IR</td>
<td>19 (4/7/7/1)</td>
<td>19 (4/7/7/1)</td>
<td>19 (4/7/7/1)</td>
<td>19 (4/7/7/1)</td>
<td>19 (4/7/7/1)</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>36.1</td>
<td>36.1</td>
<td>36.1</td>
<td>36.1</td>
<td>36.1</td>
</tr>
<tr>
<td>GC content in LSC (%)</td>
<td>33.6</td>
<td>33.6</td>
<td>33.6</td>
<td>33.7</td>
<td>33.7</td>
</tr>
<tr>
<td>GC content in SSC (%)</td>
<td>29.8</td>
<td>29.8</td>
<td>29.8</td>
<td>29.8</td>
<td>29.8</td>
</tr>
<tr>
<td>GC content in IR (%)</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.5</td>
<td>42.5</td>
</tr>
<tr>
<td>Sequencing Platform</td>
<td>Illumina Miseq</td>
<td>Illumina HiSeq</td>
<td>Illumina HiSeq</td>
<td>Illumina HiSeq</td>
<td>Illumina HiSeq</td>
</tr>
<tr>
<td>Raw reads</td>
<td>6321912</td>
<td>12382845</td>
<td>10285876</td>
<td>13320133</td>
<td>11903351</td>
</tr>
<tr>
<td>Raw Base (G)</td>
<td>1.9</td>
<td>3.1</td>
<td>2.57</td>
<td>3.33</td>
<td>2.98</td>
</tr>
<tr>
<td>Average read length (bp)</td>
<td>300</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Average insert size (bp)</td>
<td>350</td>
<td>350</td>
<td>350</td>
<td>350</td>
<td>350</td>
</tr>
<tr>
<td>Number of assembled reads</td>
<td>1846010</td>
<td>804634</td>
<td>1118104</td>
<td>669686</td>
<td>1055940</td>
</tr>
<tr>
<td>Source</td>
<td>Xi’an, Qinling</td>
<td>Lijiang, Yunnan</td>
<td>Laishui, Beijing</td>
<td>Lantian, Qinling</td>
<td>Xiaolongmen, Beijing</td>
</tr>
</tbody>
</table>

---

### Chloroplast Genome Sequencing, Assembly, and Gap Filling

Raw reads with sequences shorter than 50 bp or with more than the allowed maximum percentage of ambiguous bases (2%) were removed from the total NGS PE reads using the NGSCQ toolkit v2.3.3 (Patel and Jain, 2012) trim tool. After trimming, high-quality PE reads were assembled using MIRA v4.0.2 (Chevreux et al., 2004) assembler. Then, to further assemble the Cpg, some ambiguous regions were picked out for extension with a baiting and iteration method based on MITObim v1.8 (Hahn et al., 2013). A de novo assembly strategy combined with a reference-based assembly allowed us to reconstruct each Cpg. Reads were then remapped to references for each taxon to check for misassemblies or rearrangements using Geneious v8.0.2 (http://www.Geneious.com; Kearse et al., 2012) and reads matching the draft reference were assembled de novo, also in Geneious, using suggested settings. Inverted repeat boundaries were determined and verified by remapping reads in Geneious. Lastly, primers were developed with Primer3 (Untergrasser et al., 2012) to close low coverage gaps between contigs (for a few single end datasets). Small gaps in the assemblies were bridged by designing custom primers for PCR (Table S1) based on their flanking sequences, followed by conventional Sanger sequencing. The PCR primers were designed using *J. regia* sequences when they appeared identical to our original de novo assembly (Hu et al., 2016). Eleven primer pairs were used to validate junctions using PCR based sequencing in each of five *Juglans* Cpgs. PCR amplification was carried out on a SimpliAmp Thermal Cycler (Applied Biosystem, USA) in 20 µL reaction volumes (10 µL 2 × PCR Master Mix including 0.1 U Taq polymerase/µL; 500 µM each dNTP; 20 mM Tris- HCl (pH 8.3); 100 mM KCl; 3.0 mM Mgc2 (Tiangen, Beijing, China), 0.5 µL each primer, 2 µL BSA, 2 µL of 10 ng/µL DNA). The PCR was programmed for 3 min at 94°C followed by 35 cycles of 15 s at 93°C, 1 min at annealing temperature (60°C), 30 s at 72°C and 10 min at 72°C. After PCR amplification, fragments were sequenced by Sangon Biotech (Shanghai, China). All newly generated sequences were deposited in GenBank (Table S1).

### Genome Annotation and Analysis

The completed genome sequences were imported into the online program Dual Organellar Genome Annotator (DOGMA, Wyman et al., 2004) for annotation, coupled with manual investigation of the positions of start and stop codons and boundaries between introns and exons. Putative starts, stops, and intron positions were determined by comparison with homologous genes in other chloroplast genomes using MAFFT v7.0.0 (Katoh and Standley, 2013). All tRNA genes were further verified online using tRNAscan-SE search server (Lowe and Eddy, 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/). The circular *Juglans regia* chloroplast genome map was drawn using Organellar Genome DRAW (Lohse et al., 2013). Genome annotation was performed in Geneious, and the GC-content of protein-coding genes, tRNA genes, introns and intergenic
spacers (IGSs) was determined on the basis of their annotation. Cpg comparison among the five Juglans species was performed with VISTA (Frazer et al., 2004). Genome, protein coding gene, intron, and spacer sequence divergences were evaluated using DnaSP v5.10 (Librado and Rozas, 2009) after alignment. For the protein coding gene sequences, introns, and spacers, every gene or fragment was annotated using the software Geneious v8.0.2 (http://www.Geneious.com; Kearse et al., 2012). For purposes of the subsequent phylogenetic analysis and plant identification, the complete Cpg of each Juglans species was compared and diagrammed using VISTA to show sequence divergence.

