Fruit Research

ARTICLE

Open Access

https://doi.org/10.48130/frures-0024-0006

Fruit Research **2024**, 4: e012

Chloroplast genome sequencing of *Carya Illinoinensis* cv. Xinxuan-4, a new pecan pollinated cultivar

Yu Chen[1,2], Shijie Zhang[1], Wu Wang[1], Xinlin Chen[1,3], Yuqiang Zhao[1], Zhenghai Mo[1] and Cancan Zhu[1*]

[1] *Institute of Botany, Jiangsu Province and Chinese Academy of Sciences (Nanjing Botanical Garden Mem. Sun Yat-Sen), Nanjing 210014, China*
[2] *Jiangsu Key Laboratory for the Research and Utilization of Plant Resources, Nanjing 210014, China*
[3] *College of Forestry, Nanjing Forestry University, Nanjing 210037, China*
* Corresponding author, E-mail: zcc@cnbg.net

Abstract

Carya illinoinensis, is a highly valuable nut plant that is cultivated worldwide. As a precious pecan pollination resource, *C. illinoinensis* cv. Xinxuan-4 is protandrous, with a very early bud break in China. In this study, the chloroplast (cp) genome of 'Xinxuan-4' was sequenced and compared with closely related cultivars. The cp genome was found to be 160,819 bp in length, and it had a common quadripartite architecture with one large single copy (LSC; 90,022 bp), one small single copy (SSC; 18,791 bp), and two inverted repeats (IRs; 26,003 bp). The genome contained 132 genes, including 87 protein-coding genes, 37 tRNA genes, and eight rRNA genes, with a GC content of 36.1%. Furthermore, 278 simple sequence repeats and 59 long repeat sequences were identified, and the genome comparisons revealed that there was a greater divergence in the noncoding regions than in the coding regions. According to the gene selective pressure analysis, five genes (*petD, rpl16*, *rps12*, *rpoC2,* and *rpoC1*) were identified to be potentially under positive selection when contrasted with the other *Carya* genotypes. Phylogenetic analysis of the cp genome of 'Xinxuan-4' and 17 other species inferred that *Carya* is monophyletic and that the genetic relationship between 'Xinxuan-4' and 'Pawnee' is quite close from an evolutionary perspective. The currently characterized cp genome of 'Xinxuan-4' offers useful data for subsequent research on this pecan species.

Citation: Chen Y, Zhang S, Wang W, Chen X, Zhao Y, et al. 2024. Chloroplast genome sequencing of *Carya Illinoinensis* cv. Xinxuan-4, a new pecan pollinated cultivar. *Fruit Research* 4: e012 https://doi.org/10.48130/frures-0024-0006

Introduction

The chloroplast (cp) is involved in a variety of biological processes in the cells of plants, including photosynthesis, carbon fixation, and stress responses[1,2]. The genome of cp is smaller (75–250 kb) than the nuclear genome, and the genome sequences are more easily obtained *via* new sequencing technology; furthermore, the influence from homologous locations is lower[3]. The cp genome is a popular method used to identify differences among species, due to its short sequence length and fairly simple analysis. Its genome genes, *ndhF*, *matK*, and *trnS-trnG*, have been widely amplified for species recognition, barcoding, and phylogeny[4]. In angiosperms, a large single-copy (LSC, 80–90 kb), a small single-copy (SSC, 16–27 kb), and two copies of inverted repeats (IRa/b, 20–28 kb) make up the typical quadripartite structure of cp genomes[5,6]. Previous research verified that the gene sequence, gene information, and genome structure of the cp genomes were extremely stable in plants[7]. Moreover, the cp genome has some specific characteristics, such as uniparental inheritance, natural haploid, and a minimum number of recombination, which have assisted in understanding the phylogeny and evolution of numerous genera, such as *Oncidiinae*[8], *Camellia*[9], and *Fritillaria*[10].

C. illinoinensis, commonly known as pecan, belongs to the family *Juglandaceae*, which is located in Asia and North America's tropical and temperate zones[11]. In China, pecan is a well-known nut crop that has been grown extensively in recent years[12,13]. *C. illinoinensis* is a monoecious, dichogamous, and wind-pollinated species[14]. The timing of its pollination is crucial, as the stigma surface of the pistil only receives pollen during a relatively short period[15]. The pecan cultivar 'Pawnee' is protandrous, meaning that the pollen is shed before the pistil is receptive. It was introduced to China in 1998[16], and is the only early pollination tree; this leads to a serious pollination deficiency in the orchards of China[17]. The 'Xinxuan-4' is also protandrous, and was selected from an autochthonous individual tree growing in Nanjing Botanical Garden Men. Sun Yat-sen in the 1950s[18,19]. The maturation of its anthers in male flowers occurs two days earlier than that of the 'Pawnee', which could satisfy the early pollination of pecan trees. However, the genetic information of 'Xinxuan-4' remains exclusive.

Recently, the cp genomes of *C. illinoinensis* cv. Pawnee[20], *C. illinoinensis* cv. Wichita[21], *C. illinoinensis* cv. 87MX3-2.11, and *C. illinoinensis* cv. Lakota[22] were identified. The release of more cp genomes will help identify genetic variations, and offer new perspectives on the interspecific relationships among the *Carya* species. This research sequenced the 'Xinxuan-4' cp genome and the first comparative analysis of its sequence with other published *Carya* cp genomes.

Our main objectives in this study were to: (1) ascertain the cp genome's structure and composition; (2) carry out an analysis of codon preference; (3) detect repeats and microsatellite patterns; (4) determine highly divergent regions; (5) define the phylogenetic analysis. The research will unveil the maternal origin of the 'Xinxuan-4' by cp sequencing and will contribute to the future genetic breeding of pecan.

Materials and methods

Plant material and DNA extraction

The pecan cultivar 'Xinxuan-4' is protandrous, with a very early-season pollen shed; it was planted in Nanjing Botanical Garden Men. Sun Yat-sen, Nanjing City, Jiangsu Province, China. The 'Xinxuan-4' should be a good pollenizer for the 'Mahan', 'Wichita', and 'Mohawk' varieties, which have plenty of flowers, red stigma, and small fruit (Fig. 1). Fresh leaves of 'Xinxuan-4' were obtained and rapidly stored at −80 °C. The adjusted CTAB protocol was adopted to extract DNA[23].

Cp genome sequencing

The pecan cultivar 'Xinxuan-4' was used for the cp genome sequencing. After sequencing, the adapters of the raw data were removed and the low-quality reads were cleaned by fastp v20.0. The clean reads were obtained to assemble the 'Xinxuan-4' cp genome using SPAdes 3.11.0 software[24]. The assembly contigs were blasted to the *Carya laciniosa* cp genome, and the gaps were repaired using GapCloser 1.12 software.

Genome annotation

Prodigal v2.6.3 software was used to annotate the cp genome, and Hammer v3.1 b2 software was utilized to scan the tRNA genes. The rRNA genes were identified using Aragorn v1.2.38. Organellar Genome DRAW v1.3.1[25] was used to generate the map. The cp genomic sequence of the 'Xinxuan-4' was uploaded to GenBank with accession number PRJNA795859.

Analysis of microsatellites

A simple sequence repeat (SSR) marker is a kind of tandem repeat sequence made up of a dozen nucleotides, which has several repeat units (usually 1 to 6). CpSSR markers are SSR markers present in the genomes of cps. CpSSR analysis was performed using the Misa software[26]. The parameters used were as follows: mono-nucleotides repeated eight times; di-nucleotides repeated five times; trinucleotides repeated four times, tetra-, penta-, and hexa-nucleotides repeated three times.

