

The chloroplast genome features and phylogenetic relationships of *Platycarya longipes* (Juglandaceae), an important woody species within karst forests of eastern Asia

Yingliang Liu

Guizhou Normal University

Lijuan Hu

Guizhou Normal University

Xiaoshuang Wang

Guizhou Normal University

Ya Tan

Guizhou Normal University

Lei Gu (✉ leigu1216@nwafu.edu.cn)

Guizhou Normal University

Article

Keywords: Chloroplast, *Platycarya longipes*, Genome comparsion, Illumina reads, Juglandaceae

Posted Date: May 9th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1602797/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Platycarya longipes of the Juglandaceae family is an important woody species in maintaining the stability of community structure of karst forests. However, its phylogenetic relationship within Juglandaceae is still unclear. In this study we assembled the complete cp genome of *P. longipes*. The genome comprises a 158,592 bp quadripartite circular that includes a large single copy (LSC) region of 88,066 bp and a small single copy (SSC) region of 18,524 bp separated by a pair of inverted repeats (IRA and IRB) with 26,001 bp. The genome contains 113 unique genes, including 80 protein-coding genes, 29 tRNAs and 4 rRNAs. Additionally, we detected 49 long repeat sequences and 66 simple sequence repeats (SSRs). Analysis of the Ka/Ks substitution rate values in the comparison of *P. longipes* VS. *Platycarya strobilacea*, supported that *P. longipes* and *P. strobilacea* are two species. Compared with other species of Juglandaceae, the cp genome of *P. longipes* has a conserved gene order and structure. Phylogenetic analysis based on ML and BI methods using genomes of the Fagales order showed that *P. longipes* is most closely related to *Platycarya strobilacea*. Our research provides a critical genetic resource for *P. longipes* supporting future phylogenetic and population genetics studies.

Introduction

The karst landscape results from the action of rainfall and groundwater on carbonate bedrock¹ and is widespread globally, accounting for 12% of the world land area². More karst landscape occurs in China than anywhere else in the world, and it is mainly distributed in mountainous regions in the south-western part of the country, particularly in the province of Guizhou^{3,4}. Karst regions generally contain fragile ecosystems due to soils that form extremely slowly, have weak water retention capacity, and have shallow, patchy coverage. Karst ecosystems are maintained in part by karst forests, which provide valuable ecosystem services⁵, and within these forests, woody species comprise vital biodiversity^{6,7}. Therefore, understanding the genetic diversity and phylogenetic relationships of woody species of karst forests is critical for modern approaches to management and conservation.

Juglandaceae, the walnut family, comprises nine genera and 71 species of which seven genera and 27 species occur in karst regions of China⁸. Thus, this family plays an important role in maintaining the community structure of karst forest ecosystems especially due to adaptations of species to the challenging edaphic environment⁴. *Platycarya longipes*, as a member of Juglandaceae family, is widely distributed in karst forests of southern China and represents a critical element within the karst ecosystem⁴. Additionally, this species is valued for its bark and leaves, which is rich in gallic and ascorbic acid^{9–12} and consequently, has antioxidant and pro-oxidant properties¹³. Nevertheless, despite the ecological and medicinal importance of *P. longipes*, there have been no studies of its plastid genome, genetic diversity, or phylogenetic relationships with other species of Juglandaceae or the Fagales order to our knowledge.

The chloroplast (cp) genome, which is maternally inherited in angiosperms, is highly conserved in gene content and genome structure¹⁴ and is an ideal system for deciphering genome evolution^{15,16}, performing DNA barcoding, and inferring phylogenetic relationships in angiosperm families that have evolutionary histories recalcitrant to traditional morphological approaches or molecular phylogenetic approaches using a few DNA markers^{17–22}. The cp genome of angiosperms generally comprises a quadripartite, circular molecule including one large single copy (LSC) region and one small single copy (SSC) region, which were separated by two inverted repeat regions (IRA and IRB)²³. Most cp genomes range from 120 to 160 kb in length and harbor 110–130 unique genes that are essential to photosynthesis and the biosynthesis of starch, amino acids, fatty acids, and pigments²⁴. Recently, owing to the advances of high-throughput sequencing, thousands of cp genome sequences are now publicly available via the National Center for Biotechnology Information (NCBI), since the first complete chloroplast genome was sequenced in tobacco (*Nicotiana tabacum* L.) in 1986²⁵. Among these, the cp genome of *Platycarya strobilacea* (KX868670) has provided valuable information for resource conservation⁹. However, the cp genome of *P. longipes* has not been sequenced.

In this study, we assembled the complete cp genome of *P. longipes* *de novo* from Illumina short reads. Within the assembled cp genome, we identified a total of 66 simple sequence repeats (SSRs) loci and 49 long duplicates repeats. We used the complete chloroplast genome sequence of *P. longipes* and related species of Fagales to perform phylogenetic analysis by ML and BI methods. Overall, our results provide valuable information for the further development of genetic resources to support ecological and evolutionary studies of *P. longipes* and its close relatives.

Materials And Methods

Ethics statement

During the leaf samples collection, no harms was done to the environment, this study did not involve endangered or protected species, and no specific permits were required for collection.

Plant materials and sequencing

We collected a total of 5g of young fresh leaves of *P. longipes* on campus at Guizhou Normal University of China (26°23'.12"N, 106°38'32" E). We extracted total DNA from the leaves using the DNeasy Plant Mini Kit (Qiagen, USA) according to manufacturer instructions and assessed the quality and quantity of the DNA by agarose gel electrophoresis. We used the extracted DNA to construct a library from fragments ~ 450 bp in size for the Illumina HiSeq X Ten (Illumina, USA) platform following manufacturer's protocols.

Genome assembly and gene annotation

We obtained 150 bp paired-end reads through Illumina HiSeq X Ten sequencing. After removing sequencing adapters and low-quality reads, we selected out sequences representing the cp genome by

aligning reads to the closely related species, *P. strobilacea*⁹ using BLASR²⁶ with default parameters. We used the selected reads to construct the draft cp genome of *P. longipes* in SOAPdenovo (v2.04)²⁷, performed sequence extension in SSPACE²⁸, and accomplished gap filling in GapCloser using default parameters²⁹.

Then we employed the software of Dual Organellar GenoMe Annotator (DOGMA)³⁰ to annotate the genes within the cp genome, including protein-coding genes, tRNAs, and rRNAs, and we manually identified coding sequence boundaries according to the positions of start and stop codons. We used OGDRAW v1.2³¹ to circularize the annotated gene map, and we deposited the annotated cp genome of *P. longipes* in GenBank (accession number MT032191).

Identification of long repeat sequences and simple sequence repeats

We used the REPuter webserver (<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>)³² to identify long repeats of at least 30 bp, with sequence identity above 90% or greater including forward, palindrome, reverse, and complement repeats. We detected simple sequence repeats (SSR) using MISA-web (<https://webblast.ipk-gatersleben.de/misa/>)³³ with the following settings: ten minimal repeats for mono-nucleotides, five for di-, four for tri-, and three for tetra-, penta-, and hexa- nucleotides.

Analysis of codon usage

Analysis of codon usage not only reflects the origin, evolution and mutation mode of species or genes, but also has an important influence on gene function and protein expression^{34–36}. CodonW1.4.2 (<http://downloads.fyxm.net/CodonW-76666.html>) was used to calculate the relative synonymous codon usage (RSCU) of *P. longipes* chloroplast protein-coding genes under the default parameters.

Comparisons of the whole cp genomes of related species

We compared sequence divergence of the complete cp genome of *P. longipes* with *Carya illinoensis*, *Castanopsis echinocarpa*, *Cyclocarya paliurus*, *Juglans hopeiensis*, *Quercus acutissima* and *P. strobilacea* using mVISTA in the Shuffle-LAGAN mode³⁷. The SNPs and indels between the *P. longipes* and *P. strobilacea* cp genome were detected by MUMMER3.23 with the default settings (maxgap = 500, mincluster = 100). Additionally, we visualized comparisons of the LSC/IRB/SSC/IRA junctions in seven species of Juglandaceae, including *C. illinoensis*, *C. paliurus*, *J. hopeiensis*, *J. cinerea*, *J. major*, *P. strobilacea*, and *P. longipes*, according to their annotations of chloroplast genomes deposited in GenBank using IRSCOPE (<https://irscope.shinyapps.io/irapp/>).

Molecular evolution analysis

To assess the synonymous (Ks) and nonsynonymous (Ka) substitution rates, We calculated pairwise comparisons of 62 commonly conserved protein-coding genes between *P. longipes* and the six closely related species mentioned above in mVISTA analysis, and the Ka/Ks ratios were computed by TBtools³⁸ using the default parameters of Simple Ka/Ks calculator mode.

