scientific reports

Check for updates

OPEN

The adaptive evolution of *Quercus* section *Ilex* using the chloroplast genomes of two threatened species

Yu-Ren Zhou[1,2,5], Yu Li[2,5], Liang-Hai Yang[2], Gregor Kozlowski[2,3,4], Li-Ta Yi[1], Mei-Hua Liu[1]✉, Si-Si Zheng[2]✉ & Yi-Gang Song[1,2]

Chloroplast (cp) genome sequences have been extensively used for phylogenetic and evolutionary analyses, as many have been sequenced in recent years. Identification of *Quercus* is challenging because many species overlap phenotypically owing to interspecific hybridization, introgression, and incomplete lineage sorting. Therefore, we wanted to gain a better understanding of this genus at the level of the maternally inherited chloroplast genome. Here, we sequenced, assembled, and annotated the cp genomes of the threatened *Quercus marlipoensis* (160,995 bp) and *Q. kingiana* (161,167 bp), and mined these genomes for repeat sequences and codon usage bias. Comparative genomic analyses, phylogenomics, and selection pressure analysis were also performed in these two threatened species along with other species of *Quercus*. We found that the guanine and cytosine content of the two cp genomes were similar. All 131 annotated genes, including 86 protein-coding genes, 37 transfer RNA genes, and 8 ribosomal RNA genes, had the same order in the two species. A strong A/T bias was detected in the base composition of simple sequence repeats. Among the 59 synonymous codons, the codon usage pattern of the cp genomes in these two species was more inclined toward the A/U ending. Comparative genomic analyses indicated that the cp genomes of *Quercus* section *Ilex* are highly conserved. We detected eight highly variable regions that could be used as molecular markers for species identification. The cp genome structure was consistent and different within and among the sections of *Quercus*. The phylogenetic analysis showed that section *Ilex* was not monophyletic and was divided into two groups, which were respectively nested with section *Cerris* and section *Cyclobalanopsis*. The two threatened species sequenced in this study were grouped into the section *Cyclobalanopsis*. In conclusion, the analyses of cp genomes of *Q. marlipoensis* and *Q. kingiana* promote further study of the taxonomy, phylogeny and evolution of these two threatened species and *Quercus*.

Keywords *Quercus*, Chloroplast (cp) genome, Comparative analyses, Phylogenetics, Positive selection

The genus *Quercus* L. (oak) comprises approximately 450 species worldwide, which is widely distributed in the Northern Hemisphere[1,2]. The innovative traits of *Quercus* and geoclimatic changes have facilitated its rapid spread and differentiation in the Northern Hemisphere over the last 56 million years[3–6]. *Quercus* is one of the largest woody genera and has important ecological, economic, and cultural functions[2,7]. Eight sections belonging to two subgenera were identified and documented in 2017[1,8]. The subgenus *Quercus* contains the *Lobatae*, *Ponticae*, *Protobalanus*, *Quercus*, and *Virentes* sections, whereas the subgenus *Cerris* contains the *Cerris*, *Cyclobalanopsis*, and *Ilex* sections[1].

Due to agriculture, biological resource-use, climate change, and other factors, an estimated 31% of oak species are threatened with extinction and 41% are of conservation concern[2]. Thus, the biological conservation of *Quercus* needs to be strengthened in terms of ecosystems, species, and genetic diversity. With the rapid development of high-throughput sequencing technology, increasing amounts of genomic information are available for use in conservation efforts[9–11].

Chloroplast (cp) genomes are much smaller than nuclear genomes, normally ranging from 115 to 165 kilobase pairs (kb), with a pair of inverted repeats (IRs) separating a large single-copy (LSC) region and a small

[1]College of Forestry and Biotechnology, Zhejiang A&F University, Lin'an 311300, Hangzhou, China. [2]Eastern China Conservation Centre for Wild Endangered Plant Resources, Shanghai Chenshan Botanical Garden, Shanghai 201602, China. [3]Department of Biology and Botanic Garden, University of Fribourg, 1700 Fribourg, Switzerland. [4]Natural History Museum Fribourg, 1700 Fribourg, Switzerland. [5]These authors contributed equally: Yu-Ren Zhou and Yu Li. ✉email: mhliu@zafu.edu.cn; zhengsisi1228@163.com

single-copy (SSC) region[12]. As they are non-recombinantly and uniparentally inherited and have low rate of nucleotide substitutions, cp genomes have been increasingly targeted to resolve the deep phylogeny of plants[13,14]. The exploration of gene rearrangements, structural changes, and repeat sequences has become an ideal strategy for species identification, population genetics, and genetic engineering[15].

