Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Complete Chloroplast Genome of Chinese Hickory (*Carya Cathayensis*) and Genetic Comparison to Its Closely Related Species

Xiangtao Zhu
Jiyang College of Zhejiang A and F University
Jianshuang Shen
Jiyang College of Zhejiang A and F University
Xueqin Li
Jiyang College of Zhejiang A and F University
Xia Chen
Jiyang College of Zhejiang A and F University
Xiaoling Huang
Jiyang College of Zhejiang A and F University
Songheng Jin (shjin@zafu.edu.cn)
Zhejiang A&F University

Original Article

Keywords: Carya cathayensis, Chloroplast genome, Genome skimming, Phylogenetic relationship

Posted Date: July 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-719966/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: *Carya cathayensis*, an important economic nut tree, is narrowly endemic to Eastern China in the wild. Although the nuclear genome of this plant has been reported recently, its chloroplast (cp) genome is yet to be decrypted.

Results: Here, the complete cp genome of *C. cathayensis* was sequenced with NGS Illumina HiSeq2500, analyzed, and compared to its closely related species. The cp genome is 160,825 bp in length with an overall GC content of 36.13%. It displays a quadripartite structure with a large single copy (LSC) region of 90,127 bp and a small single copy (SSC) region of 18,760 bp, separated by a pair of inverted repeats (IRs) of 25,969 bp. The genome contains 131 genes, including 86 protein-coding genes, 37 tRNA genes, and eight rRNA genes. The codon usage frequency and repeat sequences (including 252 simple sequence repeats (SSRs) and 55 long-repeats were identified.

Conclusions: Phylogenetic analysis revealed that *Juglandaceae* is monophyletic, and that *C. cathayensis* is sister to *C. kweichowensis* and *C. sinensis*. Comparison of the *C. cathayensis* cp genome with those of the closely related species in *Juglandaceae* revealed that the noncoding regions are highly mutated, suggesting a great potential in phylogenetic researches. The newly characterized cp genome of *C. cathayensis* provides valuable information for further studies of this economically important species.

Background

Genus *Carya* belong to family *Juglandaceae*, comprises ~18 species and four variants, which are distributed in the temperate and tropical regions of East Asia and eastern North America (Lu et al. 1999; Zhang et al. 2013). *Carya* species from East Asian and eastern North American are phylogenetically separated (Zhang et al. 2013), while the relationships among some taxa within the genus have not been resolved yet.

Nuclear- and plastid-DNA are the basics for phylogenetic re-construction; the single- or low-copy nuclear genes are most suitable for systematic analyses (Naumann et al/ 2011). Till now, several plastid (*matK, rbcL-atpB, rpoC1, rps16, trnH-psbA*, and *trnL-F*) and nuclear (ITS and *phyA*) DNA markers have been used for the phylogenetic study of genus *Carya*. Identified these nuclear genes by ortholog screening, cloning, and sequencing can be costly and time-consuming. Compared with nuclear genes, chloroplast (cp) genome is an excellent alternative owing to its small size (75-250Kb) (Raman et al. 2017), easily obtainable sequences by the low-cost Next Generation Sequencing (NGS), and less interference from homologous regions. Besides the genic regions, the noncoding regions of cp genomes can also be harness for phylogenetic analysis due to a relatively high level of genetic variation resulted from the low selective pressure (Böhle et al. 1994). In addition, structural rearrangements, such as the loss of introns, genes, or even the inverted repeats, extensively occur in the plastid genomes of many flowering plants (Li et al. 2018, 2019; Zhao et al. 2018; Zeng et al. 2017; Xu et al. 2017; Yang et al. 2016). Recently, the cp genomes of *Carya kweichowensis* (Ye et al. 2018), *Carya cathayensis* (Zhai et al. 2019) and *Carya*

illinoinensis (NBCI accession NC_041449.1) have been published, and the publication of more cp genomes of *Carya* species will facilitate the identification of genetic variations via sequence comparison, providing new insights into the evolutionary history and interspecific relationships among *Carya* species.

Carya cathayensis (Chinese hickory) is naturally distributed in moist valleys at altitudes of 500–1200 m in Zhejiang, Jiangxi, and Anhui Provinces. Because of its high nutritional and economic values, *C. cathayensis* has been widely cultivated in Zhejiang (Zhang et al. 2015). *C. cathayensis* is an important economic nut tree, and vulnerable to abiotic factors (Grauke et al. 2016; Zhang et al. 2012) suggesting that suitable habitat is essential for its survival in the wild. The changes in climate and over-exploitation in recent years has made the conservation of wild *C. cathayensis* populations becomes an urgent task. Although the nuclear genome and cp genome of *C. cathayensis* had been released (Zhai et al. 2019; Huang et al. 2019) its cp genome has not been reported detailed. Information on the *C. cathayensis* cp genome is essential for the development of conservation and breeding strategies.

In this study, we present the whole plastome sequence of *C. cathayensis* and explore the utility of this new genomic resource and relationship with those of other *Carya* species. These results will lay a foundation for future phylogenetic and structural diversity studies of *Carya*.

Results And Discussion

Genome Features of the C. cathayensis

Filtering of the raw sequencing data yielded a total of 12,470,465 clean paired-end reads. There were 3.7G bases, of which 89.47% bases had a quality score higher than Q30. The whole cp genome of *C. cathayensis* is 160,825 bp in length, with a GC content of 36.13%. The genome assembly had an average read coverage of higher than 700×. The synteny was identified by comparing the *C. cathayensis* cp genome to the reference (Table 1), which presents most of sequences of the genomes were conservative. These data suggest that the complete cp genome of *C. cathayensis* is of high quality and can be used for downstream analyses.

Table 1. Statistics of the synteny between the *C. cathayensis* and *Cyclocarya paliurus* cp genomes

rstart	rend	qstart	qend	rlength	qlength	Identity (%)	qstrand
22	9509	12	9567	9488	9556	96.25	1
9509	40748	9672	40744	31240	31073	97.03	1
40678	46923	40831	47079	6246	6249	98.72	1
47055	118560	47084	118669	71506	71586	98.07	1

Note: (NCBI accession: NC_034315). rstart: starting position of the reference genome; rend: ending position of the reference genome; qstart: starting position of the query genome; qend: ending position of the query genome; rlength: length of the reference genome; qlength: length of the query genome; qstrand: strand of the query genome, 1 for the plus strand. The *Cyclocarya paliurus* cp genome was used as the reference.