Phylogenetic analysis

We obtained a total of 31 cp genomes (nucleotide level) of the Fagales including 15 species of Juglandaceae, four species of Fagaceae, and 12 species of Betulaceae from GenBank and used these together with *P. longipes* for phylogenetic reconstruction. The complete chloroplast genome sequence of these 32 species were aligned using the MAFFT software with default parameters, we performed phylogenetic reconstruction of the selected species of Fagales in MEGA7.0³⁹ using the maximum likelihood (ML) method based on the Tamura-Nei model. And 1000 bootstrap replicates were set to infer node support, branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Meanwhile, the Mrbayes 3.2.7⁴⁰ under GTRGAMMA model was used to construct a phylogenetic tree with the Bayesian inference (BI) method, four chains of the Markov Chain Monte Carlo were run each for 1,000,000 generations and were sampled every 100 generations.

Results

Assembly and features of the *P. longipes* cp genome

We obtained a total of 8.46 Gb raw reads from Illumina sequencing platform. After trimming, we retained 1.15 Gb of clean reads, from which we performed *de novo* assembly of the complete cp genome of *P. longipes*. The cp genome showed a typical circular quadripartite structure that was 158,459 Qbp in length, contains 113 unique genes, including 80 protein-coding genes, 29 tRNAs and 4 rRNAs. It included a large single copy (LSC) region of 87,898 bp, a small single copy (SSC) region of 18,521 bp, which were separated by two inverted repeats (IRa and IRb) having a total of 26,020 bp. The overall GC content of the *P. longipes* cp genome was 36.16%. The two IR regions had the highest GC content of 42.54%, followed by 33.76% in the LSC region, and 29.67% in the SSC region (Table 1; Fig. 1).

Among the 113 unique genes, ten genes, comprising four protein-coding genes and six tRNA genes, had one intron; and only two genes (ndhB and trnR-UCU) possessed two introns (Table 2).

Table 1

Summary of the complete chloroplast genomes of *P. longipes* and five closely related species

Genome Features	<i>P. longipes</i>		<i>P. strobilacea</i>	<i>C. illinoensis</i>	<i>J. hopeiensis</i>	<i>C. paliurus</i>	<i>Q. acutissima</i>
Length (bp)	158,459	160,994	160,819	159,714	160,562	161,129	
GC content (%)	36.16	36.04	36.14	36.14	36.08	36.78	
LSC length (bp)	87,898	90,225	90,042	89,316	90,007	90,423	
LSC GC content (%)	33.76	33.59	33.74	33.71	33.66	34.62	
SSC Length (bp)	18,521	18,371	18,791	18,352	18,477	19,070	
SSC GC content (%)	29.67	29.72	29.89	29.79	29.71	31.31	
IR length (bp)	26,020	26,199	25,993	26,023	26,039	25,817	
IR GC content (%)	42.54	42.47	42.58	42.56	42.55	42.77	
Total genes	113	112	107	112	116	114	
Protein genes	80	79	77	79	81	79	
tRNA genes	29	29	26	29	31	31	
rRNA genes	4	4	4	4	4	4	

Table 2
Gene composition in the chloroplast genome of *P. longipes*

Category of genes	Group of genes	Name of genes
photosynthesis	Subunits of NADH-dehydrogenase	<i>ndhJ, ndhK, ndhC, ndhB^{a,c}, ndhH, ndhA, ndhI, ndhG, ndhE, ndhD, ndhF</i>
	Large subunit of Rubisco	<i>rbcL</i>
	Subunits of photosystem II	<i>psbA, psbK, psbI, psbD, psbC, psbZ, psbF, psbE, psbB, psbH</i>
	Subunits of photosystem I	<i>psaB, psaA, psaI, psaJ, psaC</i>
	Subunits of ATP synthase	<i>atpA, atpF, atpH, atpI, atpE, atpB,</i>
	Subunits of cytochrome b/f complex	<i>petA, petB, petD, petL, petG</i>
	photosystem II assembly	<i>ycf3^b, ycf4</i>
Self-replication	Ribosomal RNA genes	<i>rrn16^a, rrn23^a, rrn4.5^a, rrn5^a</i>
	Transfer RNA genes	<i>trnG-GCC, trnS-GGA, trnL-UAA^b, trnF-GAA, trnM-CAU, trnI-GAU^b, trnA-UGC^{a,b}, trnR-ACG^a, trnN-GUU^a, trnR-UCU^a, trnC-GCA, trnT-GGU, trnS-UGA, trnE-UUC, trnY-GUA, trnD-GUC, trnS-GCU, trnQ-UUG, trnH-GUG, trnV-GAC^a, trnI-GAU^{a,b}, trnA-UGC^b, trnR-ACG, trnL-UAG, trnR-UCU^c, trnL-CAA^a, trnM-CAU, trnP-UGG, trnW-CCA, trnC-ACA^b, trnT-UGU</i>
	Small subunit of ribosome	<i>rps16^b, rps2, rps14, rps4, rps18, rps11, rps8, rps3, rps19, rps7, rps15, rps7^a, rps12^b</i>
	Large subunit of ribosome	
		<i>rpl33, rpl20, rpl14, rpl16, rpl22, rpl2^a, rpl23^a</i>
^a indicates genes duplicated in the IR regions		
^b indicates the genes containing a signal intron		
^c indicates the genes containing two signal introns		

Category of genes	Group of genes	Name of genes
	DNA-dependent RNA polymerase	<i>rpoC2, rpoC1, rpoB, rpoA</i>
	Translation initiation factor	<i>infA</i>
Other genes	Maturase	<i>matK</i>
	Subunit of acetyl-CoA	<i>accD</i>
	Protease	<i>ClpP^b</i>
	Envelope membrane protein	<i>cemA</i>
	C-type cytochrome synthesis	<i>ccsA</i>
Functionally unknown genes	Conserved Open reading frames	<i>ycf1, ycf2^a</i>
^a indicates genes duplicated in the IR regions		
^b indicates the genes containing a signal intron		
^c indicates the genes containing two signal introns		

Detection of long repeat sequences and SSRs

We detected a total of 49 long repeats in the cp genome of *P. longipes* ranging from 37 to 78 bp in length. These included 32 forward, 13 palindromic, and four reverse repeats, but we detected no complement repeat was detected. Most repeats (34, 69.39%) were located in intergenic spacer (IGS) regions, 14 repeats (28.57%) occurred within coding sequences (CDS), and 11 repeats (22.45%) were in introns (Table S1). Among these repeats, 10 were of 30–39 bp in size, 14 were 40–49 bp, 13 were 50–59 bp, nine were 60–69 bp, and three were 70–79 bp (Table S1).

In the complete cp genome of *P. longipes*, we detected 66 SSR loci of 15 different types with lengths of at least 10 bp, including 47 mononucleotides, 11 dinucleotides, three trinucleotides, four tetranucleotides, and one pentanucleotide (Table S2). Of the 47 mononucleotides, 46 were A or T types, and only one was

a G type as is consistent with observations in other cp genomes of angiosperms^{21,22,41}. Among the dinucleotide repeats, AT (6, 54.5%) was observed more frequently than TA, AG, CT and TC, the trinucleotides repeats comprised ATT and TAT, the tetranucleotides were TTTA, AATA, CTTT and AAAG, and the pentanucleotide was AATAT. Out of the 66 SSRs, 51 SSR loci occurred in the LSC region (77.27%), nine in the SSC region (13.64%), and six among the two IR regions (9.09%) (Table S2). 14 identified SSRs were within the coding regions, while 51 were located in the intergenic regions and only one was located in the intron regions.

Codon usage analysis

The codon usage frequency and RSCU were analyzed based on the sequence of 80 protein-coding genes in the *P. longipes* chloroplast genome (Figure S1), a total of 25529 codons were detected. The statistics analysis of all protein-coding cpDNA and amino acid sequences showed obvious codon preferences. Of these codons, 2693 (10.54%) encoded leucine, whereas only 298 (1.16%) encoded cysteine, indicating the most and the least frequently used amino acids in the *P. longipes* cp genome, as observed in the plastomes of other angiosperms such as the early diverging species⁴². The codon usage frequency and RSCU were used as a relative intuitionistic to measure the extent of codon bias⁴³, based on sequences of 80 distinct protein-coding genes in the *P. longipes* chloroplast genome. The results showed that the AUU had the highest frequencies and the UGC had the lowest frequencies. 20 amino acids were encoded by 61 codons, the RSCU value of 31 codons were > 1, indicating that these codons exist preference. Moreover, among the preferred codons, except UUG and UCC, all of the preferential codons ended with A/U, supporting the idea that such biased usage of certain degenerate codons was likely a result of adaptive evolution of cp genome.

Analysis of genome divergence

We determined genomic similarity and divergence among *P. longipes* and six related species in mVISTA, using the cp genome of *P. longipes* as a reference. The result showed that more than 95% of regions were well conserved among these species, indicating a high degree of sequence similarity. In addition, the non-coding regions are more variable than coding regions, however, we observed lower levels of sequence conservation in *rp122*, *rpoC1*, and *petD* (Fig. 2).