The *Quercus* section *Ilex* originated along the East Tethys seaway during the middle Eocene, followed by climate change and the Himalayan orogeny, which dominating the formation of the current distribution pattern[16]. The *Quercus* section *Ilex* can adapt to the most diverse climatic ecosystems and has the highest genetic diversity among all sections in *Quercus*[17]. Till now, previous studies have sequenced and published the cp genomes of 16% of *Quercus* species (70 species in Table S1, available until December 2023), including 49 species of the subgenus *Cerris* (30%) and 21 of the subgenus *Quercus* (8%). Currently, 22 species for which the cp genome has been sequenced in section *Ilex*, accounting for the largest proportion of all sections. However, very few cp genome resources exist for species threatened with extinction and species of conservation concern in the genus *Quercus* (< 10%). Among section *Ilex*, only three of the eight threatened species have been sequenced (Table S1). Therefore, additional genetic and genomic resources need to be explored for these threatened species.

Quercus marlipoensis (critically endangered; CR) and *Q. kingiana* (endangered; EN) are threatened species distributed in tropical area of Southeast Asia[2]. Except for phylogenetic investigations that construct nuclear gene trees, no other studies on all the other aspects of these two threatened species have been conducted. Therefore, we sequenced and assembled the cp genomes of *Q. marlipoensis* and *Q. kingiana*, and performed the comparative analyses of all cp genomes of section *Ilex*. In this study, we aimed to (1) examine abundant simple sequence repeats (SSRs) and highly variable regions in the whole cp genomes of *Q. marlipoensis* and *Q. kingiana* to identify markers for phylogenetic and genetic studies; (2) study the consistency and difference of cp genome structure within and among sections of *Quercus* to provide new insights into the evolution of cp genomes in each section; and (3) perform phylogenetic reconstruction and selection pressure analysis for all cp genomes of section *Ilex* to understand the development and evolution of this section. Our findings are expected to enrich the molecular data for phylogenetic studies and conservation of endangered species in the *Quercus* section *Ilex*.

Results

Structural characteristics of cp genomes

The length of the cp genomes for *Q. marlipoensis* and *Q. kingiana* were 160,995 and 161,167 bp, respectively. The two cp genomes had the same circular quadripartite structure, comprising a long single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeat (IR) regions (Fig. 1). The GC content of the two cp genomes were 36.8% and 36.9%, respectively. Furthermore, the GC content of the IR region (approximately 42.7%) was significantly higher than that of the LSC and SSC regions (34.7% and 31.07%, respectively) (Table 1). In addition, the name, number, and order of annotated genes in the two cp genomes were consistent. The 131 annotated genes included 86 protein-coding genes (PCGs), 37 transfer RNA (tRNA) genes, and 8 ribosomal RNA (rRNA) genes (Tables 1 and 2). Remarkably, 18 genes (seven PCGs, seven tRNAs, and four rRNAs) were duplicated in the IR regions. Eighteen genes contained introns, including 12 PCGs and 6 tRNAs. Nine PCGs (*rps16*, *rpl2*, *rpl16*, *rpoC1*, *atpF*, *ndhA*, *ndhB*, *petB*, and *petD*) and six tRNAs (*trnA-UGC*, *trnG-UCC*, *trnI-GAU* , *trnK-UUU*, *trnL-UAA*, and *trnV-UAC*) contained one intron, whereas the remaining three PCGs (*ycf3*, *clpP1*, and *rps12*) contained two introns (Table 2).