Fig. 1 *Carya illinoinensis* cv. Xinxuan-4. (a) Male flower. (b) Female flower. (c) Fruits on the tree. (d) Fruit without husk, scale bar = 2 cm.

Repeat sequence and synonymous codon usage analysis

REPuter software was employed to examine repeat structures, containing forward (F), reverse (R), complement (C), and palindromic (P) repeats[27]. The tandem repeats finder 4.07b was used to search for tandem repeats. Using CodonW1.4.4, the synonymous codon usage was characterized using relative synonymous codon usage (RSCU).

Cp comparison and phylogenetic analysis

The sequence information of five *Carya* genotypes, including *C. illinoinensis* cv. Xinxuan-4 (PRJNA795859), *C. illinoinensis* cv. Pawnee (MN9771241), *C. illinoinensis* cv. 87MX3-2.11 (MH909600), *C. illinoinensis* cv. Lakota (MH909599), and *C. cathayensis* (PE00820836)[28], were obtained from the Gene Bank for the comparative analysis. The cp genome of 'Xinxuan-4' was compared to those of four chosen *Carya* materials using the mVISTA program[29]. The nucleotide variability (Pi) in the whole cp genome was evaluated by DnaSp v6[30]. To analyze the cp genome difference between 'Xinxuan-4' and its close relatives, the relative rates of synonymous (Ks) and non-synonymous (Ka) substitution rates were determined using the Ka/Ks Calculator software[31]. The sequence data of 17 *Juglandaceae* species were retrieved from the NCBI to examine the evolutionary relations among these species. The PhyML program (v3.0)[32] was adopted to produce the phylogenetic tree.

Results and discussion

Genome features of *C. Illinoinensis* cv. Xinxuan-4

The cp genome of 'Xinxuan-4' was 160,819 bp in length (Fig. 2), which was the same as that of 'Pawnee'[20] and 'Lakota' (Table 1)[22]. *Carya* species showed similar cp genome sizes (Fig. 3), and the cp genome of 'Wichita' was the shortest genome published thus far[21]. Similarly to most angiosperms, the cp genome presented a common quadripartite architecture consisting of LSC (90,022 bp), SSC (18,791bp), and IRa/b, (each 26,003 bp). Comparatively, the lengths of LSC, SSC, and IR of 'Xinxuan-4' were the same as that of 'Pawnee', whereas the 'Wichita' had the shortest LSC (89,799 bp), SSC (18,751bp), and IR (25,991bp). The GC content is an important indicator of affinity in various species. The 'Xinxuan-4' cp genome had a total GC content of 36.1%, which means it was exactly like other species in the genus *Carya* (35.8%–36.3%)[33–35]. Specifically, the LSC, SSC, and IR locations had GC contents of 33.74%, 29.89%, and 42.58%, respectively. In the IR locations, the high GC content may be attributed to the high GC content in four rRNA genes (*rrn16*, 56.41%; *rrn23*, 55.64%; *rrn4.5*, 47.97%; *rrn5*, 52.07%) in this region (Fig. 3)[36].

The cp genome of 'Xinxuan-4' contained 132 genes, comprising 87 protein-coding genes, 37 tRNA genes, and eight rRNA genes (Table 1). Nineteen duplicate genes were discovered: eight were protein-coding genes (*rpl2*, *rpl23*, *rps7*, *rps12*, *yct1*, *ycf15*, *ycf2*, and *ndhB*), seven were tRNA (*trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*), and four were rRNAs genes (*rrn16*, *rrn23*, *rrn4.5*, and *rrn5*) (Table 2). In the 'Xinxuan-4' cp genome, there were 18 intron-containing genes, of which 15 genes (nine protein-coding genes and six tRNA genes) contained one intron, and three genes (*rps12*, *ycf3,* and

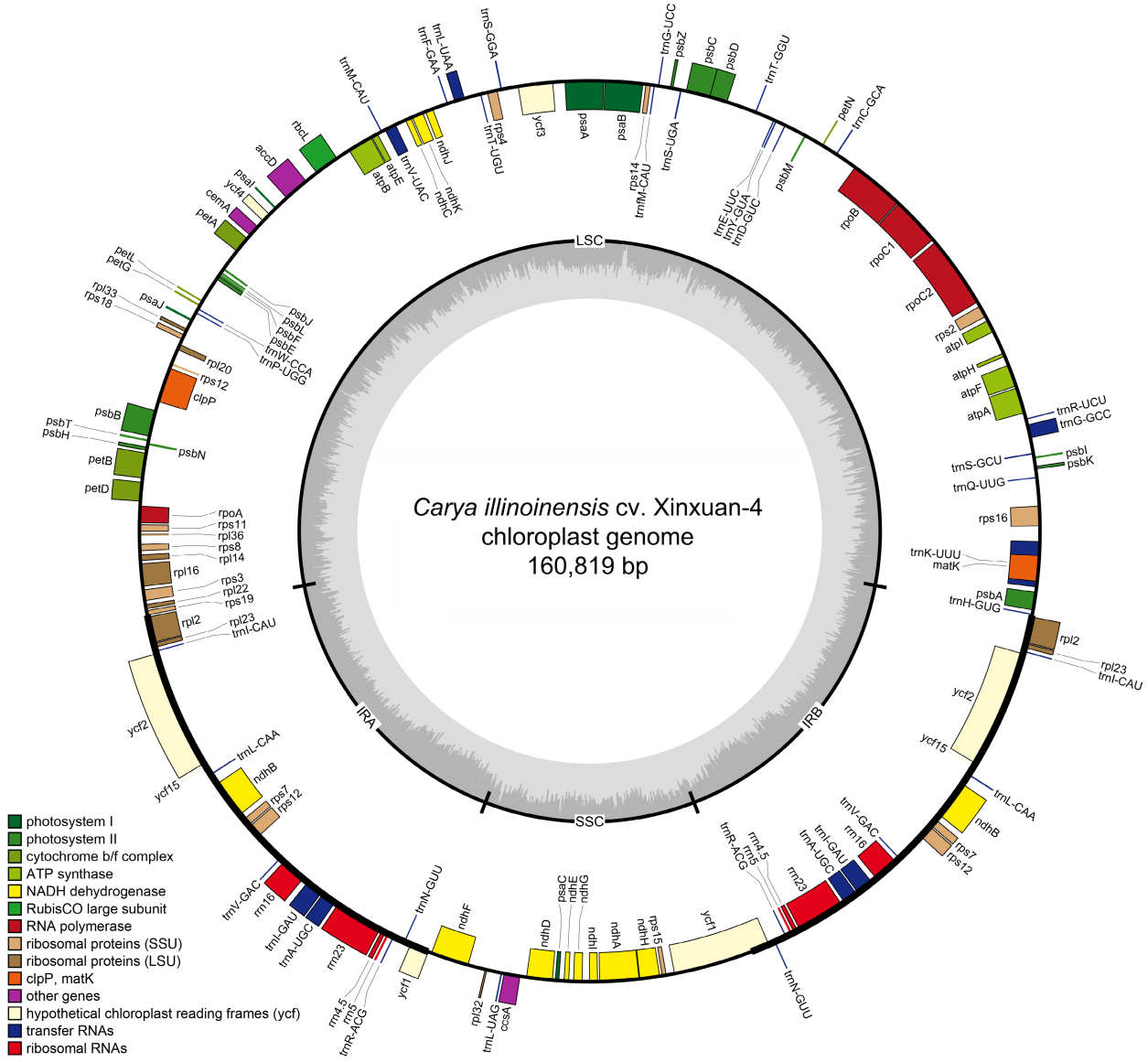

Fig. 2 Circular representation of the cp genome of *C. illinoinensis* cv. Xinxuan-4.