### Repeat Sequencing Analysis
The genomic sequences were analyzed to identify potential microsatellites (simple sequence repeats or SSRs, i.e., mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) using MISA software (http://pgrc.ipk-gatersleben.de/misa/) with thresholds of ten repeat units for mononucleotide SSRs and five repeat units for di-, tri-, tetra-, penta-, and hexanucleotide SSRs. The web-based software REPuter (Kurtz et al., 2001) (http://bibiserv.techfak.uni-bielefeld.de/reputer/) was used to analyze the repeat sequences, which included forward, reverse, complement, palindromic and tandem repeats with minimal lengths of 30 bp and edit distances of less than 3 bp. The large repeat sequences were analyzed by using the Web-based Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html). We investigated if the repeated elements identified in the chloroplast of J. regia were also present in other four other Chinese Juglans species by aligning their cp genomes using Geneious v8.0.2 (http://www.Geneious.com; Kearse et al., 2012). Tandem repeat sequences (>10 bp in length) were detected using the online program Tandem Repeats Finder (Benson, 1999), with 2, 7, and 7 set for the alignment parameters match, mismatch, and indel, respectively. The minimum alignments core and maximum period size were 80 and 500, respectively.

### Mutation Events Analysis, Substitution Rate Analyses, and Inference of Rate Changes
To identify the microstructural mutations of Juglans, the five aligned sequences were further analyzed using DnaSP v5 (Librado and Rozas, 2009) and MEGA v5.0 (Tamura et al., 2011). Indel and SNP events were counted and positioned in the cp genome using DnaSP v5. Signatures of natural selection were studied for every chloroplast genome located outside of the inverted repeats region. Selective pressures (\( K_d/K_s \)) were computed with the codeml tool from PAML package v4.0 (Yang, 2007) using a YN00 model to test every gene sequence. We used the KaKs calculator program to check the selective pressures (\( K_d/K_s \)) using same model as YN (Zhang et al., 2006). To avoid potential convergence biases, those genes with few mutations were filtered out from selective pressure analysis.

### Phylogenetic Analysis
The Juglans Cpg sequences from the finalized data set were aligned with MAFFT v7.0.0 (Katoh and Standley, 2013). The analyses were carried out based on the following three data sets: (1) the complete cp DNA sequences; (2) protein coding sequences; (3) the introns and spacers. We conducted ML analyses using each of the data sets separately. The phylogenetic analyses were carried out using the Cpgs of all five Juglans species plus eight other species with complete Cpgs (Table S2). The Maximum Likelihood (ML) phylogenetic tree analysis was conducted using RAXML v8.0 (Stamatakis, 2014) under GTR+Gamma model. For ML analysis, difference general time reversible models were performed with all three data sets. For all analyses, 10 independent ML searches were conducted, bootstrap support was estimated with 1000 bootstrap replicates, and bootstrap proportions were drawn on the tree with highest likelihood score from the 10 independent searches. The choice of substitution model for each partition was primarily determined by using Modeltest v3.7 (Posada and Crandall, 1998) with the Akaïke information criterion (AIC) (Posada and Buckley, 2004). Maximum Parsimony (MP) phylogenetic analyses were performed in MEGA v5.0 (Tamura et al., 2011) using 1000 bootstrap replicates. BL trees were produced by MrBayes v3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Altekar et al., 2004) with the setting of 1,000,000 generations and stopval = 0.1, under GTR+Gamma model with one cold and three incrementally heated Markov Chain Monte Carlo (MCMC) run simultaneously (Ronquist and Huelsenbeck, 2003) in two parallel runs sampling every 1000 generations. The first 25% of the trees were discarded as burn-in. The remaining trees were used for generating the consensus tree. The phylogenetic relationships and divergence time between lineages were estimated using Bayesian inference method BEAST v1.8.0 (Drummond et al., 2012). Calibration of the Juglandaceae and Fagaceae split (73.4 ± 0.1 Myr) was based on references in Thomas et al. (2012) and Hedges et al. (2015). The GTR+Gamma nucleotide substitution model was selected using software MODELTEST v3.7 (Posada and Crandall, 1998). A relaxed clock with lognormal distribution of uncorrelated rate variation was specified. A normal prior probability distribution was used to accommodate the uncertainly of prior knowledge. Two independent Markov chains of 10,000,000 generations, sampled every 10,000 th iteration, were generated. An adequate effective sample size (larger than 200) and convergence of the Markov chain Monte Carlo chains were diagnosed in Tracer v1.6 with the first 10% samples discarded as burn-in (Drummond et al., 2012). The phylogenetic trees were then compiled into a maximum clade credibility tree using TreeAnnotator v1.8.0 (Drummond et al., 2012) and the program FigTree v1.3.1 (Drummond et al., 2012) to visualize mean node ages and highest posterior density (HPD) intervals at 95% (upper and lower) for each node and to estimate branch lengths and divergence times.

### RESULTS
**Genome Assembly and PCR-Based Gap Filling**
Using the Illumina HiSeq system, five Juglans species were sequenced to produce a total of 10,285,876 to 13,320,133 bp paired-end raw reads from four Juglans species, while Common
walnut (J. regia) had 6,321,912 bp raw reads (Table 1). After aligning the paired-end reads with the reference Cpg (common walnut, J. regia), 689,686 to 1,118,104 bp Cpg reads were assembled (Table 1). The four Chinese Juglans Cpgs were deposited in NCBI GenBank (accession numbers, KX671976, KX671977, KX671975, and KT963008).