Like those of all other angiosperms, the circular genome of *C. cathayensis* displays a typical quadripartite structure (Daniell et al. 2016), including a pair of inverted repeats (IRs; 25,969 bp each), separated by a large single copy (LSC; 90,127 bp) and a small single copy (SSC; 18,760 bp) regions (**Fig. 1**). The overall GC content is 36.13, which is similar to that observed for other *Carya* species (35.8-36.3%). (Ye et al. 2018; Wang et al. Unpublished; Hu et al. 2016). The IRs has a relatively higher GC content compared with other regions (**Fig. 2**), a feature similar to the cp genomes of other flowering plants (Hu et al. 2016; Morton et al. 2003; Liu et al. 2018). A total of 131 annotated genes were identified, among which 86 are protein-coding genes, 37 are transfer RNA (tRNA) genes, and eight are ribosomal RNA (rRNA) genes (Table 2). We found 17 duplicated genes, including *ndhB*, *rpl2*, *rpl23*, *rps12*, *rps7*, *rrn16*, *rrn23*, *rrn4.5*, *rrn5*, *tRNA-ACG*, *tRNA-CAA*, *tRNA-CAU*, *tRNA-GAC*, *tRNA-GUU*, *tRNA-UGC*, *tRNA-UUC*, *tRNA-UGC*, and *ycf2* (Table 2). In total, 21 intron-containing genes (13 protein-coding and 8 tRNA genes) were annotated, among which only three protein-coding genes (*rps12*, *ycf3* and *clpP*) with two introns and the other genes with one intron. Gene *rps12* gene of *C. cathayensis* has its 5'-end exon situated in the LSC region and its 3'-end exons located in the IRs region (**Fig. 1**), the results was similar to the congeneric species *C. sinensis* (Hu et al. 2016). However, there is a certain difference with previous reports of the *C. cathayensis* cp genome, such as the length, GC contents, and annotated genes of the whole cp genome (Zhai et al. 2019). The difference may be due to the geographical isolation of different plants, which facilitate the identification of genetic variations via sequence comparison, providing new insights into the evolutionary history of *C. cathayensis.*

Table 2. Annotated genes in the *C. cathayensis* cp genome.

	Category	Name of Gene						
Self-replication	Ribosomal RNA	*rrn16*	*rrn23*	*rrn4.5*	*rrn5*	*rrn16*	*rrn5*	*rrn4.5*
		rrn23						
	Transfer RNA	*tRNA-GUG*	*tRNA-UUU*	*tRNA-UUG*	*tRNA-GCU*	*tRNA-CGA*	*tRNA-UCU*	*tRNA-GCA*
		tRNA-GUC	*tRNA-GUA*	*tRNA-UUC*	*tRNA-GGU*	*tRNA-UGA*	*tRNA-GCC*	*tRNA-CAU*
		tRNA-GGA	*tRNA-UGU*	*tRNA-UAA*	*tRNA-GAA*	*tRNA-UAC*	*tRNA-CAU*	*tRNA-CCA*
		tRNA-UGG	*tRNA-CAU*	*tRNA-CAA*	*tRNA-GAC*	*tRNA-UUC*	*tRNA-UGC*	*tRNA-UGC*
		tRNA-ACG	*tRNA-GUU*	*tRNA-UAG*	*tRNA-CAU*	*tRNA-CAA*	*tRNA-GAC*	*tRNA-ACG*
		tRNA-UGC	*trnA-UGC*	*tRNA-UUC*	*tRNA-GUU*			
	Small subunit of ribosome	*rps16*	*rps2*	*rps14*	*rps4*	*rps18*	*rps12*	*rps12*
		rps11	*rps8*	*rps3*	*rps19*	*rps7*	*rps15*	*rps7*
	Large subunit of ribosome	*rpl33*	*rpl20*	*rpl36*	*rpl14*	*rpl16*	*rpl22*	*rpl2*
		rpl23	*rpl32*	*rpl23*	*rpl2*			
	RNA polymerase subunits	*rpoC2*	*rpoC1*	*rpoB*	*rpoA*			
	Subunits of photosystem I	*psaB*	*psaA*	*psaI*	*psaJ*	*psaC*		
	Subunits of photosystem II	*psbA*	*psbK*	*psbI*	*psbM*	*psbD*	*psbC*	*psbZ*
		psbJ	*psbL*	*psbF*	*psbE*	*psbB*	*psbT*	*psbN*
		psbH						
Photosynthsis	Subunits of cytochrome	*petN*	*petA*	*petL*	*petG*	*petB*	*petD*	
	Subunits of ATP synthase	*atpA*	*atpF*	*atpH*	*atpI*	*atpE*	*atpB*	
	Large subunit of RuBisCO	*rbcL*						
	Subunits of NADH	*ndhJ*	*ndhK*	*ndhC*	*ndhB*	*ndhF*	*ndhD*	*ndhE*
		ndhG	*ndhI*	*ndhA*	*ndhH*	*ndhB*		
Other gene	Maturase	*matK*						

	Envelope membrane protein	*cemA*					
	Subunit of acetyl-CoA	*accD*					
	C-type cytochrome synthesis gene	*ccsA*					
	Protease	*clpP*					
Unknown function	Conserved open reading frames	*ycf3*	*ycf4*	*ycf2*	*ycf1*	*ycf2*	

The relative frequency of synonymous codons of the *C. cathayensis* cp coding sequence were estimated, the results had shown that a total of 26,476 codons were found and the four most frequently used codons were AUU (Isoleucine), AAA (Lysine), GAA (Glutamic acid), and AAU (Asparagine), pertaining to 1145 (4.32%), 1066 (4.03%), 1040 (3.93%), and 1004 (3.79%) codons, respectively (**Table S1 and Fig. 3**). The two most frequently used amino acids were leucine (2780) and isoleucine (2350), cysteine was the least abundant, with only 308 hits. Morton (2003) has proposed that the codon usage bias of cp genomes may be a result of selection and mutation. These codon usage frequencies are similar to those reported in other angiosperms (Li et al. 2018, 2019; Liu et al. 2018; Jian et al. 2018), these features of codon usage preference can help to better decipher exogenous gene expression and the mechanisms of cp genome evolution (Liu et al. 2018).

Analysis of long-repeats and simple sequence repeats (SSRs)

We identified 24 forward, nine reverse, three complement, and 13 palindrome repeats in the cp genome of *C. cathayensis* (Table S2). Most repeats ranged from 20 to 62 bp in length. The longest forward repeat with 62 bp resided in the LSC region. A total of 46, five, and four long repeats were found in the LSC, SSC, and IR regions, respectively. Three forward repeats were found in the two IRs, including one repeat associated with the *rpl14* and *tRNA-UGC* genes, one with the *IGS* genes, and one with the *tRNA-CCA* and *tRNA-GUU* genes.

A total of 252 SSRs were identified in the *C. cathayensis* cp genome (Table S3), among which 199, 12, 64, two, and one were mono-, di-, tri-, tetra-, and pentanucleotide repeats, respectively. Mononucleotide SSRs were the richest (occupied 78.97%) and the mononucleotide A+T repeat units occupied the highest portion (75.00%), the results were consist with previous study (Shen et al. 2017). The cpSSR markers are excellent tools for phylogenetic research due to several characteristics, including non-recombination, haploidy, uniparental inheritance, and low substitution rate (Ebert and Peakall 2009); they are especially valuable for intraspecific population genetic variation research (Provan et al. 2011; Diekmann et al. 2012) and interspecific evolutionary and identification studies (Singh et al. 2017; Hu et al. 2009; Deng et al. 2017; Pan et al. 2014; Huang et al. 2015). Hence, the *C. cathayensis* cpSSRs provided here offer a new avenue for the development of species protection and preservation strategies.

Phylogenetic analysis

Phylogenetic analysis was carried out based on an alignment of the concatenated nucleotide sequences of all 47 angiosperm cp genomes (**Fig. 4**). MAFFT was employed for multiple sequence alignment. The phylogenetic relationship was reconstructed using the GTR-GAMMA model by RAxML, and *Malus prunifolia*, *Ulmus gaussenii*, and *Dalbergia hainanensis* were used as outgroups. Almost all relationships inferred from the cp genome data based on the Maximum Likelihood (ML) tree received strong support, with the support values ranging from 47 to 100. The genus *Quercus* was polyphyletic in our analysis, resulting from the embedded branches of the genera *Lithocarpus* and *Castanea*, this result was in consistent with previous results (Li et al. 2018). In addition, genera *Betula*, *Corylus*, and *Ostrya* were found sister to *Juglans*, whereas *Platycarya* and *Cyclocarya* were more closely related to *Juglans*.