Table 1. Features of the cp genomes of *C. illinoinensis* cv. Xinxuan-4 and five related materials.

Species	Xinxuan-4	Pawnee	87MX3-2.11	Lakota	Wichita	Cathayensis
Genome size (bp)	160,819	160,819	160,545	160,819	160,532	160,825
LSC size (bp)	90,022	90,022	89,933	90,041	89,799	90,115
SSC size (bp)	18,791	18,791	18,576	18,790	18,751	18,760
IR size (bp)	26,003	26,003	26,018	25,994	25,991	25,975
Total genes	132	131	124	123	128	129
Protein-coding genes	87(8)	86(7)	84(6)	83(6)	83(6)	84
tRNAs	37(7)	37(8)	32(6)	32(6)	37(8)	37(7)
rRNAs	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)
GC content (%)	36.1	36.1	36.2	36.1	36.2	36.1

clpP1) possessed two introns (Table 2 & Supplemental Table S1). With a length of 2,559 bp, the intron of the *trnK-UUU* gene was the longest. The *rps12* gene, consisting of one intron in the LSC location and the other two exons in the IR locations, was a trans-spliced gene[37,38].

Genes can be gained or eliminated in the cp genomes during the process of evolution[37,39]. Comparing the cp genome of *Carya* 'Pawnee', '87MX3-2.11', 'Lakota', and '*C. cathayensis*' in GenBank, the gene *rps12* was absent in '*Lakota*', but existed in the other cp genomes, with one duplicate in the '87MX3-2.11' and two duplicates in the remaining species (Table 3 & Supplemental Table S2). Previous results verified that the 'Lakota' cp genome's *rps12* gene was missing a reading frame[22]. All *C. illinoinensis* contained the genes of *psbZ* and *ycf15*, but could not

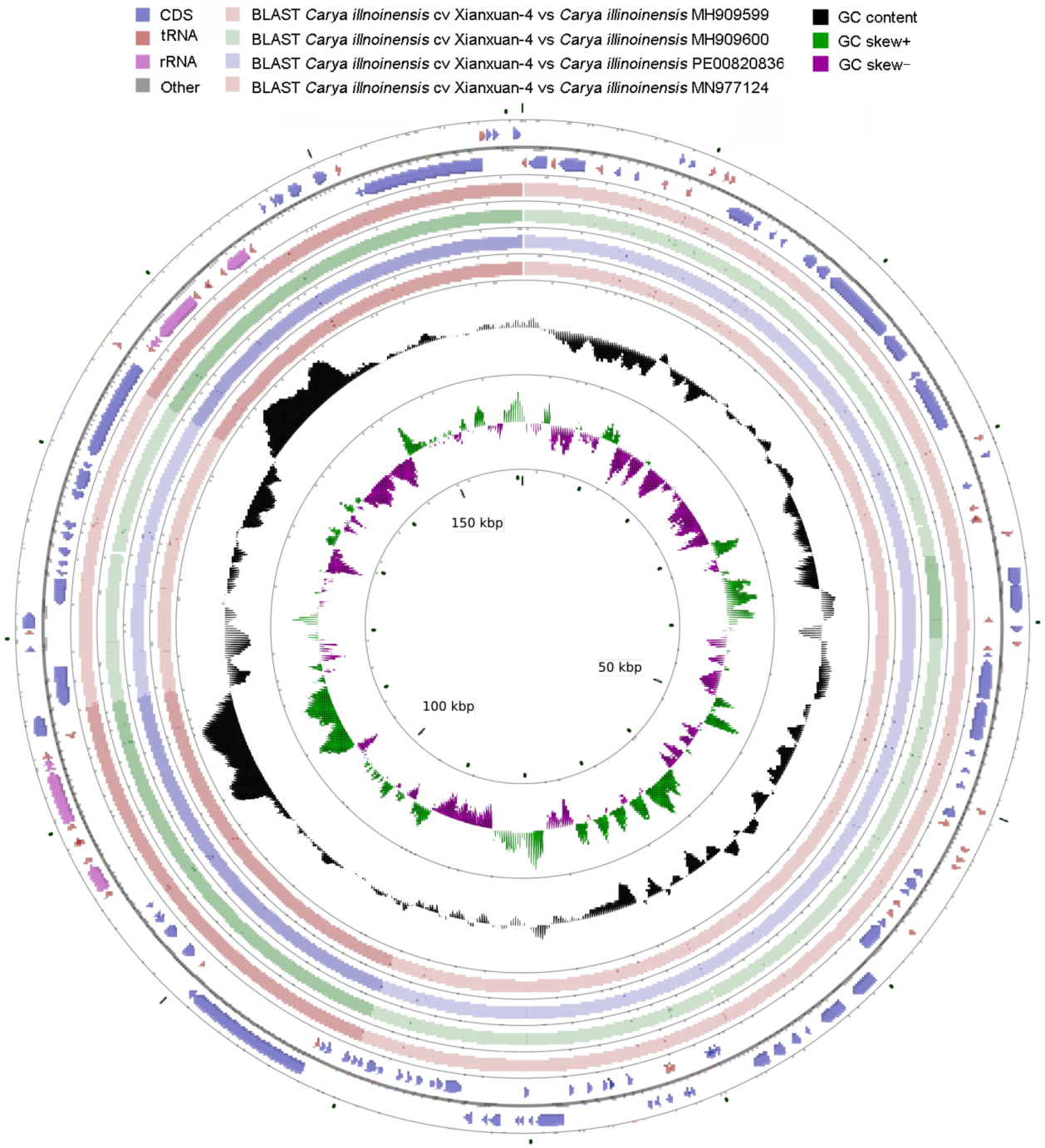

Fig. 3 GC content of the *C. illinoinensis* cv. Xinxuan-4 cp genome.

be found in *C. cathayensis*. Gene *rps16* existed in the other three *Carya* genotypes ('Xinxuan-4', 'Pawnee' and '*C. cathayensis*') with one copy; however, it was absent from cp genomes of the 'Lakota' and '87MX3-2.11'. The *rps16* gene encoded in the cp genome of most organisms was also not discovered in *Populus alba*[40]. The loss of this gene in the cp genomes was compensated by the *rps16* gene from mitochondria[41].

The most notable differences in *Carya* were in the tRNA genes, six tRNA genes (*trnA-UGC*, *trnG-UCC*, *trnI-GAU*, *trnL-UAA*, *trnV-UAC*, and *rna-UGC*) were found to be different (Table 3). All of the *C. illinoinensis* lacked duplicates of *rna-UGC*, but two duplicates were discovered in *C. cathayensis*. Both pecans varieties, '87MX3-2.11' and 'Lakota', lost copies of *trnG-UCC*, *trnI-GAU*, *trnL-UAA*, *trnV-UAC*, and *rna-UGC*, and a single duplicate of *trnG-UCC* was detected only in 'Xinxuan-4'. Approximately 110–130 individual genes were found in the majority of cp genomes, most of which were coding DNA sequence (CDS) or protein-coding genes; the rest of the genes were tRNA and rRNA genes[42]. The 'Xinxuan-4' and 'Pawnee' cp genomes included 132 and 131 genes, respectively, compared with 123 ('87MX3-2.11') or 128 ('Wichita') genes in the *Carya*. The protein-coding genes and rRNA genes remained stable among

Table 2. Annotated genes in the cp genome of *C. illinoinensis* cv. Xinxuan-4.