**General Features of the Five Chinese Walnut (Juglans) Chloroplast Genomes**

The five Juglans Cpgs ranged from 159,714 bp (J. hopeiensis) to 160,367 bp (J. regia), the average Cpg sequence length was 159,978 bp (Figure 1, Table 1). The coding sequence of the five Juglans Cpg ranged from 80,110 bp (J. cathayensis) to 80,475 bp (J. regia and J. sigillata), while the LSC length and SSC length ranged from 89,316 bp (J. hopeiensis) to 89,872 bp (J. regia and J. sigillata) and 18,351 bp (J. cathayensis) to 18,423 bp (J. regia), respectively (Table 1). For all five Cpgs the average GC content was 36.1% (Table 1). There are four introns located in the IR region and 13 introns in the LSC region in each of the Cpgs. There was only one gene (ndhA) located in SSC region (Table 2). All five Cpgs included a large single-copy (LSC) region of 89,316 to 89,872 bp, a small single-copy (SSC) region of 18,351 to 18,406 bp, and the inverted repeats (IR) were 26,023 bp (Table S1, Table 1). All five walnut Cpgs encoded 128 functional genes, including 88 protein-coding genes, 40 tRNA genes, and 8 ribosomal RNA genes (Table 1). There were 18 intron-containing genes (one class I intron in trn-UAA and 17 class II introns), of which three genes rps12, cfpP, and ycf3, contained two introns and the rest had only one intron each (Table 2). In addition, there were two pseudogenes: infA and ycf15, in which several internal stop codons were identified. The ycf15 gene displayed exactly the same structure in all five Chinese Juglans Cpgs. The pseudogene infA contained internal stop codons which differed among the five Juglans Cpgs.

**Conservation within Juglans Cpgs and Comparison with Fagaceae and Betulaceae**

When duplicated genes in IR regions were counted only once, all five Juglans Cpgs harbored 128 functional genes (except eight rRNA and pseudogenes ycf15 and infA) arranged in the same order, including 88 protein-coding genes and 40 tRNAs (Table 2). Fourteen of the protein-coding genes and six of the tRNA genes contained introns, 19 of which contained a single intron, whereas four had two introns (Table 2). The numbers of protein-coding genes in the Cpgs of the five Chinese Juglans was similar to the number of protein-coding genes in the Betulaceae and Fagaceae, two closely related plant families. As described above, ycf15 was a pseudogene in all five Chinese Juglans; it is also non-functional in the Betulaceae, and Fagaceae except in Q. rubra. We identified seven internal stop codons in the ycf15 sequence of Chinese Juglans (Figure 2B). The infA gene was also present as a pseudogene in all five Chinese Juglans Cpgs because of several stop codons. By contrast, infA appears to be a protein-coding gene in Quercus, Castanopsis, and Trigonobalanus. In Castanea, the infA gene contains a long indel (70 bp) rather than an internal stop codon (Figure 2). In this study, we identified nine internal stop codons in the infA sequence of J. regia and J. sigillata (sect. Dioscaryon). By contrast, we found five, five, and two internal stop codons in the infA sequence of J. hopeiensis, J. mandshurica, and J. cathayensis, respectively (Figure 2A).

All five Juglans Cpg IR regions were well conserved, including gene number and gene order, but they exhibited obvious differences at the single-copy (SC) boundary regions (Figure S1). The nucleotide sequence length of SSC regions ranged from 18,351 to 18,423 bp (72 bp difference), while the nucleotide sequence length of the IR regions ranged from 26,023 to 26,036 bp (13 bp difference) (Table 1). The nucleotide sequence differences were mainly found between members of the two sections (sect. Dioscaryon, and sect. Cardiocaryon). Within the IR region, the gene ycf2 had two SNPs, and ycf7 had one SNP. There were two polymorphisms (12 bp indel and 6 bp indel) in the ycf2-trnV-GAC spacers region, and one SNP in the rRNA-trnL-GAU 16s interval, one SNP in the trnL-GAU, six in the rRNA 23S, and one in rRNA-trn-RAG. The trnR-AAG-trnN-GUU spacers region had three SNPs. The gene ycf1 had six SNPs and one indel of 7 bp (Table S3). The gene ycf1 crossed into the SSC region, and the pseudogene fragment ycf1 was located in the IRA region at 1158 to 1162 bp.

The coding regions of the Cpgs were more highly conserved than the non-coding regions, as expected (Figure 3), but there were differences among the five species. The most dissimilar coding regions were ndhA and rpoC2 (Figure 3). Other evolutionary differences among the five cp genomes were inferred from differences in genome size in general and, in particular, differences in the size of the single copy (SC) region (Figure S1).

**Microsatellite Polymorphisms and Repeat Sequences**

Each Juglans Cpg contained 66 to 83 SSRs at least 10 bp in length (Table 3, Figure 4A, Table S4). Among these SSRs (about 73 SSRs per Cpg), most were located in noncoding sections of the LSC/SSC region (96.3% of the total occurrences), and about 11 per Cpg were in protein-coding genes (ycf1, rpoC1, rpoC2, rpoB, and atpB) (Table 3, Table S4). J. hopeiensis and J. mandshurica included about 17 more SSR loci in their Cpgs than the other three species. Mono-, di-, trin-, tetra-, penta-, and complex nucleotide SSRs were detected in every species, the mononucleotide, complex nucleotide, and dinucleotide SSRs averaged 64.8, 10.4, and 5.6%, of all SSRs, respectively. SSRs in walnut Cpgs are especially rich in AT. Nearly all SSRs (84.0%) were mononucleotide A/T repeats; only one or two C/G mononucleotide SSRs per genome were present. Among dinucleotide SSRs, AT/TA repeats were the most common (typically about seven per Cpg), trinucleotide SSRs (ATT/ATA) repeats were present in a small number of loci (one or three, depending on species), and depending on species, from 8 to 11 loci contained complex nucleotide repeats (Table 3, Figure S2, Table S4). AAAAT/ATTTT SSRs and AAATAT/ATATTT SSRs were only found in J. regia and J. sigillata (sect. Dioscaryon), and AAGAT/ATCTTT repeat units were only found in J. cathayensis, J. hopeiensis and J. mandshurica (Table 3, Figure S2, Table S4).
FIGURE 1 | Chloroplast genome maps of three Juglans species. (A) J. cathayensis chloroplast genome, (B) J. mandshurica chloroplast genome, (C) J. sigillata chloroplast genome. Genes drawn outside the outer circle are transcribed clockwise, and those inside are transcribed counter-clockwise. Genes belonging to different functional groups are color-coded. The dark gray in the inner circle indicates GC content of the chloroplast genomes.
### TABLE 2 | Gene contents in five *Juglans* species chloroplast genomes.