A total of 2667 (616 SNPs and 2051 indels) variable sites were observed between the *P. longipes* and *P. strobilacea* chloroplast genomes, among them, 2.40% variations (1712 SNPs and 401 indels) were within the LSC region, 2.04% (213 SNPs and 165 indels) were within the SSC region, while 0.34% (126 SNPs and 50 indels) were within the region of IRs (Figure S2). The results suggested that the IR regions were more conserved than SC regions in the cp genome of *Platycarya*. In spite of this, the chloroplast genome sequences of *P. longipes* and *P. strobilacea* still showed significant differences.

Comparison of boundaries regions

We used seven cp genomes of species of Juglandaceae to compare the boundaries of the SSC, LSC, and IR regions using the IRscope webserver. The result showed that the size of the IR was highly conserved,

ranging from 25,993 bp to 26,199 bp and that the genes located in the LSC/IRb and SSC/IRa border regions were also highly conserved. In particular, the LSC/IRb boundaries were located between *rps19* and *rpl2* genes in all seven cp genomes, and the IRa/SSC boundaries were located within the pseudogene *ycf1*. However, genes in IRb/SSC and IRa/LSC junctions were inconstant (Fig. 3). The IRa/LSC border was located between *rpl2* and *trnH* genes in five of the cp genomes, including *P. longipes*, *P. strobilacea*, *C. illinoensis*, *C. paliurus*, and *J. hopeiensis*, whereas the boundary was between *rpl23* and *trnH* in *J. cinerea* and *J. major*. In *P. longipes*, *P. strobilacea*, and *C. paliurus*, the border of IRb/SSC was located between *ycf1* and *ndhF* genes, however, either *ycf1* or *ndhF* gene was absent from IRb in the other four cp genomes.

Phylogenetic analysis

Chloroplast genomes have been widely used to determine the phylogenetic relationships because they are highly conserved in terms of gene size and content, genome structure, and linear order of the genes. We employed 32 selected species of Fagales (Table S3) for phylogenetic reconstruction. The Maximum Likelihood phylogenetic tree possessed a total of 28 branches with bootstrap values of above 85%. Among these branches, 26 branches were supported by values above 90% (Fig. 4A). As expected, *P. longipes* was most closely related to the congeneric species, *P. strobilacea*. The genus *Platycarya* formed a monophyletic clade with 100% bootstrap support, showed the most closed relationship to *Cyclocarya* genus. Moreover, both the ML and BI phylogenetic (Fig. 4B) tree showed nearly identical topologies in identifying the taxonomic status of 32 species.

Analysis of selection pressure

The Ka/Ks ratio is widely used to infer rates of genomic evolution and selection pressure on individual genes^{44–46}. The ratio of $Ka/Ks < 1$, $Ka/Ks = 1$, and $Ka/Ks > 1$ indicate that genes underwent purifying, neutral, and positive selection, respectively³⁹. In this study, we calculated the pairwise Ka/Ks ratios of 62 common protein-coding genes between the *P. longipes* cp genome and six related species (Table S4), including *C. illinoensis*, *C. echinocarpa*, *C. paliurus*, *J. hopeiensis*, *Q. acutissima* and *P. strobilacea*. Overall, the average Ka/Ks value of these genes in the seven genomes was 0.246. The majority of common genes (40 of 62 genes) had an average Ka/Ks ratio of 0 and 0.3 when compared to *P. longipes*, suggesting that these genes were subject to strong purifying selection. The average Ka/Ks ratio of all comparisons of the *atpF* gene was 1.52, ranging from 0.668 (*P. longipes* vs. *P. strobilacea*) to 1.863 (*P. longipes* vs. *C. paliurus* and *P. longipes* vs. *J. hopeiensis*), indicating that this gene has undergone strong positive selection. Moreover, *matK*, *rpoA*, *petD*, *atpF*, *rpl22*, and *ycf2* also exhibited high ratios, with $Ka/Ks > 0.5$ among the six pairwise comparisons (Table S4, Fig. 5).

Comparison analysis of SSR and long repeats

Simple sequence repeats (SSRs), also known as microsatellites, are frequently used as molecular markers in population genetics and evolutionary studies of higher eukaryote genomes¹⁵. In the present study, we detected complete SSRs among the six cp genomes of species of Fagales (Fig. 6), the results revealed a total of 66, 61, 62, 72, 78 and 83 SSRs in the *P. longipes*, *C. illinoensis*, *P. strobilacea*, *J.*

cinerea, *Corylus yunnanensis* and *Q. acutissima* cp genomes, respectively. *Q. acutissima* of Fagaceae had the largest number of SSRs, followed by *C. yunnanensis* of Betulaceae. Similarly, hexanucleotide SSRs (AACAGA and TTTTAT) were detected in the cp genome of *C. yunnanensis* and *Q. acutissima* but not in the family of Juglandaceae (*P. longipes*, *C. illinoensis*, *P. strobilacea* and *J. cinerea*). Furthermore, we observed a significantly larger number of A and T microsatellites than G and C as expected based on reports from other species of angiosperms^{47–49}. These results suggest that SSRs can be used to conduct evolutionary analysis and are powerful for identifying the genetic diversity among different species.

Longer repeat sequences facilitate base substitutions, evolution of genome size, and genomic rearrangements in cp genomes and are useful for phylogenetic studies^{50,51}. We detected a total of 294 long repeat sequences across the six genomes with a length distribution of 30–109 bp, most of them were 30–60 bp long and accounted for 87.41% of the total, and two duplicates with a length greater than 100 were only detected in *J. cinera*. Each species possessed 49 long repeats, the number of F (Forward, 156) and P (Palindromic, 110) reached 266 among four types of repetition, accounting for 90.48% of the total, and we detected only one complement repeat, which was in *C. illinoensis* (Fig. 7). The number and pattern of repeat sequences were highly similar and conserved within the six cp genomes of Fagales. Taken together, the long repeats and SSRs may represent valuable lineage-specific markers for population biology and molecular phylogenetic studies in this plant order^{41,48}.

Discussion

Genome features

In general, the size of cp genomes in photosynthetic land plants ranges from 108 kb to 165 kb^{47,52–54}, most cp genomes of the angiosperm are considered to be conserved. The size of the cp genome of *P. longipes* was 158,459 bp and is similar to the sizes of cp genomes previously reported in other species of Juglandaceae, such as *C. illinoensis* (160,819 bp), *P. strobilacea* (160,994 bp), *J. hopeiensis* (159,714 bp), and *C. paliurus* (160,562 bp). Among the species we compared, *Quercus acutissima* of Fagaceae had the largest cp genome (161,129 bp), indicate that the length of cp genomes within Juglandaceae family is conservative. The LSC regions in the genomes compared were varied from 88,066 bp to 90,423 bp in lengths, the SSC ranged from 18,352 bp to 19,070 bp, and the IR regions were from 25,817 bp to 26,199 bp (Table 1). Notably, *Q. acutissima* has the longest overall length (161,129 bp) but the shortest IR regions (25,817 bp), which may be attributed to the contraction of the IR regions. The overall GC content of these cp genomes was approximately 36% and was unevenly distributed among the LSC, SSC, and IR regions, which had 34%, 30%, and 42% GC content, respectively. Compared with the LSC and SSC regions, the GC content is greater in IR regions of all Fagales, this unequal distribution of GC content is typical for angiosperms^{55,56}, in which the presence of ribosomal RNA (rRNA) sequences appears to increase the GC content of the IR regions^{57,58}.

The expansion and contraction of the IR regions was the main reasons for variation of cp genomes size, and evaluating this difference could reveal the evolution of related taxa^{59,60}. The size of IR regions was relatively conserved, but there were some differences in adjacent genes and junctions. The junctions of *P. longipes*, *P. strobilacea* and *C. paliurus* were nearly identical with only slight differences in the distance of the boundary, whereas there were significant differences in the boundaries of genes in *P. longipes* compared to *C. illinoensis*, *J. hopeiensis*, *J. cinerea*, and *J. major*. Although there were some changes in the cp IR boundary regions, the size of the overall genome, base composition of the LSC, SSC and IR regions of *P. longipes* was similar to those closely related species. Based on comparisons of the complete cp genome of studied species, the number of genes, genome size, gene order and genome structure were similar, this further indicates that cp genomes are generally conserved.

Codon usage bias and selection pressure

Codon usage bias was considered to be the consequence of the balance between gene mutation and natural selection. Generally, the GC content at the first, second and third base positions per codon is largely different, and it is consider that the first base position has the highest GC content, following by second and third position⁶¹. Additionally, the dicot plants mostly ending with A or T, while the monocot plants mostly ending with G or C⁶². The analysis of codon usage revealed that codons encoding proteins in *P. longipes* chloroplast genomes tend to end with A/T, this result is consistent with previous studies^{63,64}. The GC content varies differently in three positions, indicating the chloroplast genome in *P. longipes* mostly affected by natural selection, while little affected by gene mutations or other factors.