Category	Gene group	
Photosynthesis	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Subunits of NADH	ndhA[b], ndhB[ab], ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Cytochrome b/f complex	petA, petB[b], petD[b], petG, petL, petN
	Subunits of ATP synthase	atpA, atpB, atpE, atpF[b], atpH, atpI
	Large subunit of rubisco	rbcL
Self-replication	Large ribosomal subunit	rpl14, rpl16[b], rpl2[ac], rpl20, rpl22, rpl23[a], rpl32, rpl33, rpl36
	Small ribosomal subunit	rps11, rps12[ab], rps14, rps15, rps16[b], rps18, rps19, rps2, rps3, rps4, rps7[a], rps8
	Subunits of RNA polymerase	rpoA, rpoB, rpoC1[b] , rpoC2
	Ribosomal RNAs	rrn16[a], rrn23[a], rrn4.5[a], rrn5[a]
	Transfer RNAs	trnA-UGC[ab], trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC[b], trnG-UCC, trnH-GUG, trnI-CAU[a], trnI-GAU[ab], trnK-UUU[b], trnL-CAA[a], trnL-UAA[b], trnL-UAG, trnM-CAU, trnN-GUU[a], trnP-UGG, trnQ-UUG, trnR-ACG[a], trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC[a], trnV-UAC[b], trnW-CCA, trnY-GUA, trnfM-CAU
Other genes	Maturase	matK
	Protease	clpP[c]
	Envelope membrane protein	cemA
	Acetyl-CoA carboxylase	accD
	c-type cytochrome synthesis gene	ccsA
Unknown function genes	Conserved open reading frames	ycf1[a], ycf15[a], ycf2[a], ycf3[c], ycf4

[a] Gene with two copies, [b] Gene with one intron, [c] Gene with two introns.

the five *Carya* genotypes (77–83 CDS and four rRNA genes). In general, the number of tRNA genes was the most variable among the annotations in the *Carya* cp genomes.

Codon preference analysis

Due to the degeneracy of codons, each amino acid is encoded by multiple codons (synonymous codons) in the organisms[43]. The utilization rate of genome codons differs greatly in various species; this inequality in the use of synonymous codons is known as RSCU. Natural selection in the organisms is believed to have generated the RSCU, which can be categorized into four models: no preference (RSCU < 1.0), low preference (1.0 < RSCU < 1.2), moderate preference (1.2 > RSCU < 1.3), and high preference (RSCU > 1.3)[44]. The RSCU of the 'Xinxuan-4' cp encoding sequence was calculated. The findings demonstrated that 26,643 codons encoded all of the genes and that 20 amino acids were encoded by 68 different kinds of codons. The most commonly utilized codons were ATT (isoleucine), AAA, (lysine), and GAA (glutamic acid), which were 1,159 (4.35%), 1,071 (4.02%), and 1,049 (3.94%) codons, respectively (Fig. 4 & Supplemental Table S3). In the 'Xinxuan-4' cp genome, 31 out of 68 codons had RSCU values > 1, of which 22 displayed a strong preference, six had a median preference and three showed a low preference. With the exception of TTG being G-ending, all of the codons in the 'Pawnee' cp genome showed a preference for an A/T ending[20]. In this study, we also detected that the codons in the 'Xinxuan-4' cp genome preferred A/ T ending. Similarly, A/T ending has been found in *C. cathayensis* and other angiosperms[45,46]. This preference of the codons may be due to the high conservation in the cp genes. In contrast, numerous codons ending in G or C showed RSCU values below 1, suggesting that these condons were less frequent in *Carya* cp genes. The specific species characteristics of synonymous codon usage could be utilized to study the regulation of gene expression, differentiation, and evolutionary processes of *Carya* in the future.

Table 3. Different genes and gene copies in the cp genomes among five *Carya* materials.

Gene group	Different gene	Xinxuan 4	Pawnee	87MX3-2.11	Lakota	*Cathayensis*
Protein-coding genes	rps12	2	2	1	0	2
	rps16	1	1	0	0	1
	psbZ	1	1	1	1	0
	lhbA	0	0	0	0	1
	ycf15	2	2	2	2	0
tRNA genes	trnA-UGC	2	2	2	2	0
	trnG-UCC	1	0	0	0	0
	trnI-GAU	2	2	0	0	0
	trnL-UAA	1	1	0	0	1
	trnV-UAC	1	1	0	0	1
	rna-UGC	0	0	0	0	2

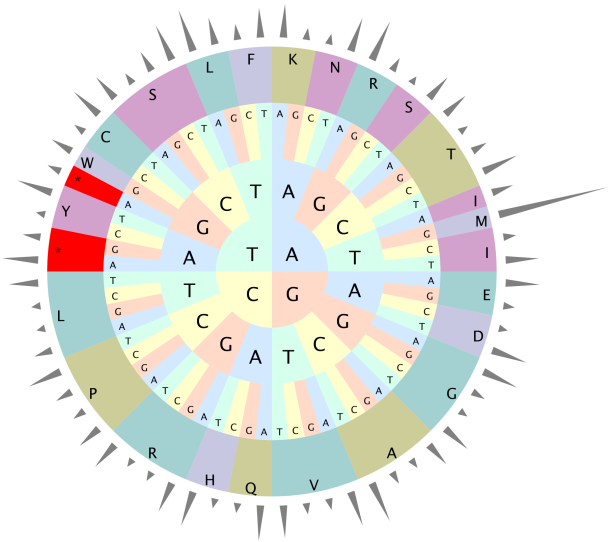

Fig. 4 Codon usage frequency of the *C. illinoinensis* cv. Xinxuan-4 cp genome.

Repeat sequence and CpSSR analysis

Repeat sequences are thought to be crucial for the rearrangement and recombination of the cp genome[47,48]. In the present study, a total of 59 long repeats, comprising 32 forward, 23 palindromic, three reverse, and one complementary repeat, were found in the cp genome (Fig. 5 & Supplemental Table S4). The lengths for the majority of the repeats varied from 30 to 90 bp. There were 46, 7, and 12 long repeats located in the LSC, SSC, and IR locations, respectively. Most of the repeats occurred in the intergenic spacers (IGS), and the remaining repeats were located in the coding regions associated with the protein-coding genes *ycf3*, *ycf2*, *rrn4.5s ndhA*, *psaB*, *psaA*, and *tRNA* genes *trns-UGA*, and *trns-GCU*. These conclusions were consistent with the findings for *C. illinoinensis* and *C. cathayensis*[20,28].