The well supported phylogenetic tree (**Fig. 4**) indicates that genus *Carya* is monophyletic and is most closely related to the cluster formed by another genus of *Juglandaceae*, which is consistent with previous studies (Zhang et al. 2013; Ye et al. 2018). *C. cathayensis* is sister to *C. kweichowensis* and they are sister to *C. illinoinensis* successively with high support scores (Bootstrap = 100; **Fig. 4**).

Comparative analysis of genome structure

The size variation of angiosperm plastid genomes is often accompanied by the expansion and contraction of the IR and SSC boundary regions (Dugas t al. 2015; Drescher t al. 2000). To further resolve the structural evolutionary history of the cp genomes of genus *Carya*, we compared the IR/SSC and IR/LSC junctions across six selected *Juglandaceae* species, including *C. cathayensis*, *C. illinoinensis*, *C. kweichowensis*, *Platycarya strobilacea*, *Cyclocarya paliurus*, and *Juglans cathayensis*. The results of the IRscope analysis are presented in **Fig 5**. We observed a wide variability of the junction sites in these cp genomes. For example, in genus *Carya*, *C. cathayensis* exhibited similar JLB, JSB, JSA junction sites compared with its elder sister species *C. illinoinensis* (Figures 4 and 5). Both species had a IRa/b region of ~25900 bp and an SSC region of ~18700 bp. By contrast, *C. kweichowensis*, which is most closely related to *C. cathayensis* and *C. illioninensis*, displayed an extremely large IRa/b region of 40943bp. In addition, the *C. kweichowensis* cp genome showed some striking structural differences compared to its sister species. For example, the *rps19* gene was shifted by 61 bp from the LSC to IRb at the LSC/IRb border, *ccsA* and *trnL* were located in the IRa/b regions instead of the SSC region, and the *ycf1* was absent from the JSA site. Moreover, we observed variations of the IR/SSC and IR/LSC junction sites across other genera in family *Juglandaceae* (**Fig. 5**).

We next performed a cp genome identity analysis on the six *Juglandaceae* species described above, with the *C. cathayensis* cp genome as reference (**Fig. 6**). This analysis revealed a relatively higher level of divergence in the noncoding than in the coding regions, similar to that has been reported for genus *Quercus* from family *Fagaceae* (Li t al. 2018), which is related to family *Juglandacea*. We also identified a considerable number of variations in the noncoding cp sequences, such as *tRNA-CGA*, *tRNA-CCC*, *tRNA-CAU*, and *tRNA-UAG*, of species in genus *Carya* (**Fig. 6**). Hence, these noncoding sites may be useful for resolving the suspending phylogenetic relationships of *Carya* species (Zhang t al. 2013).

Genes nucleotide variability (pi) values of six selected *Juglandaceae* species (including *C. cathayensis*, *C. illinoinensis*, *C. kweichowensis*, *Platycarya strobilacea*, *Cyclocarya paliurus*, and *Juglans cathayensis*) were showed in **Fig. 7**, the values of *LSC.rpl36, IR. rrn4.5, rrn23, and rrn16* were higher than 1, while the values of other genes were lower than 0.03. The results showed that there have a lower nucleotide diversity among the six *Juglandaceae* species. To test whether the remaining cp genes in these six *Juglandaceae* species have undergone selection, we estimated the synonymous (*Ks*) and nonsynonymous (*Ka*) substitution rates (Table S4). The *Ka/Ks* ratios were then categorized, with *Ka/Ks* < 1, *Ka/Ks* = 1, and *Ka/Ks* > 1 denoting purifying, neutral, and positive selections, respectively, in the context of a codon substitution model. According to our results, only seven genes of *Carya cathayensis*, *rps15, rpoA, rpoB, petD, ccsA, atpI,* and *ycf1-2* underwent positive selection compared with the other *Juglandaceae* species (Table S4). By contrast, most genes were shown to have undergone purifying selection, which was evidenced by a *Ka/Ks* ratio below 1 and the presence of negatively selected sites within some genes.

Conclusions

The newly characterized *C. catayensis* cp genome provides valuable genetic information for the phylogenetic study and the development of conservation strategies of genus *Carya*.

Methods

DNA extraction, sequencing, and cp genome assembly

The young green leaves of *C. cathayensis* were collected from the nursery of Zhejiang A&F university (stored in Institute of Botany, the Chinese academy of Sciences Mem and the specimen Accession number is PE00820836) and stored immediately at -80°C. Total genomic DNA was isolated from the leaves using a modified CTAB method (Doyle t al. 1987). After ensuring the quality of DNA, shotgun libraries (250 bp) were constructed in accordance with the standard protocol suggested by the manufacturer's instructions (Illumina Inc., San Diego, CA, USA). Sequencing was performed with an Illumina Hiseq 2500 platform (Genepioneer Biotechnologies Co., Ltd.; Nanjing, China) with the PE150 strategy.

Quality control for the raw sequencing data was carried out using package FastQC (Version 0.11.8; http://www.bioinformatics.babraham.ac.uk/proje-cts/fastqc/). High-quality clean reads were obtained by removing the adapters and low-quality reads from the raw data using Trimmomatic (version 0.35) (Bolger t al. 2014). The *C. cathayensis* cp genome was assembled using the SPAdes pipeline (Bankevich et al. 2012) with the *Cyclocarya paliurus* cp genome as the reference (NCBI accession: NC_034315).

Annotation of the C. cathayensis cp genome

C. cathayensis cp genome annotation was performed via the CpGAVAS pipeline (Liu et al. 2012). The annotated *C. cathayensis* genome was deposited to GenBank under accession number MN892516. The circular gene map was visualized in OGDRAWv1.2 (http://ogdraw.mpimp-golm.mpg.de/). Relative synonymous codon usage (RSCU) was determined by CodonW version 1.4.4 (http://codonw.sourceforge.net/).

Identification of repeats

REPuter (Kurt 2001; Kurtz and Schleiermacher 1999) was used to identify the repeat sequences (Liu et al. 2018) using the parameters reported by Li et al. (2019). Then, the online microsatellite identification tool (MISA, https://webblast.ipk-gatersleben.de/misa/) (Beier et al. 2017) was applied to predict cpSSRs with default parameters.

Phylogenetic Analysis

To determine the phylogenetic relationships among *Juglandaceae* species, a Bayesian inference (BI) tree was inferred using protocols suggested by Zou et al. (2015). An alignment of 17 cp genomic sequences (**Fig. 2**) was created using the MAFFT online version (Katoh et al. 2017; Kuraku et al. 2013) with default parameters.

Genomic comparison with related species

The online tool Irscope (Amiryousefi et al. 2018) was employed to draw the genetic architecture of the IR/SSC and IR/LSC junctions. mVISTA (Mayor et al. 2000) was used to compare the complete *C. cathayensis* cp genome to those of five related species including *C. kweichowensis*, *C. illioninensis*, *Cyclocarya paliurus*, *Juglans cathayensis*, and *Platycarya strobilacea*. The shuffle-LAGAN mode was used in mVISTA (Mayor et al. 2000), with the annotation of *Q. variabilis* as the reference. The sequences were initially aligned using MAFFT online version (Katoh et al. 2017; Kuraku et al. 2013), the pi value of each gene were calculated through alignment of each gene CDS sequences of different species using vcftools, and the ratios of non-synonymous (*Ka*) to synonymous (*Ks*) substitutions (*Ka/Ks*) in protein-coding genes were determined by the KaKs_Calculator.