The synonymous and nonsynonymous substitution incidents were widely occurred in the process of gene evolution, which can be used, to evaluate the rates of genomic evolution and determine whether the protein-coding gene has a selective effect. It is believed that the *Platycarya* genus comprises of two closely related species, *P. longipes* and *P. strobilacea*, for a long time⁶⁵. Chen et al. implemented a phylogeographical study on *P. strobilacea* using *psbA-trnH* and *atpB-rbcL* intergenic spacer sequences of cpDNA to demonstrate that *Platycarya* is likely a monotypic genus⁶⁶. But a later study which employed both nuclear genetic marker and cpDNA marker showed that the interspecific genetic divergence was more fitting with 'two species' scenario⁶⁷. In the present study, the cp genome of *P. longipes* has 158,592 bp in length, shorter than the cp genome of *P. strobilacea* (160,994 bp in length)⁹. Additionally, the Ka/Ks values of these genes (*ycf3*, *rpoB*, *rpl2*, *matK*, *accD*, *petD*, and *clpP*) in the comparison of *P. longipes* VS. *P. strobilacea* were even higher than the comparisons between *P. longipes* and other species in Juglandaceae, likely supported the idea that *P. longipes* and *P. strobilacea* are two species. We noticed that the *petD* gene, which controls the cytochrome b6/f complex, affecting photosynthetic efficiency⁶⁸, always showed a significant positive selection (average Ka/Ks value of 2.995) in *Platycarya*. This gene can be considered as a glimpse of response of *Platycarya* on the drought habitat of karst. Moreover, most genes involved in the functional category "Subunits of photosystem", such as *psbA*, *psaC*, *psbE*, *psaB*, *psbC* and *psbD* genes, have undergone lower purifying selection pressure.

Relationship analysis

Both ML and BI phylogenetic tree revealed that the 16 species representing Juglandaceae comprised of multiple clades and that *P. longipes* was most closely related to *P. strobilacea* (Fig. 4). The tree topology was consistent with the traditional tribal-level classification and nuclear RAD-Seq data of Juglandaceae^{24, 69}. Furthermore, the ML and BI tree showed that Juglandaceae was more closely related to Betulaceae than to Fagaceae, this is consistent with the findings in prior studies⁷⁰.

Conclusion

In this study, we assembled the complete chloroplast genome of *P. longipes* using a *de novo* approach and found that it was consisted of 158,459 bp in total and exhibited a typical quadripartite, circular structure comprising an LSC, SSC and two IR regions, including 80 protein-coding genes, 29 tRNAs and four rRNAs. We detected 49 long repeats and 66 SSRs in the cp genome of *P. longipes* that may be useful for development of molecular markers as well as phylogenetic and polpulation studies in *P. longipes*. Our analyses of selection pressure revealed strong positive selection on *atpF* gene in *P. longipes*. The relative high Ka/Ks values of *ycf3*, *rpoB*, *rpl2*, *matK*, *accD*, *petD*, and *clpP* were observed in the comparison between *P. longipes* and *P. strobilacea*, likely support the idea that *P. longipes* and *P. strobilacea* were two different species. The result of our phylogenetic analysis based on ML and BI method showed that *P. longipes* was most closely related to the congeneric species, *P. strobilacea*. Our results provide insight into the evolutionary relationships of Juglandaceae and genomic evolution in Fagales, as well as represent a new genetic resource for future phylogenetic, taxonomic, ecological, population biology, and conservation studies. However, it is limited to study the taxonomic status and phylogenetic relationship of Fagales only based on chloroplast genome. With the development of high-throughput sequencing technology, the nuclear genome information will also be integrated in future studies.

Declarations

Guidelines Statement: The collection of plant material is in comply with relevant institutional, national, and international guidelines and legislation.

Data Availability Statement: The annotated chloroplast genome data that support the findings of this study are openly available in GenBank of NCBI at <https://www.ncbi.nlm.nih.gov> under the accession number MT032191.

Funding: This research was Supported by National Natural Science Regional Fund Project (31760124), The Joint Fund of the National Natural Science Foundation of China and the Karst Science Research Center of Guizhou province (Grant No. U1812401).

Author Contributions

Conceptualization: Lei Gu, Yingliang Liu

Data curation: Lijuan Hu,Xiaoshuang Wang

Funding acquisition: Yingliang Liu

Resources: Xiaoshuang Wang, Ya Tan

Writing-review & editing: Yingliang Liu, Lijuan Hu, Lei Gu

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, X. Y. *et al.* Positive correlation between soil bacterial metabolic and plant species diversity and bacterial and fungal diversity in a vegetation succession on karst. *Plant and Soil*. **307**, 123-134. (2008).
2. Liu, C. C. *et al.* Comparative ecophysiological responses to drought of two shrub and four tree species from karst habitats of southwestern China. *Trees-struct Funct*. **25**, 537-549. (2011).
3. Li, Y. B., Hou, J. J. & Xie, D. T. The recent development of research on karst ecology in southwest china. *Scientia Geographica Sinica*. **22**, 365-370. (2002).
4. Zhang, Z. H., Hu, G., Zhu, J. D. & Ni, J. Stand structure, woody species richness and composition of subtropical karst forests in Maolan, south-west china. *J. Trop For. Sci*. **24**, 498-506. (2012).
5. Ran, J. C., He, S. Y., Cao, J. H., Xiong, Z. B. & Chen, H. M. Benefit of soil and water conservation at a subtropical karst forests: illustrated by Maolan National Nature Reserve, Guizhou Province, China. *J. Soil. Water. Conserv*. **16**, 92-95. (2002).
6. Noss, R. F. Indicators for monitoring biodiversity: a hierarchical approach. *Conserv. Biol*. **4**, 355-364. (1990).
7. Novotny, V. *et al.* Why are there so many species of herbivorous insects in tropical rainforests? *Science*. **313**, 1115-1118. (2006).
8. Lu, X., Huang, H., Nemchuk, N. & Ruoff, R. S. Patterning of highly oriented pyrolytic graphite by oxygen plasma etching. *Appl. Phys. Lett*. **75**, 193-195. (1999).
9. Yan, J., Han, K., Zeng, S., Zhao, P. & Liu, Z. L. Characterization of the complete chloroplast genome of *Platycarya strobilacea* (Juglandaceae). *Conserv. Genet. Resour*. **9**, 79-81. (2016).
10. Wang, M. Y., Liu, J. T. & H, N. Determination of gallic acid in *Platycarya strobilacea* Sieb. et Zucc by RP-HPLC. *China Pharm*. **13**, 378-379. (2010).
11. Yan, Y. Determination of ascorbic acid in *Platycarya longipes* by spectrophotometry. *Journal of Anhui Agricultural Science*. **18**, 149-152. (2010).
12. Yan, Y., Jian, Z., Xiao, C., Zai-Bo, Y. & Cheng, M. L. Determination of gallic acid in *Platycarya longipes*. *Chinese Journal of Experimental Traditional Medical Formulae*. **17**, 107-109. (2011).
13. Yen, G. C., Duh, P. D. & Tsai, H. L. Antioxidant and pro-oxidant properties of ascorbic acid and gallic acid. *Food Chemistry*. **79**, 307-313. (2002).

14. Wicke, S., Schneeweiss, G. M., Depamphilis, C. W. & Kai, F. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant. Mol. Biol.* **76**, 273-297. (2011).
15. Duan, R. Y., Yang, L. M., Lv, T., Wu, G. L. & Huang, M. Y. The complete chloroplast genome sequence of *Pinus dabeshanensis*. *Conserv. Genet. Resour.* **8**, 395–397. (2016).
16. Asaf, S., Khan, A. L., Khan, M. A., Imran, Q. M. & Lee, I. J. Comparative analysis of complete plastid genomes from wild soybean (*glycine soja*) and nine other glycine species. *Plos One.* **12** (8), 0182281. (2017).
17. Huang, H., Shi, C., Liu, Y., Mao, S. Y. & Gao, L. Z. Thirteen camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC. Evol. Biol.* **14**, 151. (2014).
18. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One.* **9** (11), e112963. (2014).
19. Gao, Y. X., Zhou, Y. Y., Xie, Y., Feng, L. & Shen, S. G. The complete chloroplast genome sequence of an endangered orchidaceae species *Dendrobium moniliforme* and its phylogenetic implications. *Conserv. Genet. Resour.* **10**, 397-399. (2018).
20. Zhu, B. *et al.* The complete chloroplast genome sequence of garden cress (*Lepidium sativum* L.) and its phylogenetic analysis in Brassicaceae family. *Mitochondrial DNA Part B.* **4**, 3601-3602. (2019).
21. Du, X. Y. *et al.* The complete chloroplast genome sequence of yellow mustard (*Sinapis alba* L.) and its phylogenetic relationship to other Brassicaceae species. *Gene.* **10** (731), 144340. (2020).
22. Zhu, B. *et al.* Chloroplast genome features of an important medicinal and edible plant: *Houttuynia cordata* (Saururaceae). *PloS One.* **15** (9), e0239823. (2020).
23. Kang, H. *et al.* Complete Chloroplast Genome of *Pinus densiflora* Siebold & Zucc. and Comparative Analysis with Five Pine Trees. *Forests.* **10** (7), 600. (2019).
24. Rodriguezpeleta, N. *et al.* Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* **15**, 1325-1330. (2005).
25. Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T. & Sugiura, M. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal.* **5**, 2043-2049. (1986).
26. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics.* **13**, 238-238. (2012).
27. Gogniashvili, M. *et al.* Complete chloroplast genomes Of *Aegilops tauschii* Coss. and *Ae. cylindrica* host sheds light on plasmon devolution. *Curr. Genet.* **62**, 791-798. (2016).
28. Boetzer, M. & Pirovano, W. SSPACE-longread: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics.* **15**, 211-211. (2014).
29. Acemel, R. D. *et al.* A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nature Genetics.* **48**, 336-341. (2016).