Repeated sequences

In the two cp genomes, 113 (*Q. marlipoensis*) and 122 (*Q. kingiana*) SSRs were detected (Table S2). The number of mononucleotide repeats was the highest, accounting for approximately 68% of the total, while the number of pentanucleotide repeats was the lowest. There were a slightly more tetranucleotide than trinucleotide repeats. No hexanucleotide repeats were detected in either species (Fig. 2A). Further, the proportion of A/T bases in the SSRs was significantly higher than that of G/C, indicating a strong A/T bias (Fig. 2B). Additionally, from the distribution of SSRs, most of the SSRs were located in the LSC and inter-genic spacer (IGS) regions, while a few were located in the IR and gene regions (Fig. 2C,D).

Twelve and eight tandem repeats were also detected in the cp genomes of *Q. marlipoensis* and *Q. kingiana*, respectively (Table S3). The size of the basic repeat units mainly ranged between 20 and 29 bp (Fig. 3B). Among the four types of scattered repeats detected, the numbers of forward repeats (17 and 20) and palindromic repeats (19 and 24) were the highest, whereas the numbers of reverse repeats (3 and 4) and complementary repeats (0 and 1) were smaller in the *Q. marlipoensis* and *Q. kingiana* cp genomes (Fig. 3A). The scattered repeats ranged in size from 30 to 80 bp, mainly concentrated between 30 and 39 bp (Fig. 3C).

Codon usage *bias*

According to the results of CodonW analysis, 17,070 and 16,929 codons were encoded by the PCGs extracted from the *Q. marlipoensis* and *Q. kingiana* cp genomes, respectively. The GC content and codon preference indexes of the two cp genomes indicated a weak usage bias in the codons (Table 3). The relative synonymous codon usage (RSCU) analysis of the two cp genomes indicated that 30 of 59 synonymous codons had RSCU values > 1, with only two codons ending in G/C (UUG and UCC) and the remaining 28 codons ending in A/U (Table S4). The results showed that the codon usage pattern of the cp genomes was inclined toward A/U endings. In addition, the RSCU value of the codons encoding leucine (Leu) was the highest (Fig. 4).

We also analyzed the factors influencing codon usage bias. In the PR2-bias-plot analysis, the four bases of the codon at the third position were unevenly distributed in the four regions of the PR2-bias-plot (A3 ≠ T3, G3 ≠ C3), indicating that codon usage bias varied with base position (Fig. 5A,B). In the neutrality-plot analysis, the GC12 and GC3 values were positively correlated, but the correlation was not significant. This suggested that the codon