<table>
<thead>
<tr>
<th>Category of genes</th>
<th>Group of gene</th>
<th>Name of gene</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-replication</strong></td>
<td>Ribosomal RNA genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>trnA-UGC&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>trnA&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>trnM-CAU&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnC-GCA</td>
</tr>
<tr>
<td></td>
<td>trnL-GUC&lt;sup&gt;b&lt;/sup&gt;</td>
<td>trnG-GCC&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>trnL-CAU&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnM-CAU&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>trnT-GGU&lt;sup&gt;a&lt;/sup&gt;</td>
<td>trnT-GUA</td>
</tr>
<tr>
<td></td>
<td>rps2</td>
<td>rps3</td>
</tr>
<tr>
<td></td>
<td>rps11</td>
<td>rps12&lt;sup&gt;ac&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>rps18</td>
<td>rps19</td>
</tr>
<tr>
<td></td>
<td>rpl2&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>rpl2&lt;sup&gt;ab&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>rpl32</td>
<td>rpl32</td>
</tr>
<tr>
<td></td>
<td>rpoA</td>
<td>rpoB</td>
</tr>
<tr>
<td></td>
<td>infA&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>Subunits of NADH-dehydrogenase</strong></td>
<td>ndhA&lt;sup&gt;b&lt;/sup&gt;</td>
<td>ndhA&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ndhF</td>
<td>ndhG</td>
</tr>
<tr>
<td></td>
<td>ndhK</td>
<td>ndhH</td>
</tr>
<tr>
<td><strong>Subunits of photosystem I</strong></td>
<td>psaA</td>
<td>psaB</td>
</tr>
<tr>
<td></td>
<td>ycf4&lt;sup&gt;c&lt;/sup&gt;</td>
<td>ycf4</td>
</tr>
<tr>
<td><strong>Subunits of photosystem II</strong></td>
<td>psbA</td>
<td>psbB</td>
</tr>
<tr>
<td></td>
<td>psbH</td>
<td>psbH</td>
</tr>
<tr>
<td></td>
<td>psbM</td>
<td>psbN</td>
</tr>
<tr>
<td></td>
<td>petA</td>
<td>petB</td>
</tr>
<tr>
<td><strong>Subunits of ATP synthase</strong></td>
<td>atpA</td>
<td>atpB</td>
</tr>
<tr>
<td></td>
<td>atpF</td>
<td>atpE</td>
</tr>
<tr>
<td><strong>Subunits of rubisco</strong></td>
<td>rbcL</td>
<td></td>
</tr>
<tr>
<td><strong>Tanslational initiation factor</strong></td>
<td>rpoC1&lt;sup&gt;f&lt;/sup&gt;</td>
<td>rpoC2</td>
</tr>
<tr>
<td></td>
<td>rpoC2</td>
<td></td>
</tr>
<tr>
<td><strong>Maturase</strong></td>
<td>matK</td>
<td></td>
</tr>
<tr>
<td><strong>Protease</strong></td>
<td>clpP&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>Envelope membrane protein</strong></td>
<td>cemA</td>
<td></td>
</tr>
<tr>
<td><strong>Subunit of Acetyl-CoA-carboxylase</strong></td>
<td>accD</td>
<td></td>
</tr>
<tr>
<td><strong>C-type cytochrome synthesis gene</strong></td>
<td>ccsA</td>
<td></td>
</tr>
<tr>
<td><strong>Genes of unknown function</strong></td>
<td>Conserved open reading frames</td>
<td>ycf1&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>Two gene copies in IRs.  
<sup>b</sup>Gene containing a single intron.  
<sup>c</sup>Gene containing two introns.  
<sup>d</sup>Pseudogene.

---

**Long Repeat Analysis**

*Juglans* Cpgs contained numerous forward repeats, palindromic repeats, and reverse repeats of at least 30 bp with a sequence identity ≥ 90% (Figure 4B, Table S5). These "long repeats" ranged from 30 to 44 bp in length and were repeated twice. Protein-coding genes (e.g., *rpoC1*, *psaB*, *petB*, and *ycf2*) contained a range of five to seven long repeat sequences (across species). Species also varied somewhat for number of long repeat sequences located in the intergenic regions (*J. regia* n = 24; *J. sigillata* n = 22; *J. hopeiensis* n = 21; *J. mandshurica*, n = 20; *J. cathayensis* n = 19; Table S5).

Depending upon species, we observed 12 or 13 forward repeats, 11 to 16 palindromic repeats, one or two reverse repeats, and one complementary repeat (only seen in *J. hopeiensis*)(Table 4, Table S5). The longest forward repeat unit was 44 bp; it was located in the *psbT-psbN* intergenic spacer of the LSC region of *J. regia* and *J. sigillata*. A different 44 bp repeat was located in the protein-coding genes *psaB-psaA* in the LSC of *J. cathayensis*, *J. hopeiensis*, and *J. mandshurica* (Table S5). In the sections *Juglans/Dioscaryon*, *J. sigillata* and *J. regia* each contained 13 forward repeats and two reverse repeats, and 16 (*J. regia*) or 13 (*J. sigillata*) palindromic...
repeats (Table 4, Table S5). In the section Cardiocaryon, J. cathayensis contained 13 forward and 11 palindromic repeats, J. hopeiensis contained 13 forward, 11 palindromic, onereverse, and one complementary repeat, and J. mandshurica contained 12 forward, 12 palindromic, and 1 reverse repeat (Table 4, Table S5). Tandem repeats of more than 20 bp and 100% sequence identity were identified in the intergenic spacers of trnK-UUU-rps16 (one repeat each in J. hopeiensis, J. mandshurica, and J. cathayensis); trnE-UUC-trnT-GGU (J. regia, 1; J. sigillata, 1; J. hopeiensis, 2; J. mandshurica, 1; J. cathayensis, 1); trnT-GGU-psbD (J. regia, 1; J. sigillata, 1; J. hopeiensis, 1; J. mandshurica, 2; J. cathayensis, 1); lhhA-trnG-UCC (J. hopeiensis, 1; J. mandshurica, 1; J. cathayensis, 1); ndhC-trnV-UAC (every Juglans species had one repeat); trnF-GAA-ndhJ (J. regia, 1; J. sigillata, 1); and trnG-UCC-trnM-CAU (J. regia, 1; J. sigillata, 1). Two identical tandem repeats were found in the protein-coding regions of all five Juglans Cpgs (Table S6).