30. Wyman, S., Jansen, R. & Boore, J. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*. **20**, 3252-3255. (2004).
31. Lohse, M., Drechsel, O. & Bock, R. Organellar Genome DRAW (ogdraw): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* **52**, 267-274. (2007).
32. Liu, X. *et al.* Complete Chloroplast Genome Sequence and Phylogenetic Analysis of *Quercus bawanglingensis* Huang, Li et Xing, a Vulnerable Oak Tree in China. *Int. J. Mol. Sci.* **10** (7), 0587. (2019).
33. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. **33**, 2583-2585. (2017).
34. Quax, T. E., Claassens, N. J., Söll, D. & Van der Oost, J. Codon bias as a means to fine-tune gene expression. *Mol. Cell.* **59**, 149–161. (2015).
35. Wang, Z. *et al.* Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ*. **8**, e8251. (2020).
36. Li, Y., Kuang, X. J., Zhu, X. X., Zhu, Y. J. & Chao, S. Codon usage bias of *Catharanthus roseus*. *China Journal of Chinese Materia Medica*. **41** (22), 4165-4168. (2016).
37. Dubchak, I. & Ryaboy, D. V. Vista family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods in Molecular Biology*. **338**, 69-89. (2006).
38. Chen, C., Chen, H., He, Y. & Xia, R. TBtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface. *BioRxiv*. **10** (1101), 289660. (2018).
39. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874. (2016).
40. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61** (3), 539-542. (2012).
41. Kuang, D. Y. *et al.* Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome*. **54**, 663-673. (2011).
42. Li, W. *et al.* Interspecific chloroplast genome sequence diversity and genomic resources in *Diospyros*. *BMC Plant Biol.* **18**, 1-11. (2018).
43. Sharp, P. M. & Li, W. H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic. Acids. Res.* **15** (3), 1281–1295. (1987).
44. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32-43. (2000).
45. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. *In International Conference on Machine Learning*. **70**, 1321-1330. (2017).
46. Gao, K. *et al.* Comparative genomic and phylogenetic analyses of *Populus section Leuce* using complete chloroplast genome sequences. *Tree. Genet. Genomes*. **15** (3), 1-12. (2019).

47. Kim, K. J. & Lee, H. L. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA. Res.* **11**, 247-261. (2004).
48. Jansen, R. K., Saski, C., Lee, S. B., Hansen, A. K. & Daniell, H. Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. *Mol. Biol. Evol.* **28**, 835-847. (2011).
49. Cavalier-Smith, T. Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr. Biol.* **12**, 62-64. (2002).
50. Nie, X. *et al.* Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PloS One.* **7**, e36869. (2012).
51. Liu, W. *et al.* Complete chloroplast genome of *Cercis chuniana* (Fabaceae) with structural and genetic comparison to six species in Caesalpinioideae. *Int. J. Mol. Sci.* **19**, 1286-1297. (2018).
52. Palmer, J. D. Comparative Organization of Chloroplast Genomes. *Annu. Rev. Genet.* **19**, 325-354. (1985).
53. Palmer, J. D. Plastid chromosomes: structure and evolution. *The Molecular Biology of Plastids.* **7**, 5-53. (1991).
54. Sugiura, M. The chloroplast genome. *Plant Mol. Biol.* **19**, 149–168. (1992).
55. Terakami, S. *et al.* Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. *Tree. Genet. Genomes.* **8**, 841-854. (2012).
56. Qian, J. *et al.* The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PloS One.* **8** (2), e57607. (2013).
57. Asaf, S. *et al.* The complete chloroplast genome of wild rice (*Oryza minuta*) and its comparison to related species. *Front. Plant. Sci.* **8**, 304-304. (2017).
58. Boudreau, E. & Turmel, M. Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. *Plant. Mol. Biol.* **27** (2), 351-364. (1995)
59. Nazareno, A., Carlsen, M. & Lohman, L. Complete Chloroplast Genome of *Tanaecium tetragonolobum*: The First Bignoniaceae Plastome. *PLoS One.* **10** (6), e0129930. (2017).
60. Raubeson, L. A. *et al.* Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics.* **8**, 174-174. (2007).
61. Liu, H., Lu, Y., Lan, B. & Xu, J. Codon usage by chloroplast gene is bias in *Hemiptelea davidii*. *J. Genetics.* **99** (1), 1-11. (2020).
62. Wang, L. & Roossinck, M. J. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant Mol. Biol.* **61** (4), 699-710. (2006).
63. Zhou, M., Long, W. & Li, X. Analysis of synonymous codon usage in chloroplast genome of *Populus alba*. *J. Forestry Res.* **19** (4), 293-297. (2008).

64. Fu, J. M., Suo, Y. J., Liu, H. M. & Tan, X. F. Analysis on codon usage in the chloroplast protein-coding genes of *Diospyros* spp, *Nonwood Forest Research*. **35** (2), 38-44. (2017).
65. Kuang, K. R. & Lu, A. M. Juglandaceae. In: *Flora Reipublicae Popularis Sinica*. *Beijing: Science Press*. **21**, 8–9. (1979).
66. Chen, S. C. *et al.* Geographic variation of chloroplast DNA in *Platycarya strobilacea* (Juglandaceae). *J. Syst. Evol.* **50** (4), 374-385. (2012).
67. Wan, Q., Zheng, Z., Huang, K., Erwan, G. & Remy, P. Genetic divergence within the monotypic tree genus *Platycarya* (Juglandaceae) and its implications for species' past dynamics in subtropical China. *Tree. Genet. Genomes*. **13**, 1-11. (2017).
68. Xiao, J., Li, J., Ou, Y. M., Yun, T. & He, B. DAC is involved in the accumulation of the cytochrome b6/f complex in *Arabidopsis*. *Plant. Physiol.* **160** (4), 1911-1922. (2012).
69. Mu, X. Y. *et al.* Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Mol. Phylogenetics and Evol.* **147**, 106802. (2020).
70. Li, R. *et al.* Phylogenetic Relationships in Fagales Based on DNA Sequences from Three Genomes. *Int. J. Plant. Sci.* **165**, 311-324. (2004).