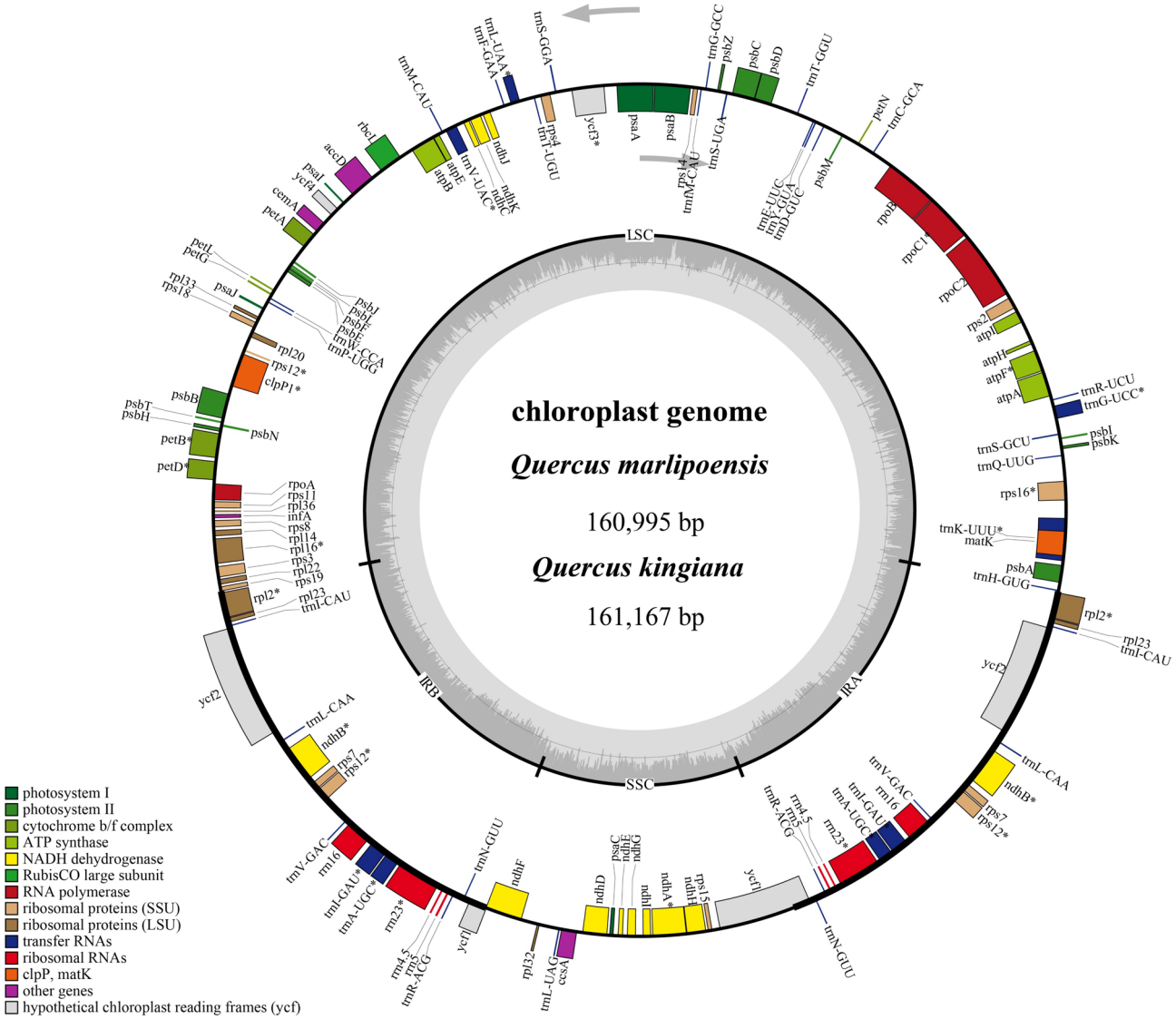

Fig. 1. Gene map of the cp genomes of *Q. marlipoensis* and *Q. kingiana*. The outermost circle represents the genes annotated in the chloroplast genomes. Genes outside the circle are transcribed in the counter-clockwise direction, whereas those inside the circle are transcribed in the clockwise direction. Genes of different functional groups are identified by color. The length and boundary of the LSC, SSC, and two IR regions are indicated in the inner circle. Inside the inner circle, the dark gray area indicates GC content while the lighter gray corresponds to the AT content of the genome. *LSC* large single copy, *SSC* small single copy, *IR* inverted repeat.

Species	Length (bp)				GC content (%)				Gene number			
	Total	LSC	IR	SSC	Total	LSC	IR	SSC	Gene	PCG	rRNA	tRNA
Q. marlipoensis	160,995	90,318	25,875	18,927	36.8	34.7	42.72	31.07	131	86	8	37
Q. kingiana	161,167	90,538	25,846	18,937	36.9	34.7	42.73	31.0	131	86	8	37

Table 1. Basic features of the cp genomes of *Q. marlipoensis* and *Q. kingiana*. *LSC* large single copy, *SSC* small single copy, *IR* inverted repeat, *PCG* protein-coding gene, *tRNA* transfer RNA gene, *rRNA* ribosomal RNA gene.

usage bias was influenced more by natural selection than other factors (Fig. 5C,D). Additionally, in the ENC-plot analysis, some genes were distributed along or near the standard curve, while others were located farther below the curve, indicating that codon usage bias was affected by natural selection and mutations (Fig. 5E,F). Through the above three analyses, we found that the codon usage bias of cp genomes was affected by base mutations, natural selection, and other factors.