Divergence Hotspots
The coding genes, non-coding regions, and introns were compared among the five Chinese Juglans species for divergence hotspots. The level of sequence divergence among all five species was estimated as the nucleotide variability value (Pi = 0.00219). The number of parsimony informative sites in coding genes, non-coding regions, and the complete Cpg was 192, 342, and 534, respectively (Table S7). The protein-coding CDS region was much more conserved than the IGS regions (i.e., LSC and SSC is much more conserved than the IR region). Within the CDS region, the ten genes with the greatest variability were rps3, psbL, petD, rpl22, psaJ, ndhD, rps19, rpoA, rpl32, and ndhA (Figure 5A), and the twelve least variable genes in CDS were petA, psbC, atpB, psbD, ndhG, ndhK, rps2, psbA, rbcL, psi, psaB, rrn23, and ycf2 (Figure 5A). Some IGS were quite conserved; rpl12-trnH-GUG, atpA-atpE, trnL-UGA-ccsA, psbC-trnS-UGA, ndhE-ndhG, rps19-rpl2, rpl14-rpl16, psi-psbT, lhhA-trnG-UCC, trnG-GCC-trnR-UCU, trnT-GGU/trnM-CAU-psbD, and trnP-UGG/trnP-GGG-psaJ showed lower levels of variation than genes located in the CDS region (Figure 5B). Across all five species, the regions with greatest sequence divergence were rps16-trnQ-UUG, trnE-UUC-trnT-GGU, trnT-GGU-psbD, petN-psbM, petB intron, rpoC2, ndhA, and ycf1. These intergenic regions were also generally rich in SSRs; rps16-trnQ-UUG had four SSRs [(T)10, (A)10, (T)11, and (A)11]; trnE-UUC-trnT-GGU had three SSRs [(T)10, (A)11, and (AT)11]; trnT-GGU-psbD had one SSR [(AT)5]; petN-psbM, one SSR [(T)10]; petB intron, two SSRs [(A)10 and (A)10]; rpoC2, three SSRs [(T)11, (T)11, (T)11]; ndhA intron, four SSRs [(A)15, (T)13aattg...T11, (AT)6]; and ycf1 had six SSRs [(T)11, (T)10, (T)12, (A)10, and (T)12]. Within section Juglans/Dioscaryon, rps4-trnT-UUG (1 SNP), ndhC-trnV-UAC (1 SNP), ycf1 (1 SNP, IRa), ccsA-ndhD, ycf1 (3 SNP, IRb) were variable. Within section Cardiocaryon, trnC-GCA-petN, trnE-UUC-trnT-GGU, trnT-GGU-psbD, and trnF-GAA-ndhJ were most variable (Figure 3). In total, we identified 610 SNPs or indels that were distinct between Juglans/Dioscaryon and Cardiocaryon.

Selective Pressures in the Evolution of Juglans
A total of 79 protein-coding genes were used to analyze synonymous and nonsynonymous change rates in Juglans. We identified five genes (matK, ycf1, accD, rps3, and rpoA) under positive selection (K_A/K_S ratio > 1; Figure S3; Table S8). The K_A/K_S ratio for accD for all five species was 1.23. The K_A/K_S ratio for matK for all five species was 1.34, for rpoA it was 1.17, and for rps3 it was 1.38 (Table S8). Interestingly, these five genes were previously found to present above average SNV and indel densities in exons (Table S8). All five genes were under positive pressure exclusively between sect.
Hu et al. Chloroplast Genomes of Five Juglans Species

FIGURE 3 | Sequence identity plot comparing the five Juglans chloroplast genomes with J. regia as a reference by using mVISTA. Vertical scale indicates the percentage of identity ranging from 50 to 100%. Coding regions are marked in blue and non-coding regions are marked in red. Gray arrows indicate the position and direction of each gene.

TABLE 3 | Summary of the simple sequence repeats (SSRs) in five Juglans species.

<table>
<thead>
<tr>
<th>Species</th>
<th>SSR Loci (N)</th>
<th>P1 Loci (N)</th>
<th>P2 Loci (N)</th>
<th>P3 Loci (N)</th>
<th>P4 Loci (N)</th>
<th>P5 Loci (N)</th>
<th>Pc Loci (N)</th>
<th>LSC</th>
<th>SSC</th>
<th>IRa</th>
<th>IRb</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. cathayensis</td>
<td>66</td>
<td>57</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>/</td>
<td>53</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>J. hopeiensis</td>
<td>83</td>
<td>62</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>67</td>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>J. mandshurica</td>
<td>83</td>
<td>62</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>67</td>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>J. regia</td>
<td>67</td>
<td>48</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>57</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>J. sigillata</td>
<td>66</td>
<td>49</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>56</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

aP1 to P5 indicate SSR loci with mono-, di-, tri-, tetra-, and pentanucleotide repeats, respectively. Pc indicates complex nucleotide repeats.

Cardiocaryon and sect. Dioscaryon; none of these five genes showed evidence of positive selection within either section (Figure S3; Table S8).