Figures

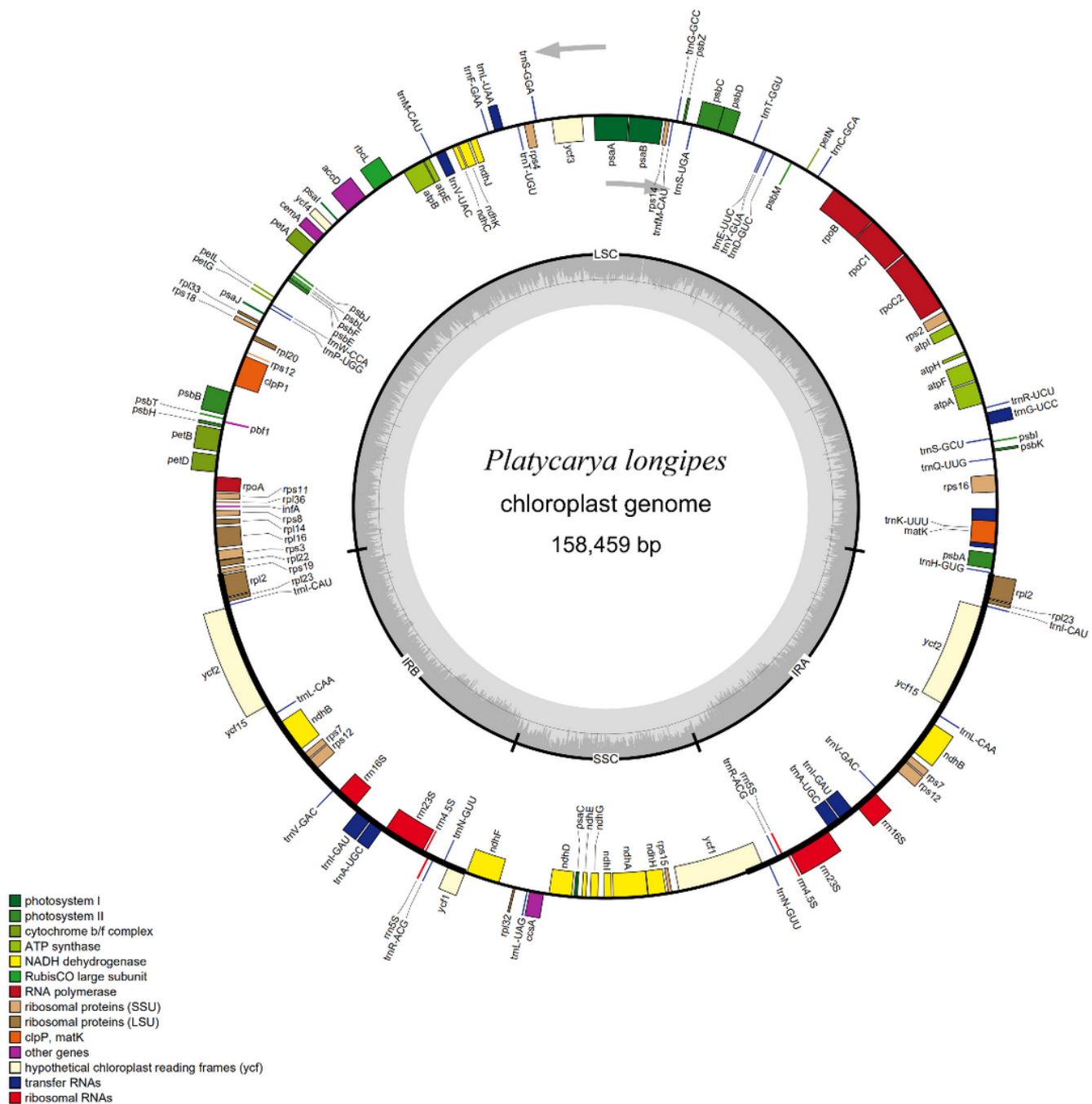


Figure 1

Gene map of the complete chloroplast genome of *P. longipes*. Genes on the outside of the circle are transcribed clockwise, while genes inside are counterclockwise. Genes belonging to different functional groups are shown in different colors. The thick lines indicate the extent of inverted repeats (IRa and IRb), which separated the genomes into large signal-copy (LSC) and small signal-copy (SSC) regions. In the inner circle, the dark gray corresponds the GC content and the light gray corresponds to the AT content.

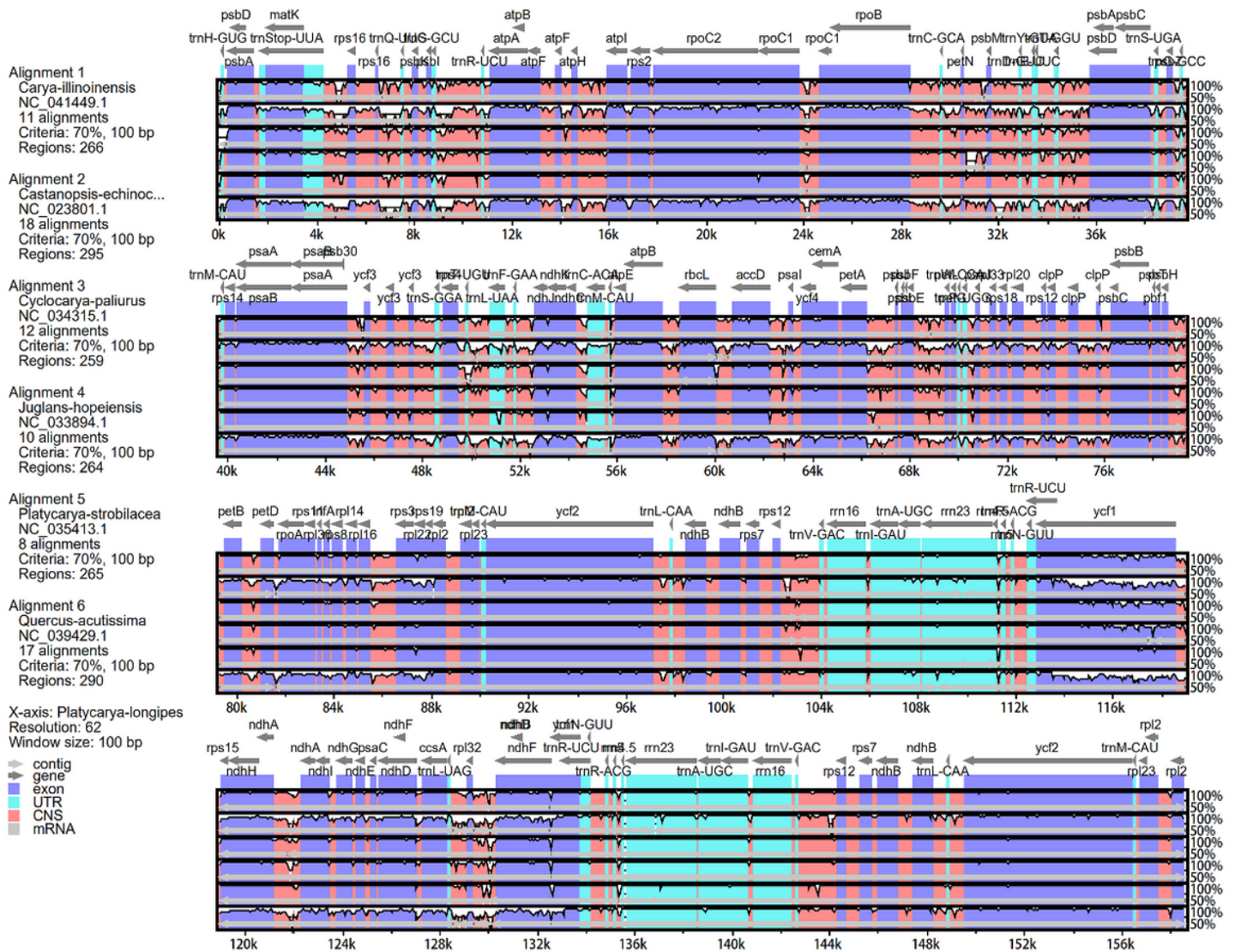


Figure 2

Visualization of alignment of the seven Fagales chloroplast genome sequences using *P. longipes* as a reference. Gray arrows and thick black lines above the alignment indicate gene orientation. Purple bars represent exons, blue bars represent untranslated regions (UTRs), pink bars represent non-coding sequences (CNSs), gray bars represent mRNA, and white peaks represent differences in genomics. The y-axis represents the percentage identity (shown: 50–100%).