Short DNA sequences, or SSRs, exhibit high polymorphism in related species. CpSSR markers are a wonderful tool for studying interspecific evolution and identification, as well as intraspecific population genetic variation[49,50]. According to a prior study, 213 SSRs were examined in the cp genome of *C. illinoinensis*[20]; in the present study, a total of 278 SSRs were detected in the 'Xinxuan-4' cp genome using MISA. There were six different types of SSR, the most prominent of which were mononucleotide repeats (189, 67.98%), followed by trinucleotide (70, 25.18%) and di-nucleotide (15, 5.39%) repeats. The other three kinds of SSRs were less prevalent: tetranucleotide (2, 0.72%), pentanucleotide (1, 0.36%), and hexanucleotide (1, 0.36%) repeats (Fig. 6a). These results were also reported in *C. illinoinensis* [20] and *C. cathayensis* [28], in which mono- and tri-nucleotide type cpSSRs were found at high rates, while at low frequencies were di-, tetra-, penta-, and hexanucleotide type cpSSRs. Moreover, the majority of cpSSRs belonged to the A/T types (65.11%), while only eight C/G types (2.87%) cpSSRs were identified in the cp genome, indicating that short A or T repeats made up the largest number of cpSSRs in the 'Xinxuan-4' cp genome (Fig. 6b & Supplemental Table S5). These results confirmed the hypothesis of G or C repeats that were uncommon in cpSSRs, which consisted of short A or T repeats[28].

In this study, we also evaluated the distribution of 278 cpSSRs in the cp genome, which were predicted to be 191 (68.71%), 45 (16.18%), and 42 (15.11%) in the LSC, SSC, and IR locations, respectively (Fig. 6c). In addition, 167, 66, and 45 SSRs existed in the intergenic regions, introns, and coding sequences, respectively (Fig. 6d). The noncoding regions in the cp genome of the 'Xinxuan-4' contained the majority of the cpSSRs, similar distribution preferences of cpSSRs have been observed in *C. illinoinensis*, *C. cathayensis*, and *Avena sativa*[20,28,51]. The majority of genes had mono- or trinucleotide SSRs, whereas only one protein-coding gene, *ycf1*, contained tetranucleotide SSR (Supplemental Table S5). Therefore, the specific SSR to the 'Xinxuan-4' in different gene regions could be used as a molecular marker to choose an appropriate cultivar for early pollination breeding material and the management of the pure line.

IR contraction and expansion

In the cp genomes of angiosperms, the expansion and contraction of the IR and SSC boundaries are frequently observed, giving rise to size differences among cp genomes[52,53]. To further explore the structural characteristics of the cp genome of 'Xinxuan-4', we examined the IR/SSC and IR/LSC junctions using four different *Carya* materials, namely 'Pawnee', '87MX3-2.11', 'Lakota', and *C. cathayensis*. The results are demonstrated in Fig. 7. Our results showed that the 'Xinxuan-4', 'Pawnee', and 'Lakota' had the same size of chloroplast genomes (160,819 bp), while *C. cathayensis* had the largest cp genomes (160,825 bp). All five materials had IR, SSC, and LSC regions of similar sizes, and their IRb boundaries all reached into the *ycf1* gene, with lengths ranging from 1,093 bp (*C. illinoinensis*) to 1,109 bp (*C. cathayensis*). Correspondingly, the 'Xinxuan-4', 'Pawnee', and 'Lakota' varieties had the same IR sequence (26,003 bp), which was highly conserved. However, '87MX3-2.11' contained a slightly longer IR region (26,030 bp), and *C. cathayensis* possessed a smaller IR location (25,975 bp). It was suggested that the IRa/b region of the '87MX3-2.11' had experienced expansion and *C. cathayensis* underwent contraction during evolution. In angiosperm plastomes, the SSC/IR border is

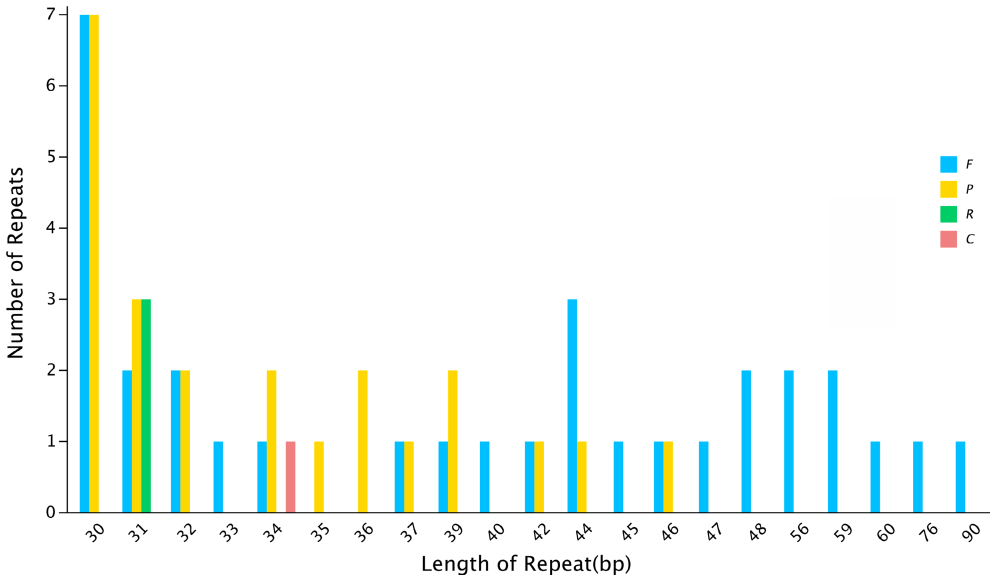

Fig. 5 Size and type of the long repeats located in the *C. illinoinensis* cv. Xinxuan-4 cp genome.

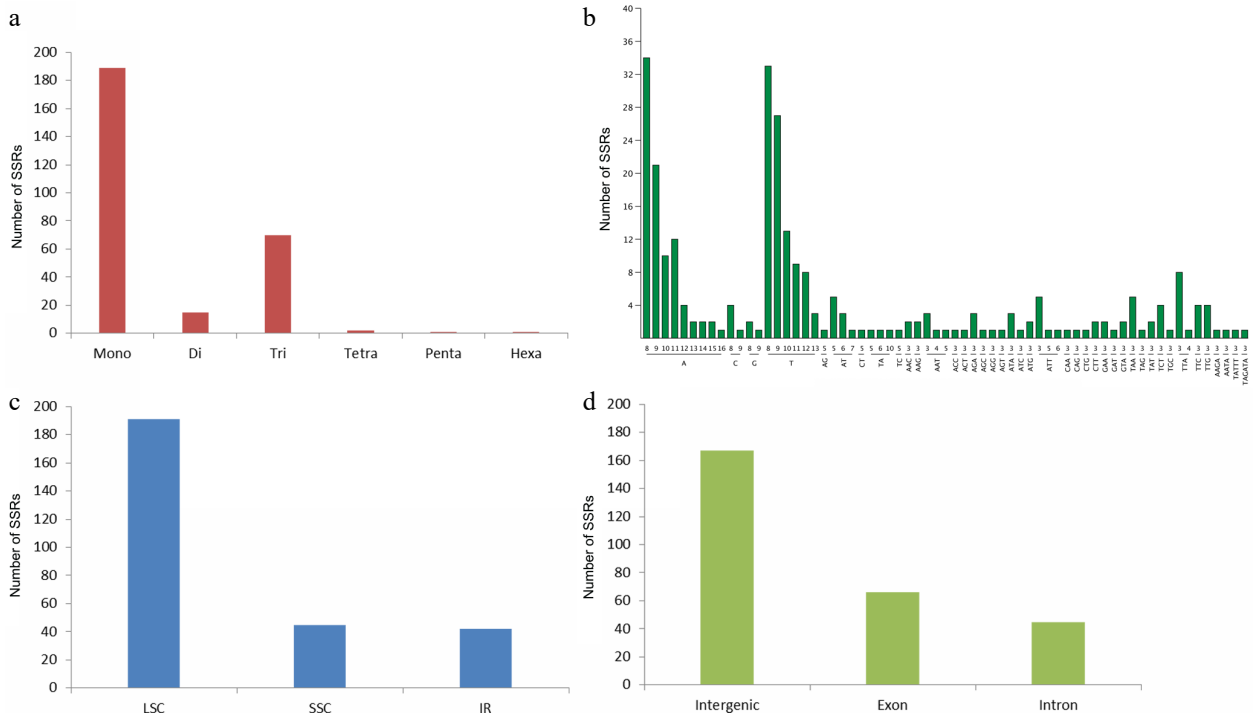

Fig. 6 SSRs distribution in the cp genomes of the *C. illinoinensis* cv. Xinxuan-4. (a) Type and number of SSRs. (b) Type and number of SSR repeats. (c) Number of SSRs in various location. (d) Number of SSRs in three locations.