Phylogenetic Analysis

We used three datasets (whole complete Cpg, protein-coding exons, and non-coding region) to analyze the phylogenetic
relationships among members of two sections of Juglans and closely related species in the Betulaceae and Fagaceae. Arabidopsis thaliana and Populus alba were used as outgroups. Among the three datasets, complete Cpgs contained the greatest number of parsimony informative characters (531, 0.33%), followed by no-coding region (342, 0.42%) and protein-coding exons (192, 0.24%). The reconstructed phylogeny divided into four clades (Figure 6; Figures S4, S5, with members of the Betulaceae (Ostrya rehderiana and Betula nana) and I. rehderiana and Betula nana) joined to the five Juglans species and distinct from the other Fagaceae, irrespective of dataset. Within Juglans, the five Chinese species were divided into two clades corresponding to the two sections (Juglans/Dioscaryon and Cardiocaryon) with 100% bootstrap (BS) support based on Maximum Likelihood (ML) and Maximum parsimony (MP) analysis (Figure 6A; Figures S4A,B). Analysis of the whole cp genomes of the five Chinese walnut species and 10 eudicot outgroups using Bayesian inference (BI) resulted in cladograms with topology similar to ML and MP, and strongly supported phylogenetic trees based on each of three datasets (whole cp genome sequences, protein coding sequences, and the introns and spacers) (Figure 6B; Figures S4C,D). In section Juglans/Dioscaryon, J. regia and J. sigillata were split with a 100% BS, while the Cardiocaryon clade (J. cathayensis and J. hopeiensis, J. mandshurica) diverged from sect. Juglans with 100% BS value (Figure 6; Figures S4, S5). J. hopeiensis was closer to J. mandshurica than to J. cathayensis (Figure 6; Figures S4, S5). We constructed the divergence time tree among five Chinese walnut species based on whole chloroplast genome sequences. The results showed that the divergence time between two sections was 7.91Myr, while J. regia and J. sigillata diverged much more recently (0.05 Myr), and J. cathayensis diverged from J. mandshurica and J. hopeiensis before 3.51Myr (Figure S5).

DISCUSSION

Chloroplast Sequence Variation and Evolution

In the present study, we sequenced the chloroplast genomes of five Juglans species, annotated the chloroplast genomes, identified SSR and tandem repeats within the genomes, and carried out a phylogenetic analysis comparing them to ten other chloroplast genomes. Our results have laid the foundation for future studies on the evolution of chloroplast genomes of walnuts and butternuts, as well as the molecular identification of Juglans species.

Most angiosperm chloroplasts contain 74 protein-coding genes, while an additional five are present in few species (Millen et al., 2001). The five Juglans Cpg we sequenced revealed 88 protein-coding genes (79 unigenes were protein-coding), 40 tRNA genes, and 8 rRNA genes, which is similar to Quercus (Du et al., 2015; Lu et al., 2016; Yang et al., 2016). The number of tRNA genes and rRNA genes in Juglans was the same as in five Quercus species (Yang et al., 2016). Moreover, the total number of introns in the Juglans Cpg was the same as Quercus rubra (Alexander and Woeste, 2014), Ampelopsis (Raman and Park, 2016), and Saxifragales (Dong et al., 2013). Several lineages of angiosperms have independently lost introns from the ribosomal protein genes rps16, rps12, and rpl16 (Downie et al., 1991; Downie and Palmer, 1992), including Geraniaceae and Caryophyllales (Logacheva et al., 2008). The five Chinese Juglans species have not lost introns in any of these genes, however, a characteristic they have in common with the woody plant family Vitaceae (Raman and Park, 2016).

The gene infA encodes translation initiation factor 1. It has been lost completely in some angiosperms (Millen et al., 2001; Steane, 2005), is present as a pseudogene in the majority of

### TABLE 4 | Summary of the long repeat sequences in five Juglans species chloroplast genomes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Forward</th>
<th>Palindromic</th>
<th>Reverse</th>
<th>Complement</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. cathayensis</td>
<td>13</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J. hopeiensis</td>
<td>13</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>J. mandshurica</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>J. regia</td>
<td>13</td>
<td>16</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>J. sigillata</td>
<td>13</td>
<td>14</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

*Long repeat sequences were at least 30 bp with a sequence identity ≥ 90%.*
angiosperm (Millen et al., 2001; Steane, 2005), and is present and presumed functional in Quercus robur and Quercusrubra (Alexander and Woeste, 2014). In this study, we identified nine internal stop codons in *Juglans/Dioscaryon* versus five, five, and three internal stop codons in the *infA* sequence of *J. hopeiensis*, *J. mandshurica*, and *J. cathayensis* Cpgs, respectively. Thus, although *infA* is a pseudogene in all *Juglans/Dioscaryon* and *Cardiocaryon* for which there are data, there are inter-sectional differences that deserve additional study (Figure 2A), and *infA* may reveal important phylogenetic information concerning section *Rhysocaryon*. We also observed that the hypothetical gene *ycf15* was truncated in *Dioscaryon* species and *Cardiocaryon* species by five and three internal stop codons, respectively (Figure 2B). A similar truncation was seen in *Quercus aliena* (Lu et al., 2016, *ycf15*) and *Quercus spinosa* (Du et al., 2015) of Fagaceae, in *Lilium* (Liu et al., 2012b), Kiwi fruit (*Actinidia chinensis* var. *chinensis*) (Yao and Huang, 2016), and *Vaccinium macrocarpon* (Fajardo et al., 2013). *ycf15* is a pseudogene in all families of Saxifragales (Dong et al., 2013), but may be a functional protein coding gene in *Thalictrum coreanum* (Ranunculaceae, Park et al., 2015). The role of *ycf15* as a protein coding gene remains unclear and requires further study.

Variability in copy number of simple sequence repeats (SSRs) in the chloroplast makes them important molecular markers for distinguishing lower taxonomic levels (Yang et al., 2011; Xue et al., 2012). Cp SSRs have been used widely in plant population genetics (Doorduin et al., 2011; He et al., 2012), polymorphism investigations (Xue et al., 2012), and ecological and evolutionary studies (Roullier et al., 2011; Wang et al., 2013). The SSRs in the five *Juglans* Cp genomes we investigated were AT rich. Poly (A)/(T) SSRs are more common than poly (G)/(C) in many plant families (Melotto-Passarin et al., 2011; Nie et al., 2012; Martin et al., 2013). The cpSSRs of the five *Juglans* we studied are expected to be useful for assays detecting polymorphisms at population-level as well as comparing more distantly phylogetic relationships among *Juglans* species.