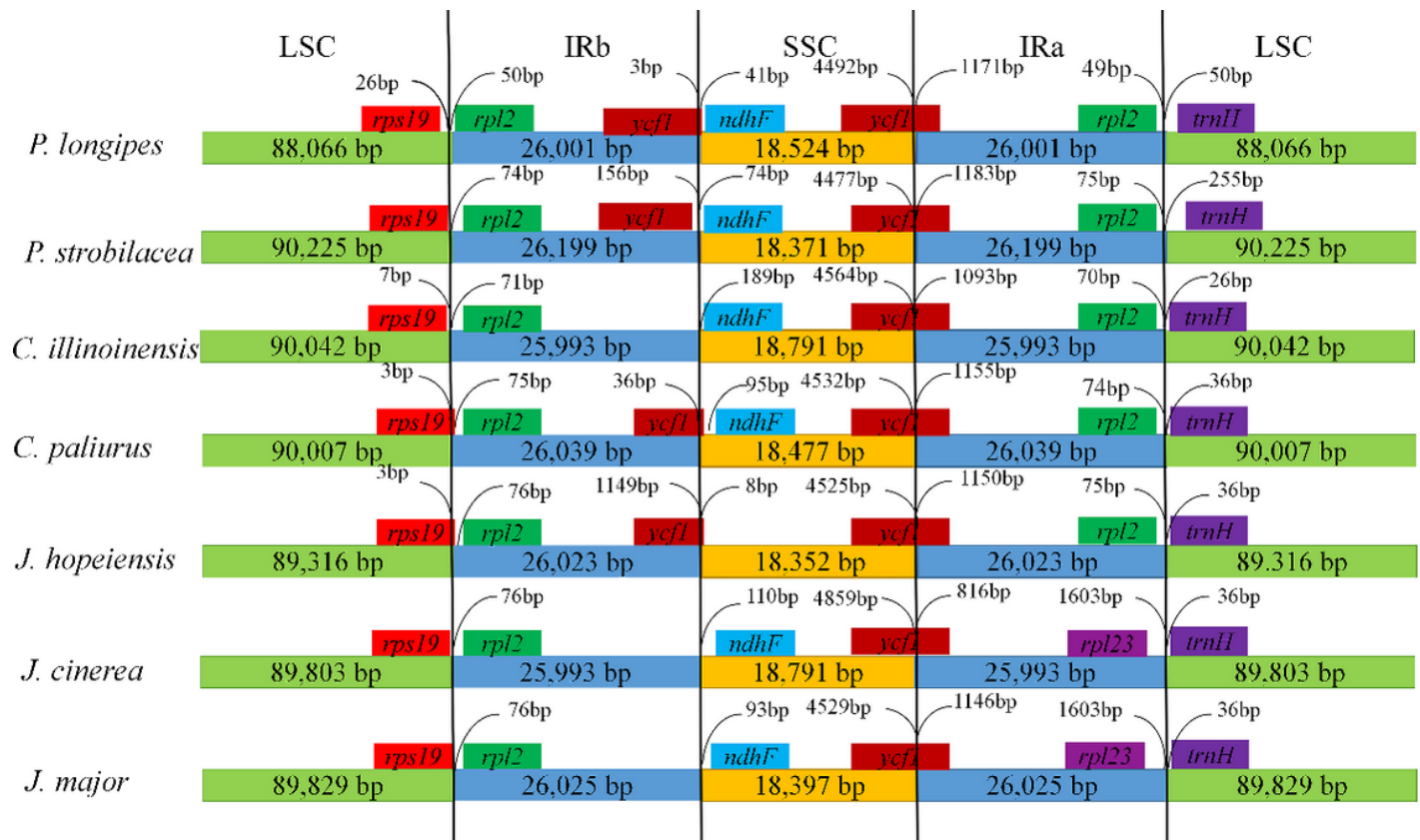


Figure 3

Comparison of the borders of large signal-copy (LSC), inverted repeats (IRa and IRb) and small signal-copy (SSC) between *P. longipes* and other six related species. Boxes above the main line indicate the adjacent border genes. The figure is not to scale regarding sequence length, and only shows relative changes at or near the IR/SC borders.

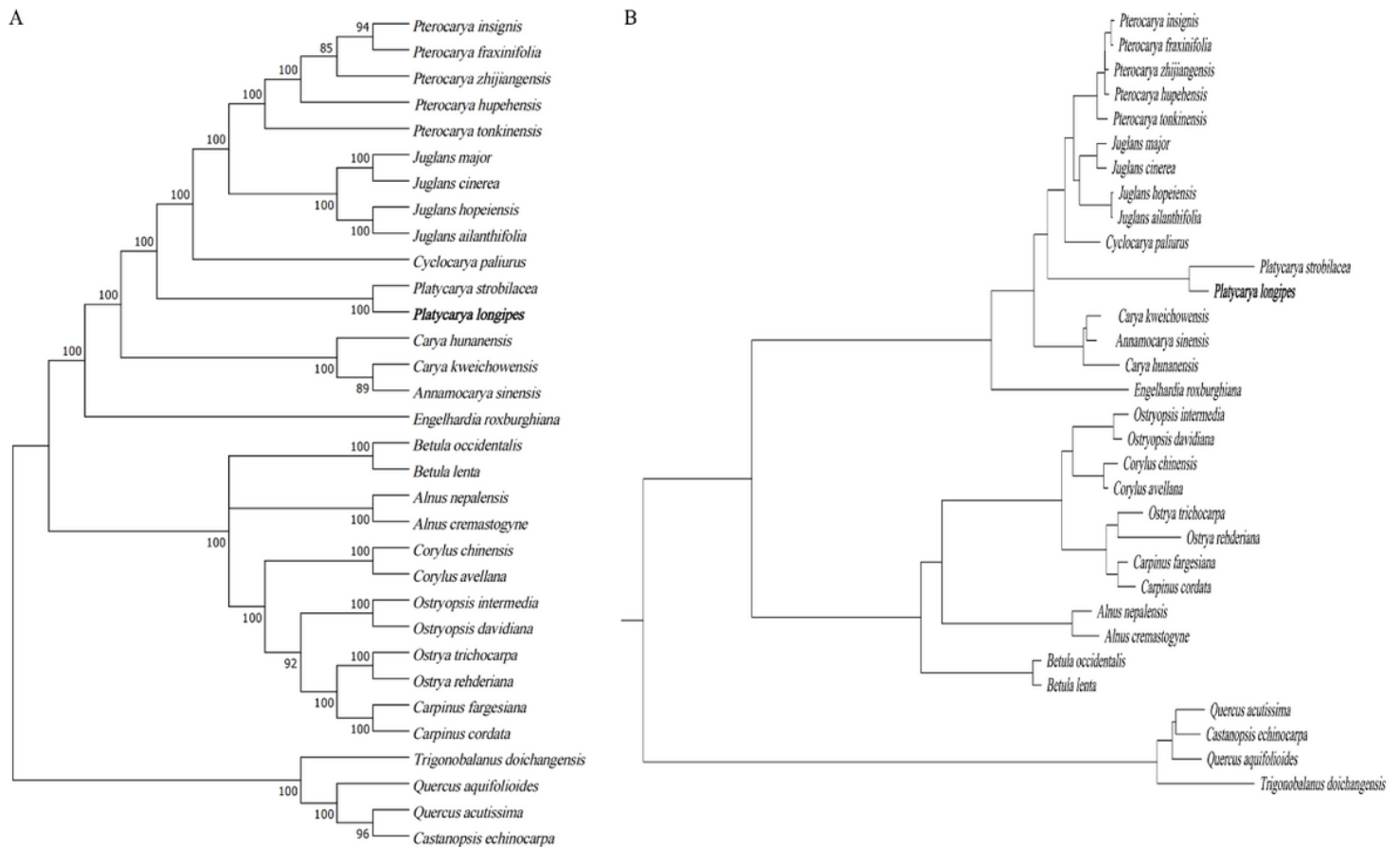


Figure 4

The Maximum Likelihood (ML) phylogenetic tree (A) and Bayesian Inference (BI) phylogenetic tree (B) were constructed based on complete chloroplast genome sequence of 32 species. 1,000 replicates were tested to confirm the stability of each tree node, numbers at the left of nodes are bootstrap support values. Four chains of the Markov Chain Monte Carlo were run each for 1,000,000 generations and were sampled every 100 generations.

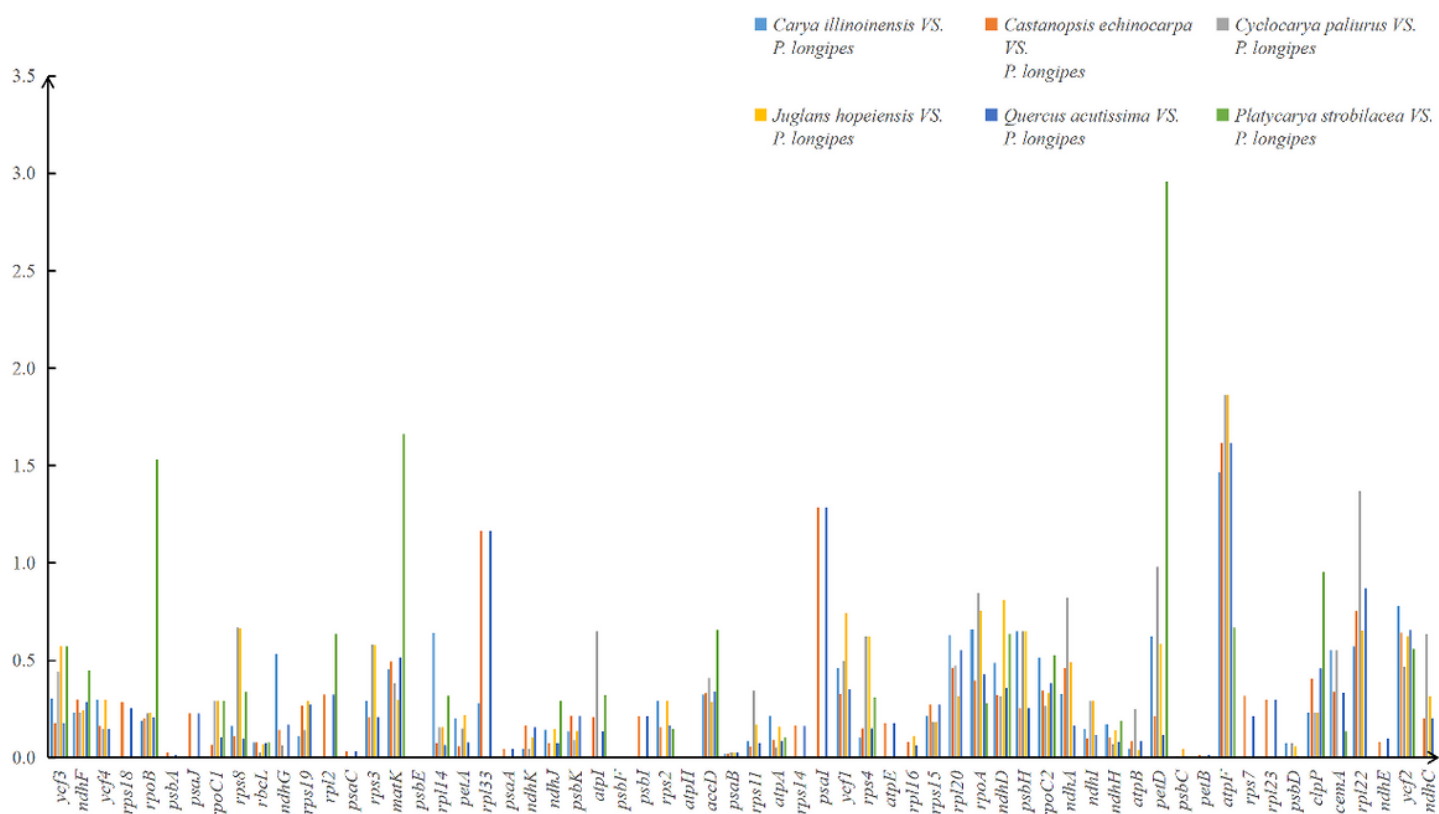


Figure 5

The Ka/Ks ratios of 64 protein-coding genes of the *P. longipes* cp genome versus six closely related species of Fagales.