Abbreviations

LSC: large single copy; SSC: small single copy; IRs: inverted repeats; SSR: simple sequence repeats; *Ks*: synonymous; *Ka*: nonsynonymous; pi: Genes nucleotide variability; RSCU: Relative synonymous codon usage.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable

Availability of data and material: The data used in our study has been submitted to NCBI Genbank (Accession number: MN892516). The related species and their Genbank accessions (Web: https://www.ncbi.nlm.nih.gov/) in this study are list as follow: *Betula nana* (KX703002), *Castanopsis concinna* (NC_033409), *C. echinocarpa* (NC_023801), *C. hainanensis*(NC_037389), *Castanea henryi* (NC_033881), *C. mollissima* (KY951992), *C. pumila* (KM360048), *C. seguinii* (NC_039749), *Dalbergia hainanensis* (NC_036961), *Fagus crenata* (NC_041252), *F. engleriana* (NC_036929), *F. sylvatica* (NC_041437), *Juglans major* (NC_035966), *J. hindsii* (NC_035965), *J. cinereal* (NC_035960), *J. nigra* (NC_035967), *J. cathayensis* (MF167457), *J. mandshurica* (MF167461), *J. sigillata* (MF167465), *J. hopeiensis* (NC_033894), *J. regia* (NC_028617), *Lithocarpus balansae* (NC_026577), *Malus prunifolia* (NC_031163), *Carya illinoinensis* (NC_041449), *C. kweichowensis* (NC_040864), *Cyclocarya paliurus* (NC_034315), *Platycarya strobilacea* (NC_035413), *Quercus acutissima* (NC_039429), *Q. aliena* (NC_026790), *Q. baronii* (NC_029490), *Q. chenii* (NC_039428), *Q. dentata* (NC_039725), *Q. dolicholepis* (KU240010), *Q. obovatifolia* (NC_039972), *Quercus rubra* (JX970937), *Q. sichourensis* (NC_036941), *Q. spinosa* (NC_026790), *Q. tarokoensis* (NC_036370), *Q. variabilis* (KU240009), *Trigonobalanus doichangensis* (NC_023959), and *Ulmus gaussenii* (NC_037840).

Competing interests: The authors declare that they have no competing interests.

Funding: The work was partly supported by the National Natural Science Foundation of China (31971641), National Key Research and Development Project (2019YFD1001504) and the Zhejiang Provincial Natural Science foundation of China (LY16C160011).

Authors' contributions: Conceptualization, ZXT and JSH; Methodology, All authors; Writing, ZXT and SHJ.

Acknowledgements: Not applicable

References

1. Lu A, Stone D, Grauke L (1999) *Juglandaceae*. Flora China 4: 277-285.
2. Zhang JB, Li RQ, Xiang XG, Manchester SR, Lin L, Wang W et al (2013) Integrated fossil and molecular data reveal the biogeographic diversification of the Eastern Asian-Eastern North American Disjunct Hickory genus (*Carya Nutt.*). PLoS One 8: e70449.
3. Naumann J, Symmank L, Samain MS, Kai FM, Wanke S (2011) Chasing the hare - evaluating the phylogenetic utility of a nuclear single copy gene region at and below species level within the species rich group *Peperomia* (*Piperaceae*). BMC Evol Biol 11: 357.
4. Raman G, Park V, Kwak M, Lee B, Park SJ (2017) Characterization of the complete chloroplast genome of *Arabis stellari* and comparisons with related species. PLoS One 12: e0183197.
5. Böhle UR ,Hilger H, Cerff R, Martin W (1994) Noncoding chloroplast DNA for plant molecular systematics at the infrageneric level. In: Schierwater B, Streit GP, Desalle R, editors. In molecular

ecology and evolution: approaches and applications. Birkhäuser: Basel 391-403.

6. Li X, Li Y, Zang M, Li M, Fang Y (2018) Complete chloroplast genome sequence and phylogenetic analysis of quercus acutissima. Int J Mol Sci 19: 1-17.
7. Li Y, Sylvester SP, Li M, Zhang C, Li X, Duan Y et al (2019) The complete plastid genome of *Magnolia zenii* and genetic comparison to *Magnoliaceae* species. Molecules 1-16.
8. Zhao J, Xu Y, Xi L, Yang J, Chen H, Zhang J (2018) Characterization of the chloroplast genome sequence of *Acer miaotaiense*: comparative and phylogenetic analyses. Molecules 23: 1740.
9. Zeng S, Zhou T, Han K, Yang Y, Zhao J, Liu ZL (2017) The complete chloroplast genome sequences of six *Rehmannia* species. Genes (Basel) 8: 103.
10. Xu C, Dong, W, Li W, Lu Y, Xie X, Jin X et al (2017) Comparative analysis of six *Lagerstroemia* complete chloroplast genomes. Front Plant Sci 8: 1-12.
11. Yang Y, Zhou T, Duan D, Yang J, Feng L, Zhao G (2016) Comparative analysis of the complete chloroplast genomes of five *Quercus* species. Front Plant Sci 7: 1-13.
12. Ye L, Fu C, Wang Y, Liu J, Gao L (2018) Characterization of the complete plastid genome of a Chinese endemic species *Carya kweichowensis*. Mitochondrial DNA Part B Resour 3: 492-493.
13. Zhai DC, Yao Q, Cao XF, Hao QQ, Ma MT, Pan J, Bai XH (2019) Complete chloroplast genome of the wild-type Hickory (*Carya cathayensis*) Mitochondrial DNA Part B 4: 1457-1458.
14. Zhang R, Peng F, Li Y (2015) Pecan production in China. Sci Hortic (Amsterdam) 197: 719-727.
15. Grauke LJ, Wood BW, Harris MK (2016) Crop vulnerability: *Carya*. Hortscience 51: 653-663.
16. Zhang B, Wang ZJ, Jin SH, Xia GH, Huang YJ, Huang JQ (2012) A pattern of unique embryogenesis occurring via apomixis in *Carya cathayensis*. Biol Plant 56: 620-627.
17. Huang Y, Xiao L, Zhang Z, Zhang R, Wang Z, Huang C et al (2019) The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. Gigaence 1-17.
18. Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 17: 134.
19. Hu Y, Chen X, Feng X, Woeste KE, Zhao P (2016) Characterization of the complete chloroplast genome of the endangered species *Carya sinensis* (*Juglandaceae*). Conserv Genet Resour 8: 467-470.
20. Morton BR (2003) The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. J Mol Evol 56: 616-629.
21. Liu HY, Yu Y, Deng YQ, Li J, Huang ZX, Zhou SD (2018) The chloroplast genome of *Lilium henrici*: genome structure and comparative analysis. Molecules 23: 1-13.
22. Jian HY, Zhang YH, Yan HJ, Qiu XQ, Wang QG, Li SB et al (2018) The complete chloroplast genome of a key ancestor of modern *Roses*, *Rosa chinensis* var. spontanea, and a comparison with congeneric species. Molecule 23: 1-13.
23. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S et al (2017) Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. Molecules 22: 1330.

24. Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9: 673-690.
25. Provan J, Powell W, Hollingsworth PM (2011) Chloroplast microsatellites: New tools for studies in plant ecology and evolution. Trends Ecol Evol 16: 142-147.
26. Diekmann K, Hodkinson TR, Barth S (2012) New chloroplast microsatellite markers suitable for assessing genetic diversity of *Lolium perenne* and other related grass species. Ann Bot 110: 1327-1339.
27. Singh N, Pal AK, Roy RK, Tamta S, Rana TS (2017) Development of cpSSR markers for analysis of genetic diversity in *Gladiolus* cultivars. Plant Gene 10: 31-36.
28. Hu JB, Li JW, Zhou XY (2009) Analysis of cytoplasmic variation in a cucumber germplasm collection using chloroplast microsatellite markers. Acta Physiol Plant 31: 1085-1089.
29. Deng Q, Zhang H, He Y, Wang T, Sun Y (2017) Chloroplast microsatellite markers for *Pseudotaxus chienii* developed from the whole chloroplast genome of *Taxus chinensis* var. Mairei (*Taxaceae*) . Appl Plant Sci 5: 1600153.
30. Pan L, Li Y, Guo R, Wu H, Hu Z, Chen C (2014) Development of 12 chloroplast microsatellite markers in *Vigna unguiculata* (*Fabaceae*) and amplification in *Phaseolus vulgaris*. Appl Plant Sci 2: 1300075.
31. Huang J, Yang X, Zhang C, Yin X, Liu S, Li X (2015) Development of chloroplast microsatellite markers and analysis of chloroplast diversity in Chinese *Jujube* (*Ziziphus jujuba* Mill.) and Wild *Jujube* (*Ziziphus acidojujuba* Mill.). PLoS One 10.
32. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE et al (2015) Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. Sci Rep5: 16958.
33. Drescher A, Stephanie R, Calsa T, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes.Plant J 22: 97-104.
34. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19: 11-15.
35. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30: 2114-2120.
36. Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov AS et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19: 455-477.
37. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X et al (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13: 715.
38. Kurt S (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633-4642.

39. Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426-427.
40. Liu L, Wang Y, He P, Li P, Lee J, Soltis DE et al (2018) Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia* (*Saxifragaceae*), using genome skimming data. BMC Genomics 19: 1-17.
41. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33: 2583-2585.
42. Zou LH, Huang JX, Zhang GQ, Liu ZJ, Zhuang XY (2015) A molecular phylogeny of *Aeridinae* (Orchidaceae: *Epidendroideae*) inferred from multiple nuclear and chloroplast regions. Mol Phylogenet Evol 85: 247-254.
43. Katoh K, Rozewicki J, Yamada KD (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 1-7.
44. Kuraku S, Zmasek CM, Nishimura O, Katoh K (2013) aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res 41: W22.
45. Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34: 3030-3031.
46. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA et al (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-1047.

Figures

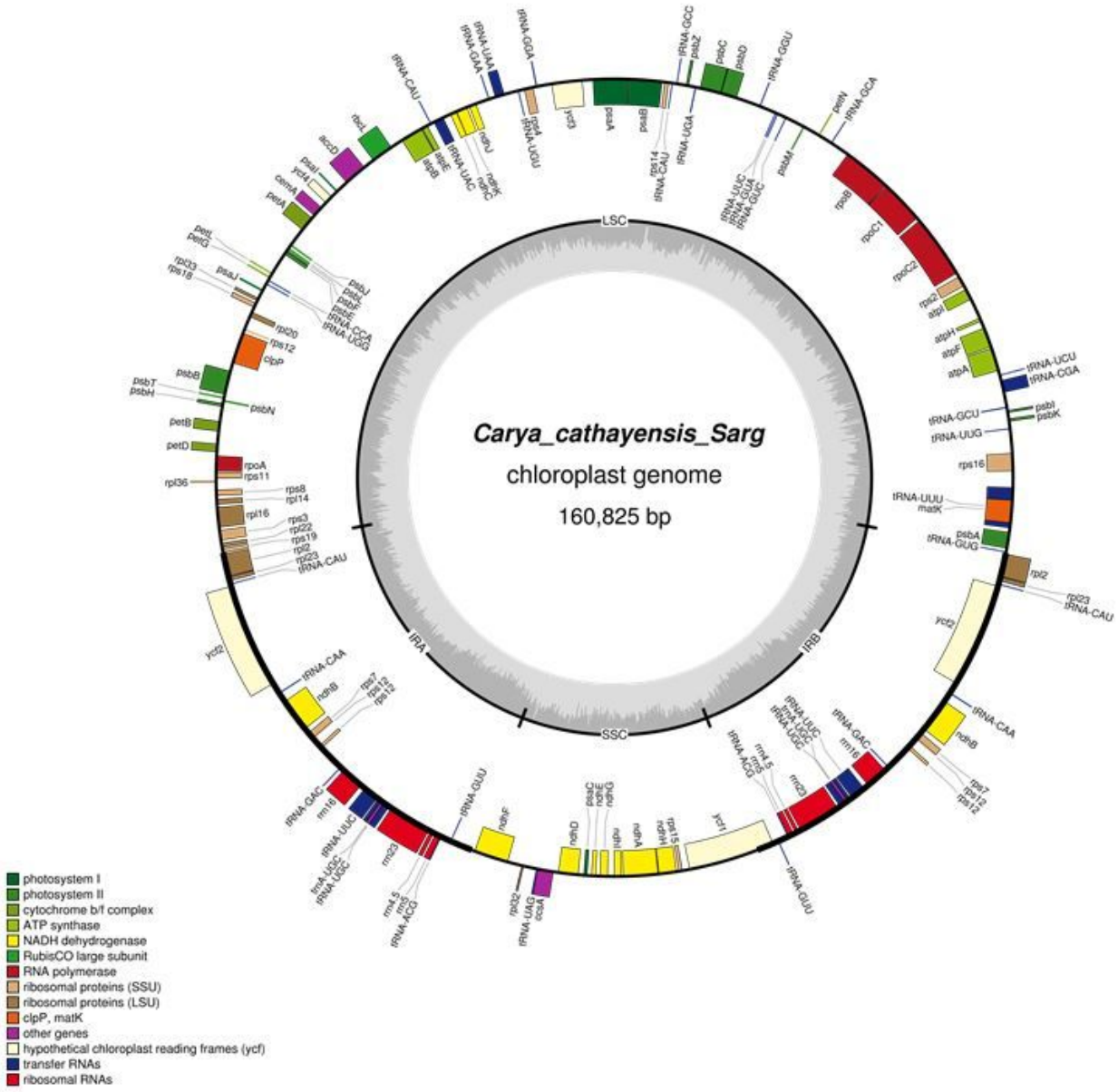

Figure 1

The complete Carya cathayensis chloroplast (cp) genome. Genes shown outside the outer circle are transcribed clockwise, whereas those shown inside are transcribed counterclockwise. The gray plots in the inner circle represent GC contents. The circular gene map was drawn using OGDRAWv1.2.