Large and complex repeat sequences may play an important role chloroplast genome arrangement and sequence divergence (Timme et al., 2007; Guisinger et al., 2011; Weng et al., 2013). We found numerous repeated sequences in the Cpgs of *Juglans,*
particularly in the intergenic spacer regions, similar to those reported in other angiosperm lineages (Yang et al., 2016). We found that repeats in petB, psaA, and ycf2 differed between species in different sections of Juglans, and the same was true of repeats in the gene junctions (trnK-UUU-rps16, trnV-GAC-rps7, trnT-GGU-psbD, and trnT-GGU-psbD) (Table S5). These divergence hotspots within Juglans Cpg sequences are potentially important resources for developing molecular markers for phylogenetic analyses and identification of Juglans species (Stanford et al., 2000; Aradhya et al., 2007).

Phylogenetic Analysis

The classical taxonomy of Juglans based on non-coding regions of the Cpg supported the separation of J. regia and J. sigillata into Sec. Juglans/Dioscaryon and other three Juglans species (J. cathayensis, J. hopeiensis, J. mandshurica) into Sec.
Cardiocaryon (Stanford et al., 2000; Aradhya et al., 2007). Whether J. regia and J. sigillata are legitimately distinct taxa in China has been controversial; Iron walnut (J. sigillata) could be an independent species based on RAPD and EST-SSR data (Wu et al., 2000; Qi et al., 2011) and based on RFLP and Cp DNA fragments (92% bootstrap value) (Aradhya et al., 2007). Our data support their maintenance as distinct taxa.

Members of the Cardiocaryon are morphologically distinct from other Juglans in that they have red stigmas, number of leaflets per leaf, and in the number of fruits typically found in a cluster, but the phylogenetic relationships within sect. Cardiocaryon are unsettled. J. hopeiensis is sympatric with J. mandshurica, and based on data from AFLPs and isozymes, some have concluded that J. hopeiensis is a hybrid species between J. regia and J. mandshurica (Wenheng, 1987; Zhang et al., 2009), consistent with the interpretation of floral evolution in the genus by Xi (1987). All phylogenetic trees based on our data indicate that J. hopeiensis is closer to J. mandshurica than J. cathayensis, and that the latter two species are distinct, in contrast to the Flora of China (1999), which relies exclusively on morphological data. The relationship between J. hopeiensis and J. ailantifolia, the only other Asian member of the Cardiocaryon, is now an important question. These results showed that the Stanford et al. (2000) and Aradhya et al. (2007) taxonomy of Juglans is reasonable on the whole. In this study, J. regia and J. sigillata were divided from each other with a 100% BS, while J. cathayensis, J. hopeiensis, and J. mandshurica diverged from sect. Juglans with 100% BS value (Figure 6). Each of the five species is supported as independent species based on whole chloroplast genome sequences.

In this study, the five Chinese walnut species and 10 eudicot outgroups were represented with well-supported cladograms with highly similar topology and strongly supported phylogenetic trees using Maximum Likelihood (ML), Bayesian inference (BI), and Maximum parsimony (MP) analysis. Analysis using whole Cpg sequences, protein coding sequences, and the introns and spacers resulted in consistent and strongly supported results (Figure 6, Figure S4). Our results confirmed that the phylogenetic relationships among the five Chinese Juglans based on chloroplast sequences only are in congruence with those reported by Stanford et al. (2000) and Aradhya et al. (2007). Each of the two sections was confirmed to be monophyletic (Dode, 1909; Manning, 1978). Within sect. Dioscaryon, division of the two species was highly supported, as suggested by Aradhya et al. (2007). With the exception of section Cardiocaryon (Dode, 1909; Manning, 1978), relationships among three Chinese walnuts were fully resolved and statistically supported (P = 0.95; BS = 100%). Stanford et al. (2000) and Aradhya et al. (2007) recovered an unsupported sister relationship between J. mandshurica and J. cathayensis because J. hopeiensis was not included in those analyses (Stanford et al., 2000). Previously suggested relationships among members of section Cardiocaryon were confirmed by our data with even higher support than in Stanford et al. (2000) and Aradhya et al. (2007), although our analysis did not include Japanese walnut (J. ailantifolia), the final member of Cardiocaryon. The chloroplast-based phylogeny presented in this work and by others is not a complete understanding of the evolutionary relationships among these five Chinese Juglans because events we did not consider, including incomplete lineage sorting, chloroplast capture, horizontal transfer, and local fixation of Cpg haplotypes can all influence phylogeny (Stegemann et al., 2012; Mariac et al., 2014; Novikova et al., 2016).

The divergence time between the two Asian Juglans sections was estimated at 7.91 Myr, although several Juglans species diverged quite recently within each section (Figure 6, Figure S4). The deep evolutionary relationships and divisions within the two Asian sections needs further investigation. The molecular phylogeny of the entire genus (Juglans) and its relationship to other genera in the Juglandaceae also awaits more evidence. These Cpg sequences will provide genetic information necessary to understand the evolution of plastid genomes via phylogenomics.

DATA ARCHIVING STATEMENT

The chloroplast genome sequences of Chinese walnut (Juglans) species were submitted on the National Center for Biotechnology Information (NCBI), the accession numbers were: KT820730, KT820731, and KT820732, KT820733.

ETHICS STATEMENT

This article does not contain any studies with human participants performed by any of the authors.

AUTHOR CONTRIBUTIONS

PZ, YH, and KW designed and performed the experiment as well as drafted the manuscript. YH and PZ collected the samples. YH and PZ completed the sequence assembly and analyzed the data. KW and PZ conceived the study and revised the manuscript. All the authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 41471038; No. 31200500; No. J1210063), the Program for Excellent Young Academic Backbones funding by Northwest University, the Northwest University Training Programs of Innovation and Entrepreneurship for Graduates (No. YZZ15062), Changjiang Scholars and Innovative Research Team in University (No. IRT1174). 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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01955/full#supplementary-material
Figure S1 | Comparisons of LSC, SSC, and IR region borders among the five Chinese Juglans chloroplast genomes.