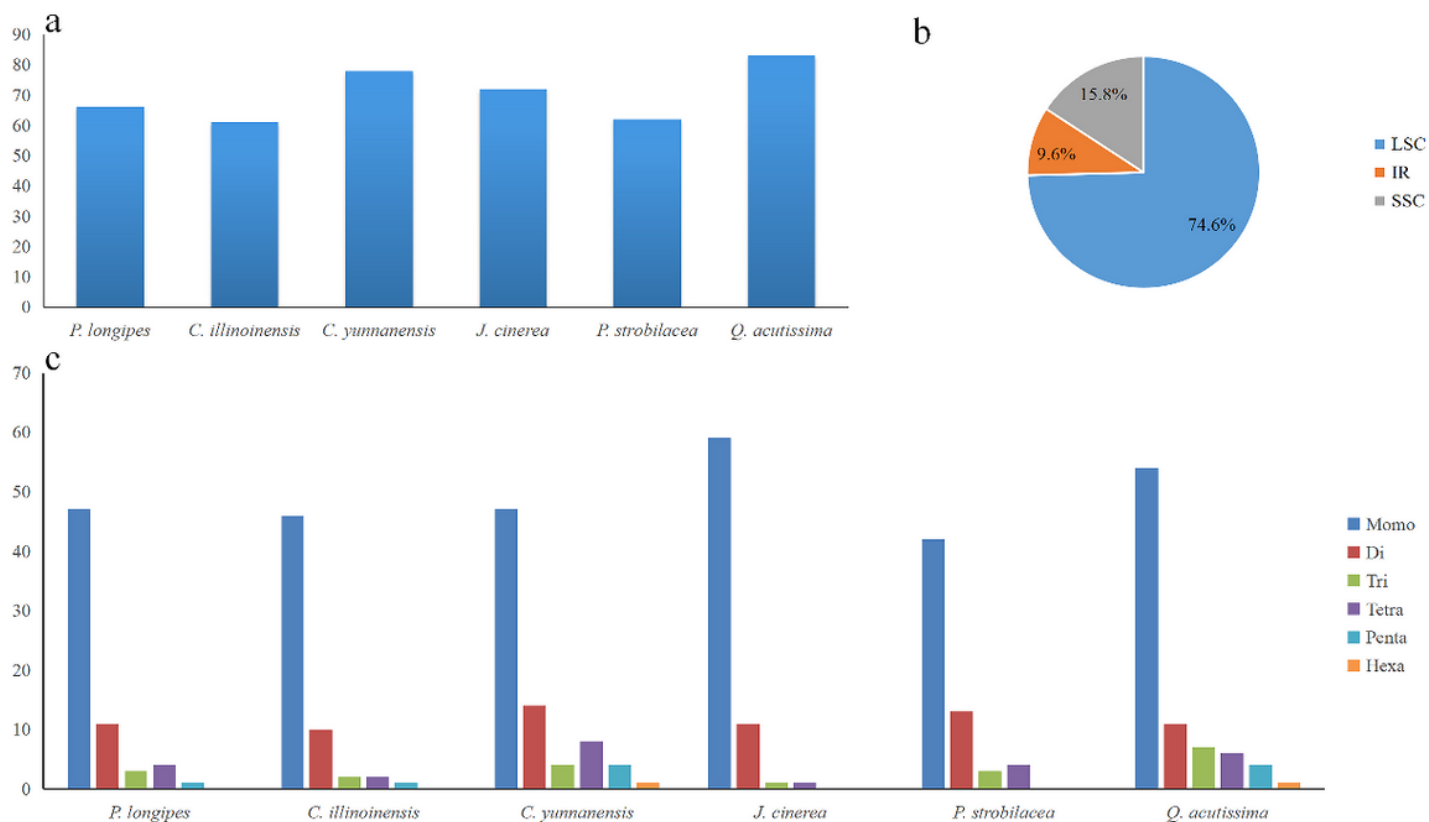


Figure 6

The type and size of simple sequence repeats among six chloroplast genomes. a. Numbers of SSRs detected in six Fagales chloroplast genomes, b. Frequencies of identified SSRs in LSC, IR and SSC regions, c. Numbers of SSR types detected in six Fagales chloroplast genomes.

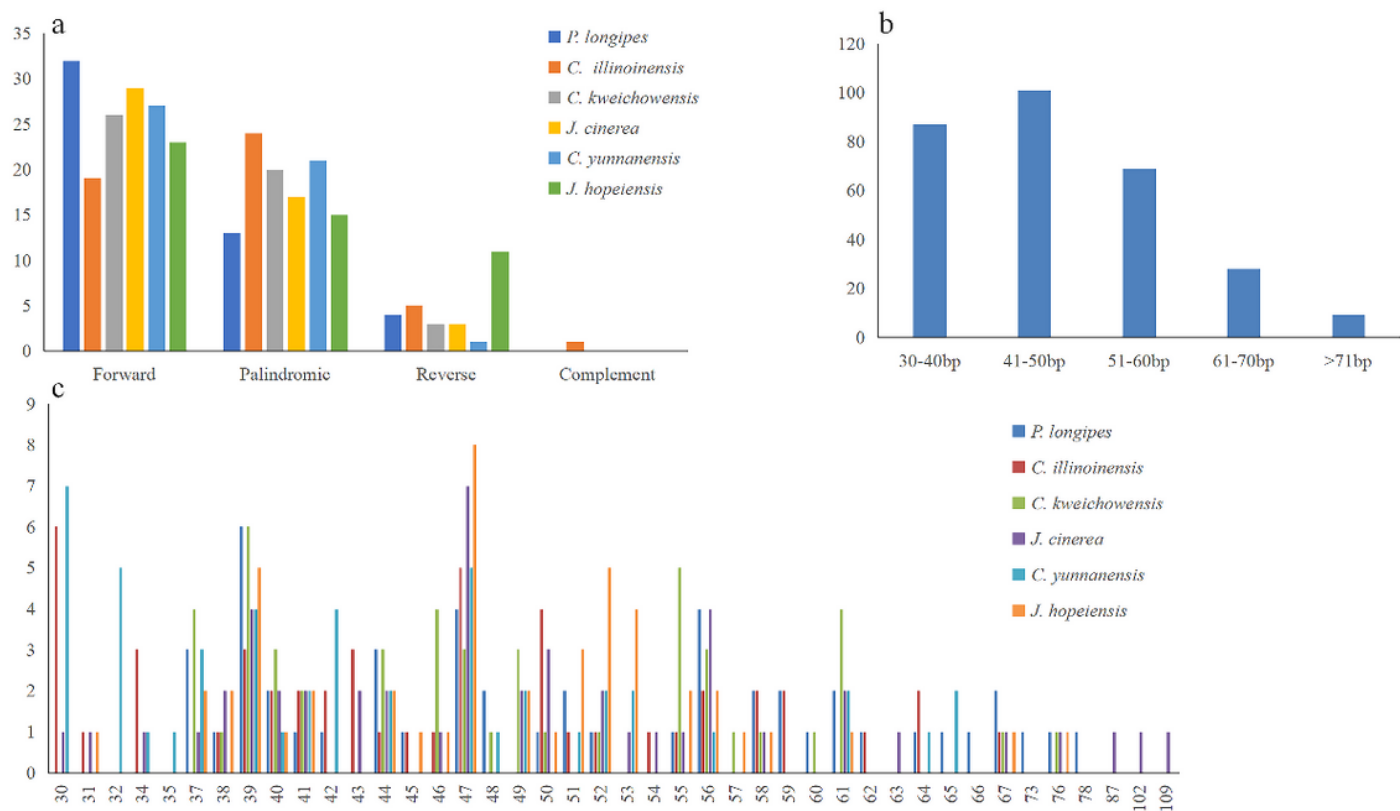


Figure 7

Analysis of long repeated sequences in the chloroplast genomes between *P. longipes* and other five Fagales species. a. frequency of repeat type, b,c. frequency of repeat length.

Supplementary Files

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

- [SupplementaryFiles.rar](#)