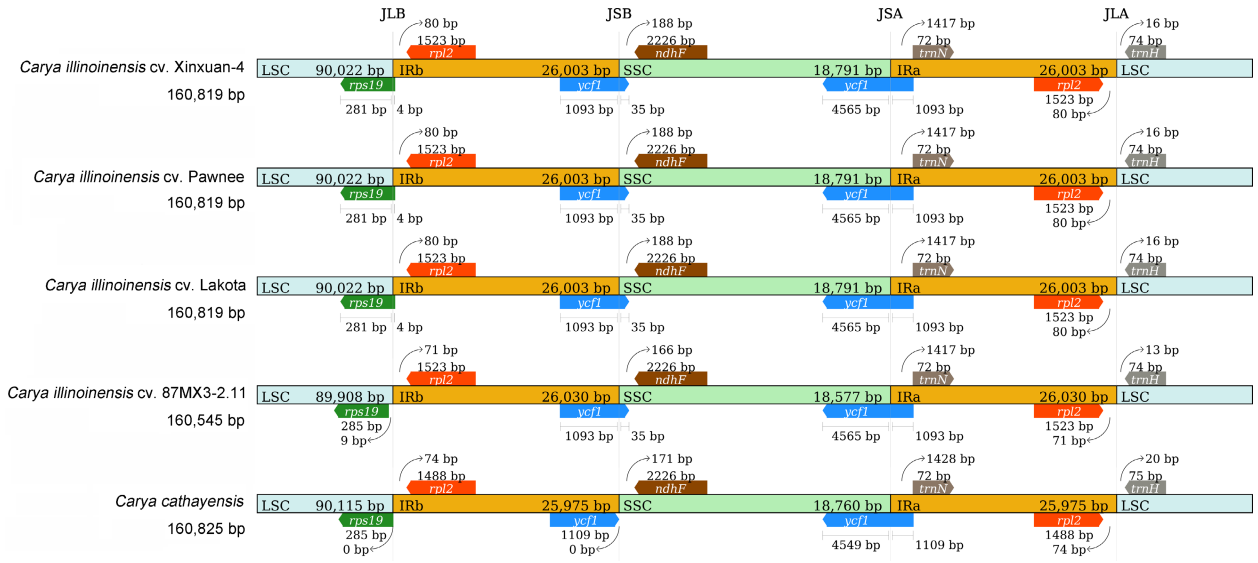

Fig. 7 Comparing the LSC, SSC, and IR locations of five selected cp genomes in the *Carya*.

relatively conserved and mostly located within *ycf1*[54]. Similar expansions or contractions have been published in *Jasminum nudiflorum* Lindl[55] and *Avena sativa*[51].

Comparative analysis of cp genome structure

Further investigation of the variations in the cp sequences was conducted with the five *Carya* genera using mVISTA, with 'Xinuan-4' as a reference. The findings showed that there were extremely high sequence similarities among the cp genome sequences of 'Xinuan-4', 'Pawnee', and 'Lakota', as demonstrated in Fig. 8. Compared to the encoding regions, the non-coding locations showed a comparatively higher level of divergence. Some notable divergences in the non-coding locations included the following: *trnS-GCU–trnG-GCC*, *trnR-UCU–atpA*, *atpF–atpH*, *trnD-GUC–trnE-UUC*, *trnT-GGU–psbD*, *trnG-UCC–trnfM-CAU*, *ndhC-trnV-UAC*, *accD–pasI*, *rpl32–trnL-UAG*, and *ndhG–ndhI*. Genes such as *matK*, *rpoC2*, and *ycf1* were discovered to contain variation coding genes. These findings were consistent with those reported for the related family *Juglandacea*, the genus *Quercus*, in the *Fagaceae* family[56].

In order to test whether the cp genes of *C. illinoinensis* underwent selection, the Ka/Ks were computed to identify the variations among genes. The results showed that the Ka/Ks values of most genes were NA, and only 14 genes had the values (Supplemental Table S6). Most of the (seven of 12) genes had

Fig. 8 Sequence identity plot comparing the cp genomes among *Carya* with *C. illinoinensis* cv. Xinxuan-4 set as a reference.

values below 1 in the cp genomes of *C. illinoinensis*, suggesting these kinds of genes were the target of purifying selection. The remaining five genes, which were *petD*, *rpl16*, *rpoC2*, *rpoC1*, and *rps12*, had Ka/Ks ratios that were generally greater than 1, meaning positive selection in comparison to the other *C. illinoinensis* species.

Moreover, the gene nucleotide variability value (pi) can offer prospective molecular markers for genetics in population applications and show variations in nucleic acid sequences of various species[51]. In this study, the Pi values of five *Carya* genotypes, including 'Xinxuan-4', 'Pawnee', '87MX3-2.11', 'Lakota', and C. cathayensis, are shown in Fig. 7. The figure demonstrated that the nucleotide diversity of the SSC and LSC locations was significantly higher than that of the IR locations (Fig. 9 & Supplemental Table S7). Gene nucleotide variability values of LSC. rps12, psbL, petD, and IR. trnV-GAC were higher genes. The remaining genes' values were less than 0.003, which suggested that the *Carya* species had a low nucleotide diversity.

Phylogenetic analysis

The encoding sequences from the cp genomes of 18 species were used to create a phylogenetic tree. As shown in Fig. 10, the phylogenetic tree indicated that the genera *Carya* and *Juglandaceae* were both monophyletic and that *Carya* proved more related to the group formed by the genus *Juglandaceae*, which was in line with earlier research[33,57]. Interestingly, *C. cathayensis* was grouped with *C. sinensis* instead of *C. illinoinensis*. Previous research demonstrated that *C. cathayensis* belonged to one of the typical species of the Asian sect. Sinocarya, while *C. illinoinensis* represented one of the typical species of the North American sect. Apocarya[20]. This could be the cause of the separated between the groups of *C. illinoinensis* and *C. cathayensis*. In addition, the 'Xinxuan-4' and 'Pawnee' varieties formed a single clade, suggesting that the 'Xinxuan-4' variety was more closely related to 'Pawnee', which inferred that 'Xinxuan-4' seedling may have come from North America.