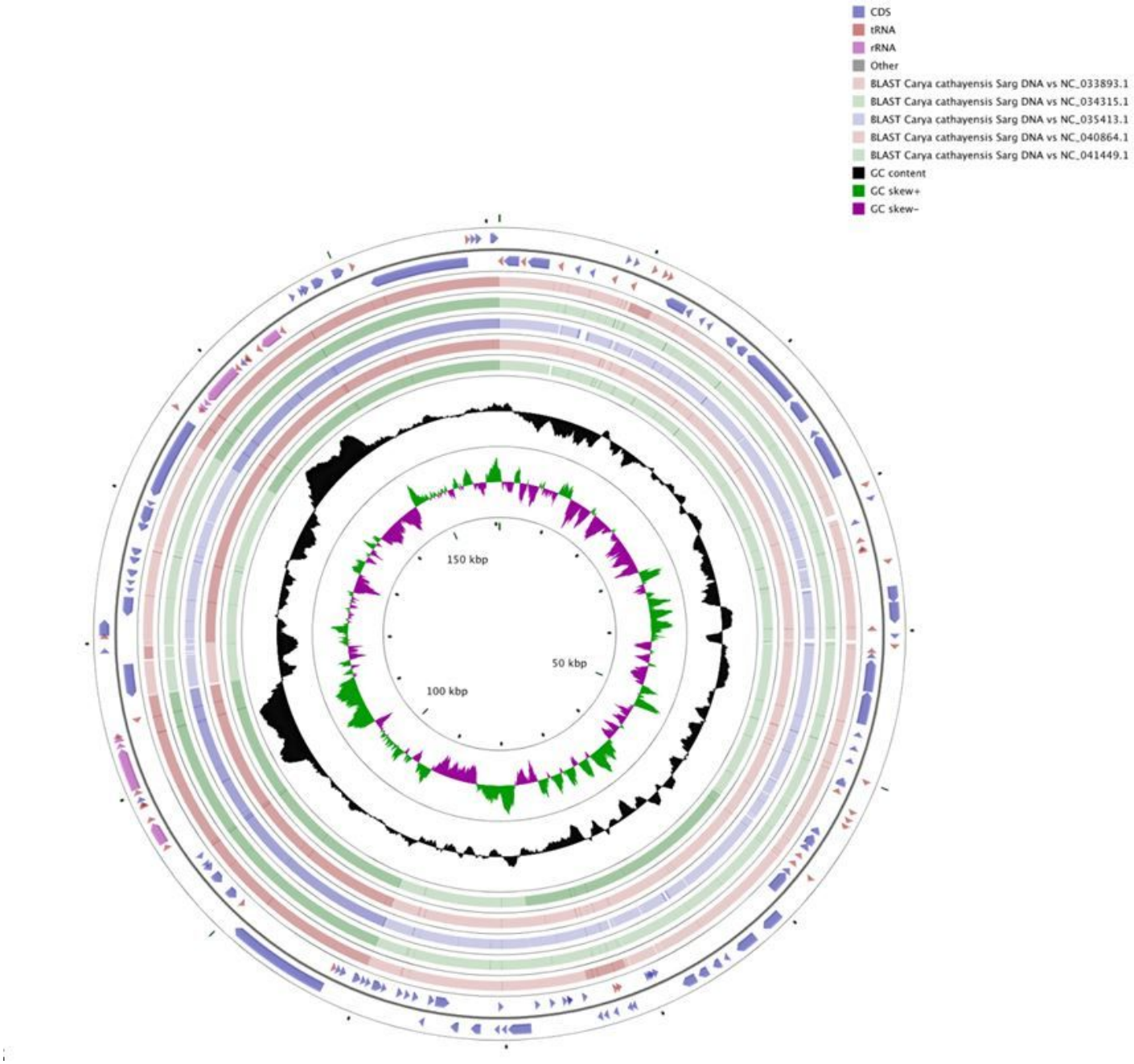

Figure 2

GC content of the C. cathayensis cp genome.