Figure S2 | Frequency distribution of major SSRs based on main motif type in the five Chinese Juglans cp genomes. Jh, Juglans hopeiensis; Jc, J. cathayensis; Jm, J. mandshurica; Jr, J. regia; Js, J. sigillata.

Figure S3 | Gene-specific $K_r/K_S$ values between the chloroplast genomes of two Juglans species (J. regia and J. cathayensis) representing section Juglan/ Dioscaryon and section Cardiocaryon, respectively. Five genes (matK, rbcL, accD, rps3, and rpoA) returned $K_r/K_S$ values greater than 0.8, whereas the $K_r/K_S$ values of the other genes were below 0.8.

Figure S4 | Phylogenetic tree construction of five Juglans species plus eight other taxa. (A) Maximum Likelihood (ML) tree and Maximum parsimony (MP) tree based on protein coding sequences. (B) Maximum Likelihood (ML) tree and Maximum parsimony (MP) tree based on the introns and spacers. (C) Bayesian inference (BI) tree based on protein coding sequences. (D) Bayesian inference (BI) tree based on the introns and spacers. Numbers above branch indicate the bootstrap (BS) support value.

Figure S5 | Phylogenetic timetree construction of five Chinese Juglans species plus eight other taxa based on whole cp genome sequences. Blue bars and the numbers at the nodes indicate 95% highest posterior densities (HPDs) of time estimates (million years ago, Myr).

Table S1 | Primers used for genome sequence validation.

Table S2 | The information of a total of 15 species used for phylogenetic analysis.

Table S3 | Indels and single nucleotide polymorphisms (SNP) in the five Chinese Juglans chloroplast genomes.

Table S4 | Simple sequence repeats in each of five Chinese Juglans species.

Table S5 | The information of the function nucleic acid sequences for five Chinese Juglans species.

Table S6 | The length of tandem repeats distribution in five Chinese Juglans species.

Table S7 | The number of variable sites in five Chinese Juglans species.

Table S8 | $K_r/K_S$ ratio for protein coding sequences for five Chinese Juglans species. Jh, Juglans hopeiensis; Jc, J. cathayensis; Jm, J. mandshurica; Jr, J. regia; Js, J. sigillata.

REFERENCES


30, 772–780. doi: 10.1093/molbev/msm010
Kezar, M., Moir, R., Wilson, A., Stones-Havas, S., Che勒rmacher, M., Stoye, J., and
Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on
a genomic scale. Nucleic Acids Res. 29, 4633–4642. doi: 10.1093/nar/29.22.4633
integrated web server for the annotation, visualization, analysis, and GenBank
submission of completely sequenced chloroplast genome sequences. BMC Genomics
Liu, J., Qi, Z. C., Zhao, Y. P., Fu, C. X., and Xiang, Q. Y. (2012b). Complete cpDNA
sequence of Smilax china and phylogenetic placement of Liliales–Influences
Comparative chloroplast genomics and phylogenetics of Fagopyrum esculentum
Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW—a suite of tools for generating physical maps
of plastid and mitochondrial genomes and visualizing expression data sets.
detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25,
955–964. doi: 10.1093/nar/25.3.955
of the Oriental white oak: Quercus aliena Blume. Mitochondrial DNA A 27,
2802–2804. doi: 10.1080/19401736.2015.1053074
Gard. 65, 1058–1087.
Cost-effective enrichment hybridization capture of chloroplast genomes at deep
multiplexing levels for population genetics and phylogeography studies. Mol.
Zingiberales): insight into plastid monophyly evolution. PLoS ONE
8:e53350. doi: 10.1371/journal.pone.0053350
Melotto-Passarin, D. M., Tambarussi, E. V., Dressano, K., De Martin, V. F.,
and Carrer, H. (2011). Characterization of chloroplast DNA microsatellites from
doi: 10.4238/vol10-3mn1019
Milen, R. S., Olmstead, R. G., Adams, K. L., Palmer, J. D., Lao, N. T., Heggie,
angiosperm evolution with multiple independent transfers to the nucleus. Plant
Cell 13, 645–658. doi: 10.1105/tpc.13.3.645
species, croton weed (Ageratina adenophora). PLoS ONE 7:e36869. doi: 10.1371/journal.pone.0036869
Novikova, P. Y., Hohnmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G.,
et al. (2016). Sequencing of the genus Arabidopsis identifies a complex history
of nonfertilizing speciation and abundant trans-specific polymorphism. Nat.
Genet. 48, 1077–1082. doi: 10.1038/ng.3617
Thalictrum coreanum (Ranunculaceae) and transfer of the p32 gene to the
nucleus in the ancestor of the subfamily Thalictridoideae. BMC Plant Biol.
control of next generation sequencing data. PLoS ONE 7:e36019. doi: 10.1371/journal.pone.0036019
Pollegioni, P., Woeste, K. E., Chiochini, F., Del Lungo, S., Olimpieri, I., Tortolano,
V., et al. (2015). Ancient humans influenced the current spatial genetic
in phylogenetics: advantages of Akaikie information criterion and
Bayesian approaches over likelyhood ratio tests. Syst. Biol. 53, 793–808.
doi: 10.1080/10635150490252304
Raman, G., and Park, S. (2016). The complete chloroplast genome sequence of
Rouiller, C., Rossel, G., Tay, D., McKey, D., and Lebot, V. (2011). Combining chloroplast and nuclear microsatellites to investigate origin and dispersal
Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington,
704–730. doi: 10.3732/ajb.1000406
Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis
of Juglans (Juglandaceae) based on matK and ITS sequence data.
Stegemann, S., Keuthe, M., Greiner, S., and Bock, R. (2012). Horizontal transfer
of chloroplast genomes between plant species. Proc. Natl. Acad. Sci. USA. 109,
2434–2438. doi: 10.1073/pnas.1114076109
null