Conclusions

In this study, we published the cp genome sequence of 'Xinxuan-4', which came from a seedling selection. The genome showed similar features to those of 'Pawnee'. The genome was estimated to contain 112 unique genes, comprising 79 protein coding genes, 29 tRNAs, and four rRNAs. We identified 278 cpSSRs and 59 long repeats that can be applied as prospective

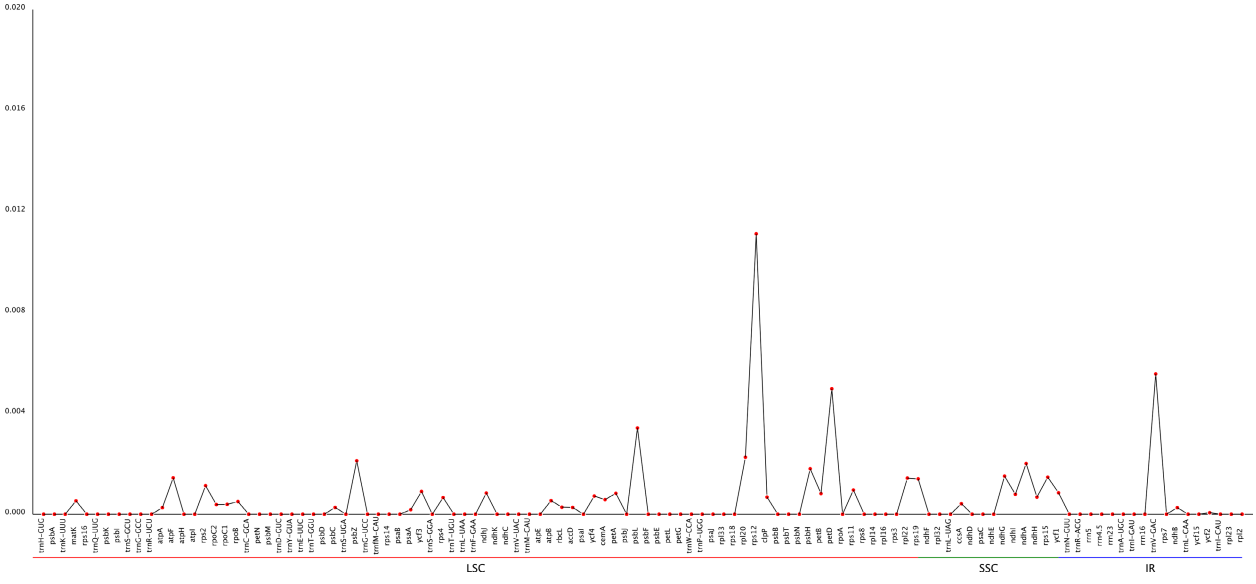

Fig. 9 Sliding window analysis of the cp genome for nucleotide diversity (pi) of three species in *Carya*.

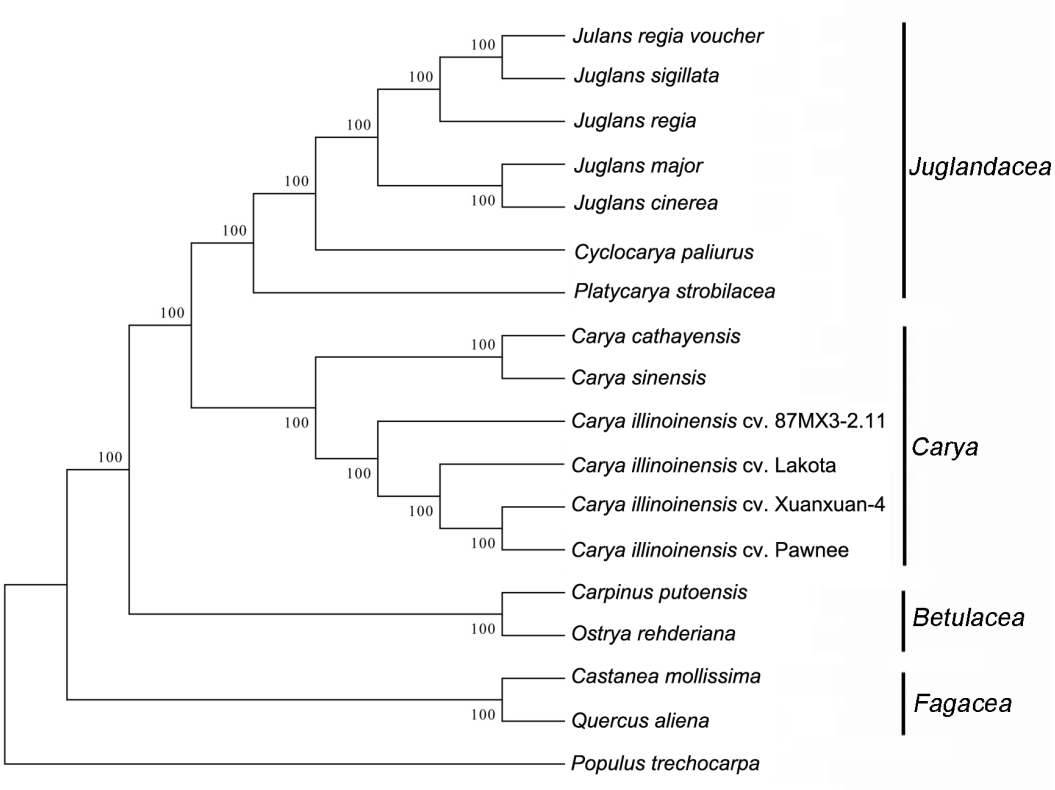

Fig. 10 Phylogenetic tree of 18 related species based on the whole cp genome.

molecular markers for genetics of the population and evolutionary studies. The RSCU analysis showed that all of the genes were encoded by 26,643 codons, and 68 kinds of codons encoded 20 amino acids. We also detected that the codons in the genome preferred A/T endings. Phylogenetic analysis showed that the 'Xinxuan-4' variety was more closely related to 'Pawnee'. These results not only offer a useful database to recognize the maternal origin of the 'Xinxuan-4' cultivars, but also contribute to the analysis of the phylogenetic relationship and application of germplasm resources for *Carya* species.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Chen Y; data collection: Zhang S, Chen Y; software, visualization: Wang W, Mo Z; draft manuscript preparation: Chen Y, Chen X; review and editing: Zhao Y, Zhu C. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are openly available in the GenBank of NCBI at www.ncbi.nlm.nih.gov (accessed on 10 September 2022), reference number (PRJNA 795859).

Acknowledgments

This research was supported by the National Natural Science Foundation of China (32001344), the Natural Science Foundation of Jiangsu Province, China (BK20200290, BK20210166), Key Research and Development Plan of Jiangsu Province (BE2021406).

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information accompanies this paper at (https://www.maxapress.com/article/doi/10.48130/frures-0024-0006)