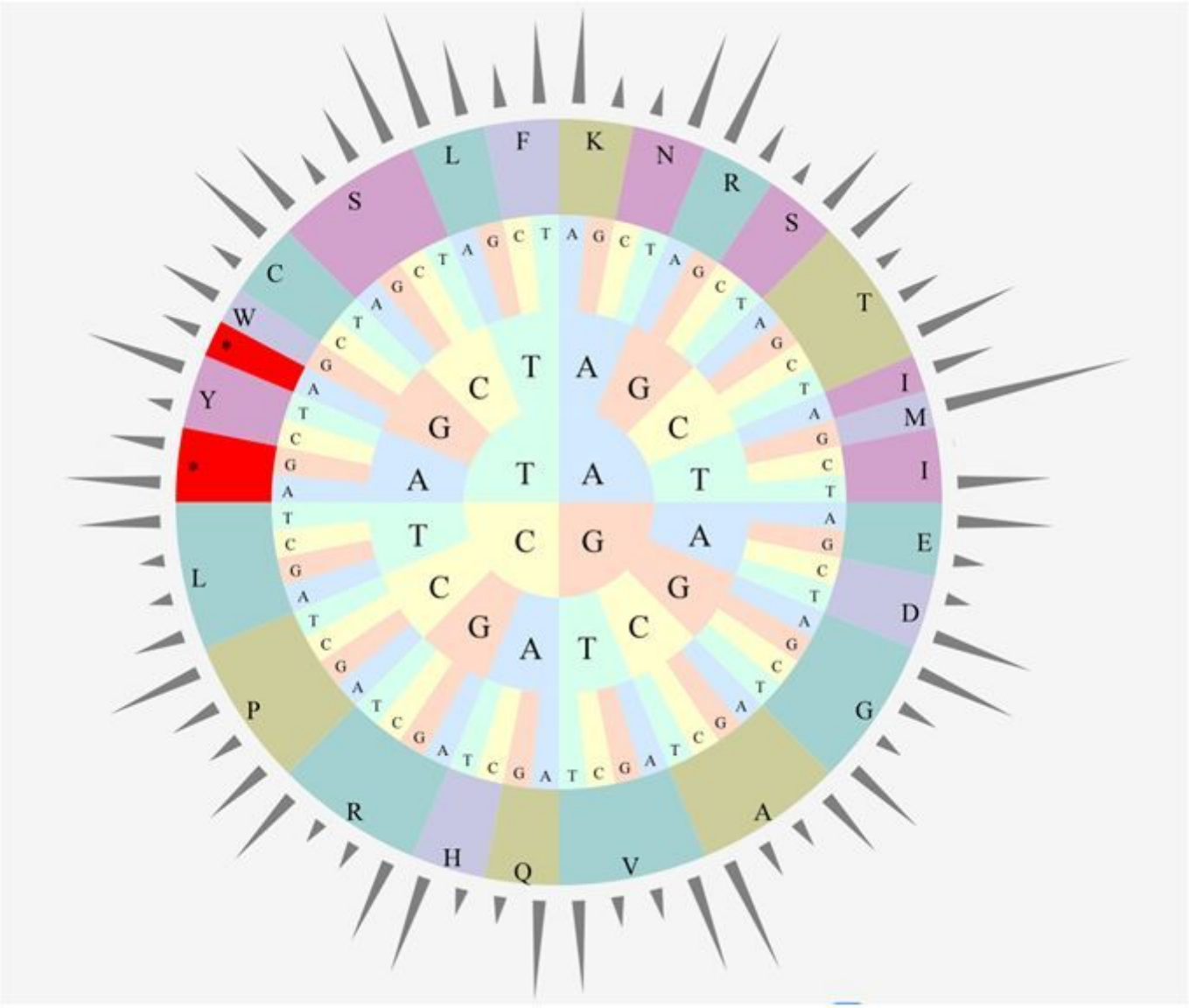

Figure 3

Codon usage frequency of the C. cathayensis cp genome

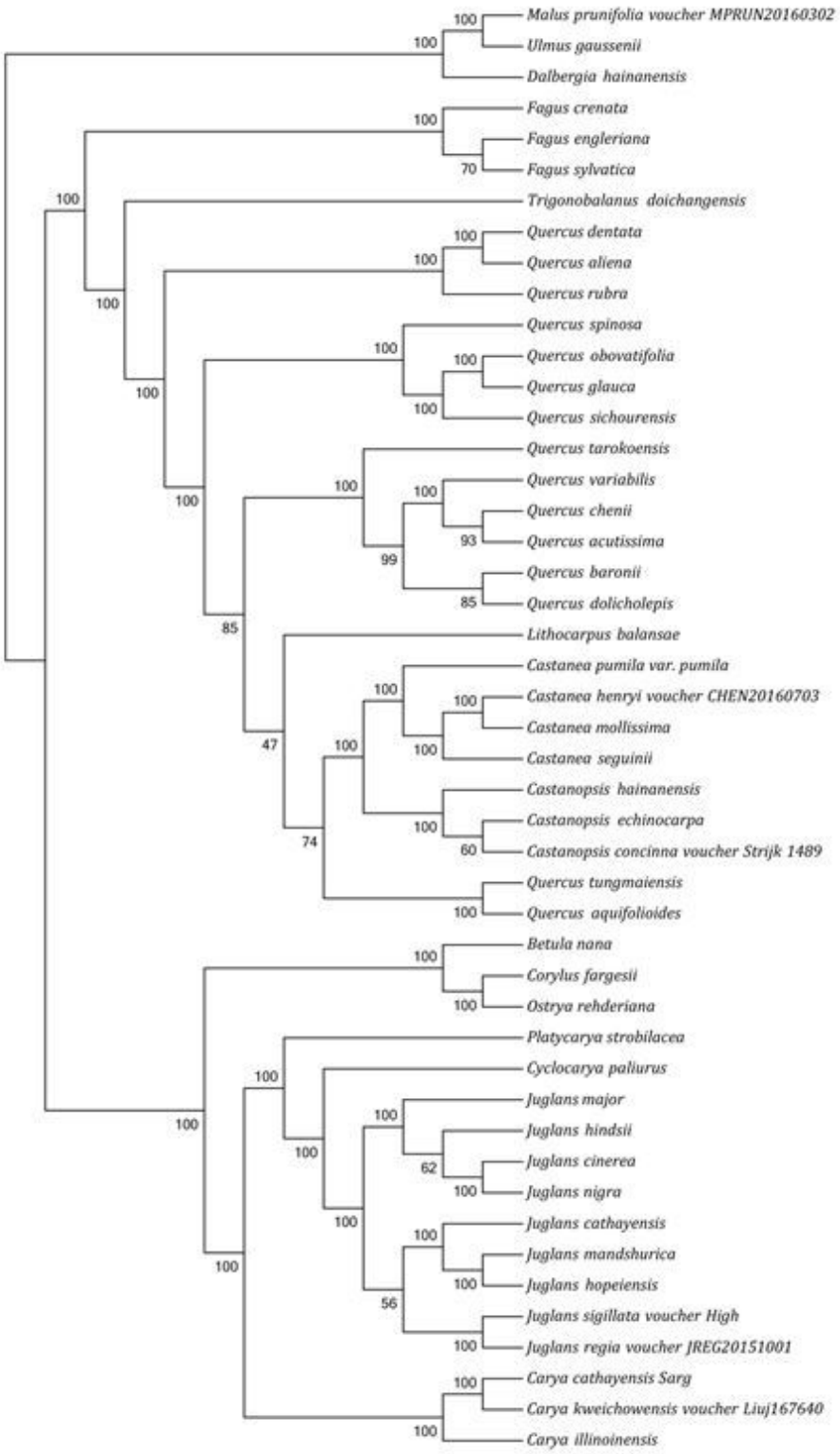

Figure 4

ML phylogenetic tree of 46 complete cp genomes resolved by Raxml. (Bootstrap values are shown near each node.)

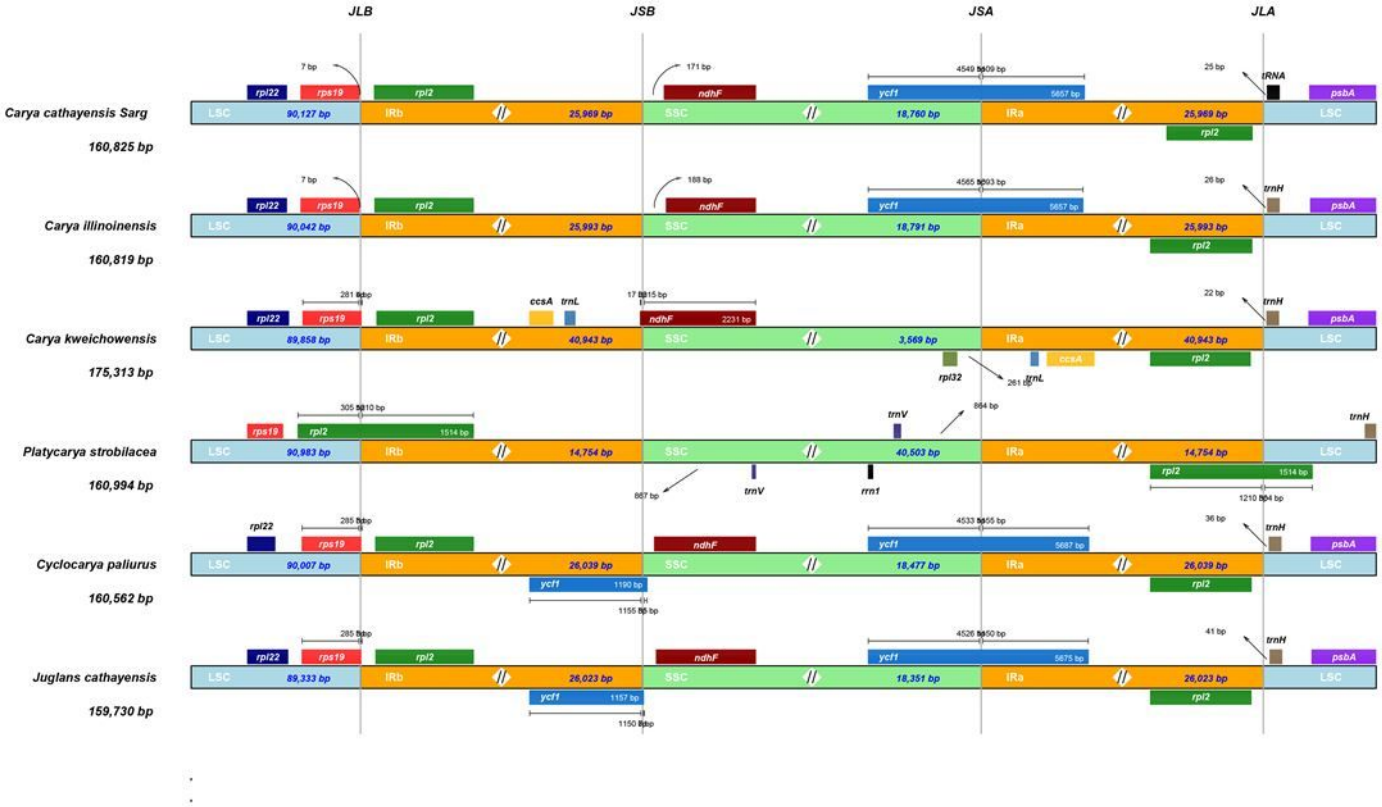

Figure 5

Comparison of the LSC, SSC, and IR regions among six selected cp genomes in family Juglandaceae. (Genes are denoted by colored boxes. The gaps between the genes and boundaries are proportional to the distances in bps.)

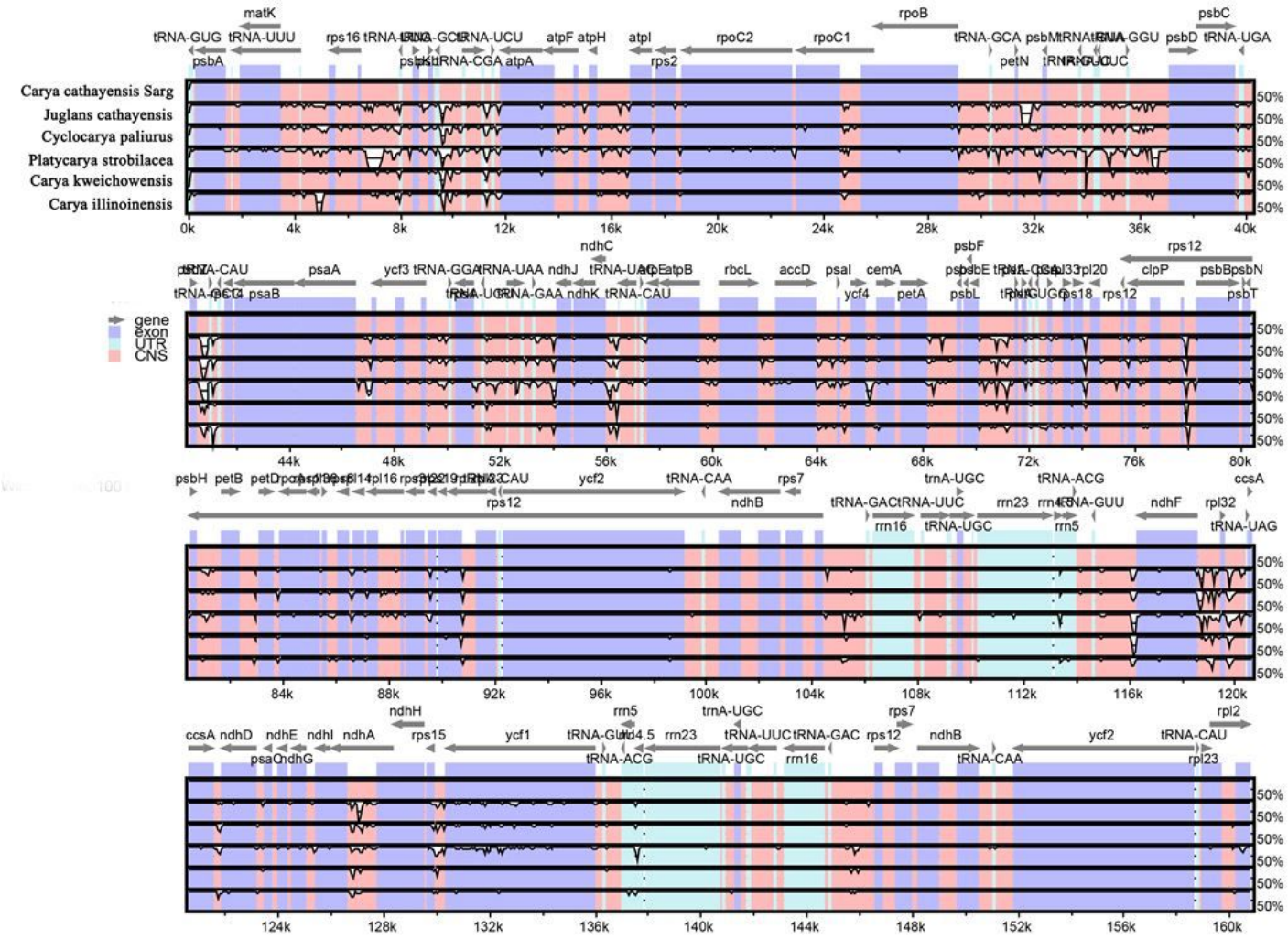

Figure 6

Variable characters in homologous regions among Carya cathayensis and five related species. (The homologous regions are oriented according to their locations in the cp genome. The grey arrows above the alignment indicate the gene orientations. The Y-axis shows the identity from 50% to 100%.)

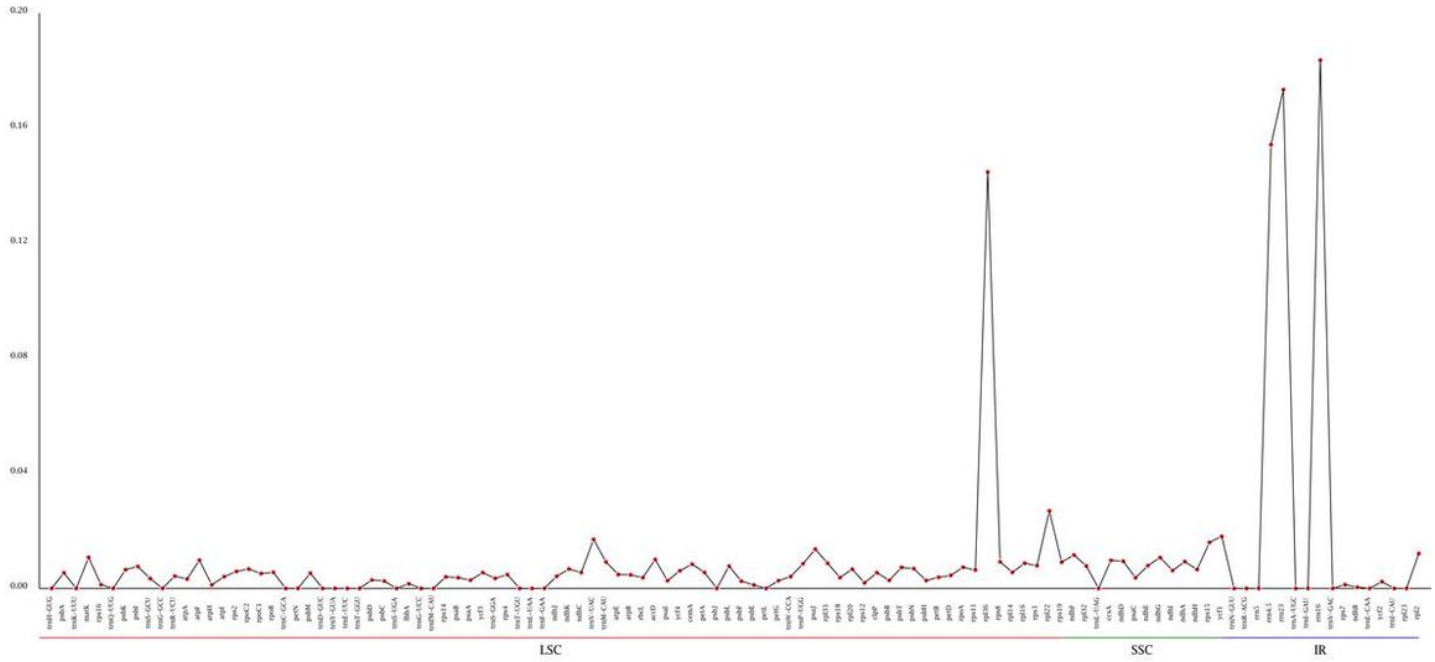

Figure 7

Genes nucleotide variability (pi) values of six Juglandaceae species. (The Y-axis shows Pi value; The X-axis shows Gene.)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xls