Dates

Received 27 June 2023; Accepted 17 November 2023; Published online 7 March 2024

References

1. Howe CJ, Barbrook AC, Koumandou VL, Nisbet RER, Symington HA, et al. 2003. Evolution of the chloroplast genome. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358:99–107
2. Daniell H, Lin CS, Yu M, Chang WJ. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* 17:134
3. Li D, Zhao C, Liu X. 2019. Complete chloroplast genome sequences of *Kaempferia Galanga* and *Kaempferia Elegans*: molecular structures and comparative analysis. *Molecules* 24:474
4. Qin M, Zhu C, Yang J, Vatanparast M, Schley R, et al. 2022. Comparative analysis of complete plastid genome reveals powerful barcode regions for identifying wood of *Dalbergia odorifera* and *D. tonkinensis* (Leguminosae). *Journal of Systematics and Evolution* 60:73–84
5. Shetty SM, Md Shah MU, Makale K, Mohd-Yusuf Y, Khalid N, et al. 2016. Complete chloroplast genome sequence of *Musa balbisiana*

corroborates structural heterogeneity of inverted repeats in wild progenitors of cultivated bananas and plantains. *The Plant Genome* 9:plantgenome2015.09.0089
6. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, et al. 2006. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution* 23:2175–90
7. Yang J, Tang M, Li H, Zhang Z, Li D. 2013. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology* 13:84
8. Wu FH, Chan MT, Liao DC, Hsu CT, Lee YW, et al. 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biology* 10:68
9. Huang H, Shi C, Liu Y, Mao S, Gao L. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology* 14:151
10. Bi Y, Zhang M, Xue J, Dong R, Du Y, et al. 2018. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on *Fritillaria*. *Scientific Reports* 8:1184
11. Mo Z, Feng G, Su W, Liu Z, Peng F. 2018. Transcriptomic analysis provides insights into grafting union development in pecan (*Carya illinoinensis*). *Genes* 9:71
12. Chen Y, Wang M, Zhu C, Zhao Y, Wang B, et al. 2018. Field investigation of resistance against black spot of different pecan varieties in Jintan, Changzhou. *Journal of Jiangsu Forestry Science & Technology* 45:26–29
13. Wu J, Lin H, Meng C, Jiang P, Fu W. 2014. Effects of intercropping grasses on soil organic carbon and microbial community functional diversity under Chinese hickory (*Carya cathayensis* Sarg.) stands. *Soil Research* 52:575–83
14. Manos PS, Stone DE. 2001. Evolution, phylogeny, and systematics of the Juglandaceae. *Annals of the Missouri Botanical Garden* 88:231–69
15. Thompson TE, Romberg LD. 1985. Inheritance of heterodichogamy in pecan. *Journal of Heredity* 76:456–58
16. Zhang R, Peng F, Li Y. 2015. Pecan production in China. *Scientia Horticulturae* 197:719–27
17. Mo Z, Zhang J, Zhai M, Xuan J, Jia X, et al. 2013. Observation and comparison of flowering phenology of *Carya illinoensis* in Nanjing. *Journal of Plant Resources and Environment* 22:57–62
18. Zhang R, Lv F, Zhang X, He F, Wang L. 2005. Feasibility study for extension of pecan cultivars introduced from America. *Economic Forest Researches* 23:1–10
19. Chen Y, Zhang S, Zhao Y, Mo Z, Wang W, et al. 2022. Transcriptomic analysis to unravel potential pathways and genes involved in pe can (*Carya illinoinensis*) resistance to *Pestalotiopsis microspora*. *International Journal of Molecular Sciences* 23:11621
20. Mo Z, Lou W, Chen Y, Jia X, Zhai M, et al. 2020. The chloroplast genome of *Carya illinoinensis*: genome structure, adaptive evolution, and phylogenetic analysis. *Forests* 11:207
21. Feng G, Mo Z, Peng F. 2020. The complete chloroplast genome sequence of *Carya illinoinensis* cv. wichita and its phylogenetic analysis. *Mitochondrial DNA Part B* 5:2235–36
22. Wang X, Rhein HS, Jenkins J, Schmutz J, Grimwood J, et al. 2020. Chloroplast genome sequences of Carya illinoinensis from two distinct geographic populations. *Tree Genetics & Genomes* 16:48
23. Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19:11–15
24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19:455–77
25. Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* 52:267–74
26. Beier S, Thiel T, Münch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33:2583–85
27. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* 29:4633–42
28. Shen J, Li X, Chen X, Huang X, Jin S. 2022. The complete chloroplast genome of *Carya cathayensis* and phylogenetic analysis. *Genes* 13:369
29. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32:W273–W279
30. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, et al. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution* 34:3299–302
31. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics* 8:77–80
32. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307–21
33. 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* 3:492–93
34. Zhai D, Yao Q, Cao X, Hao Q, Ma M, et al. 2019. Complete chloroplast genome of the wild-type Hickory *Carya cathayensis*. *Mitochondrial DNA Part B* 4:1457–58
35. Hu Y, Chen X, Feng X, Woeste KE, Zhao P. 2016. Characterization of the complete chloroplast genome of the endangered species *Carya sinensis* (Juglandaceae). *Conservation Genetics Resources* 8:467–70
36. Biju VC, Shidhi PR, Vijayan S, Rajan VS, Sasi A, et al. 2019. The complete chloroplast genome of *Trichopus zeylanicus*, and phylogenetic analysis with *Dioscoreales*. *The Plant Genome* 12:190032
37. Liu X, Zhu G, Li D, Wang X. 2019. Complete chloroplast genome sequence and phylogenetic analysis of *Spathiphyllum* 'Parrish'. *PLoS ONE* 14:e0224038
38. Wang W, Yu H, Wang J, Lei W, Gao J, et al. 2017. The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *International Journal of Molecular Sciences* 18:2288
39. Dong W, Xu C, Li W, Xie X, Lu Y, et al. 2017. Phylogenetic resolution in *Juglans* based on complete chloroplast genomes and nuclear DNA sequences. *Frontiers in Plant Science* 8:1148
40. Okumura S, Sawada M, Park YW, Hayashi T, Shimamura M, et al. 2006. Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts. *Transgenic Research* 15:637–46
41. Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura SI, et al. 2008. Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. *Molecular Biology and Evolution* 25:1566–75
42. Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H. 2011. Complete plastid genome sequences of three Rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. *Molecular Biology and Evolution* 28:835–47
43. Wald N, Alroy M, Botzman M, Margalit H. 2012. Codon usage bias in prokaryotic pyrimidine-ending codons is associated with the degeneracy of the encoded amino acids. *Nucleic Acids Research* 40:7074–83
44. Zuo L, Shang A, Zhang S, Yu X, Ren Y, et al. 2017. The first complete chloroplast genome sequences of *Ulmus* species by *de novo* sequencing: genome comparative and taxonomic position analysis. *PLoS ONE* 12:e0171264

45. Li Y, Sylvester SP, Li M, Zhang C, Li X, et al. 2019. The complete plastid genome of *Magnolia zenii* and genetic comparison to Magnoliaceae species. *Molecules* 24:261
46. Liu H, Yu Y, Deng Y, Li J, Huang Z, et al. 2018. The chloroplast genome of *Lilium henrici*: genome structure and comparative analysis. *Molecules* 23:1276
47. Wang X, Zhou T, Bai G, Zhao Y. 2018. Complete chloroplast genome sequence of *Fagopyrum dibotrys*: genome features, comparative analysis and phylogenetic relationships. *Scientific Reports* 8:12379
48. Weng ML, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution* 31:645–59
49. 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
50. Deng Q, Zhang H, He Y, Wang T, Su Y. 2017. Chloroplast microsatellite markers for *Pseudotaxus chienii* developed from the whole chloroplast genome of *Taxus chinensis* var. *mairei* (Taxaceae). *Applications in Plant Sciences* 5:1600153
51. Liu Q, Li X, Li M, Xu W, Schwarzacher T, et al. 2020. Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biology* 20:406
52. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, et al. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Scientific Reports* 5:16958
53. Kim KJ, Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research* 11:247–61
54. Downie SR, Jansen RK. 2015. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. *Systematic Botany* 40:336–51
55. Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Molecular Biology and Evolution* 24:1161–80
56. Li X, Li Y, Zang M, Li M, Fang Y. 2018. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *International Journal of Molecular Sciences* 19:2443
57. Zhang J, Li R, Xiang X, Manchester SR, Lin L, 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