Category	Group	Name
Transcription and translation	Translational initiation factor	*infA*
	Ribosomal RNAs	*rrn16*(×2), *rrn23*(×2), *rrn4.5*(×2), *rrn5*(×2)
	Transfer RNAs	*trnA-UGC**(×2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC*, *trnG-UCC**, *trnH-GUG*, *trnI-CAU*(×2), *trnI-GAU**(×2), *trnK-UUU**, *trnL-CAA*(×2), *trnL-UAA**, *trnL-UAG*, *trnM-CAU*, *trnN-GUU*(×2), *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*(×2), *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*(×2), *trnV-UAC**, *trnW-CCA*, *trnY-GUA*, *trnfM-CAU*
	Small subunit of ribosome (SSU)	*rps11*, *rps12***(×2), *rps14*, *rps15*, *rps16**, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*(×2), *rps8*
	Large subunit of ribosome (LSU)	*rpl14*, *rpl16**, *rpl2**(×2), *rpl20*, *rpl22*, *rpl23*(×2), *rpl32*, *rpl33*, *rpl36*
	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
Photosynthesis	Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbT*, *psbZ*, *psbN*
	Subunit of cytochrome	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	RubisCO large subunit	*rbcL*
	NADH dehydrogenase	*ndhA**, *ndhB**(×2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Biosynthesis	Maturase	*matK*
	ATP-dependent Protease	*clpP1***
	Acetyl-CoA-carboxylase	*accD*
	Envelop membrane protein	*cemA*
	C-Type cytochrome synthesis	*ccsA*
Unknown	Hypothetical chloroplast reading frames(*ycf*)	*ycf4*, *ycf3***, *ycf1*(×2), *ycf2*(×2)

Table 2. Genetic classification of the cp genomes. The asterisk symbol is used to indicate genes with single (*) or double (**) introns. The duplicated genes located in IR regions are marked as (×2).

Comparative chloroplast genomic analyses of *Quercus* section *Ilex*

The online software IRscope (https://irscope.shinyapps.io/irapp/) was used to assess the expansion and contraction of IR regions in 19 species of *Quercus* section *Ilex*. The junction of LSC/IRb (JLB) was located in the IGS of the *rps19* and *rpl2* genes. The *rps19* gene in most species was shifted 11 bp away from the JLB, except for *Q. semecarpifolia*, which had a 12 bp shift. Both SSC/IRb (JSB) and SSC/IRa (JSA) boundaries were located within the *ycf1* gene. The *ycf1* gene at the JSB boundary expanded (2–93 bp), while the *ycf1* gene at the JSA boundary contracted (4567 to 4636 bp). In addition, *ndhF* gene was distributed at the JSB boundary in *Q. acrodonta*, *Q. bawanglingensis*, and *Q. franchetii*; where it contracted to the IRb region by 22–28 bp (Fig. 6).

The cp genomes of *Quercus* section *Ilex* were highly conserved. The genome structures and gene sequences of cp genomes had strong collinearity without gene rearrangement or inversion (Fig. S1). However, sequences in the non-coding regions were more variable than those in the coding regions. In addition, sequence variability in the SC regions was significantly higher than that in the IR regions. Further, we found high variability in the exon regions of three PCGs (*rpoC2*, *rps19*, and *ycf1*) and in the conserved non-coding segments of four IGS regions (*rps16/psbK*, *psbM/trnD-GUC*, *psbZ/trnG-UCC*, and *rpl32/trnL-UAG*) (Fig. S2).

Within *Quercus* section *Ilex*, polymorphic site analysis detected 2437 variable sites, including 1174 singleton variable sites and 1263 parsimony informative sites. We also calculated nucleotide diversity (Pi) values to determine the diversity level of the cp genomes. The Pi values ranged from 0 to 0.01673 with the mean value of 0.00392. We detected eight highly variable regions with Pi values greater than 0.01, with the highest Pi value of 0.01673 occurring in the IGS region between *accD* and *psaI* genes (Fig. 7).

Comparison of chloroplast genome among different sections of *Quercus*

To study the consistency and difference of cp genome in each section from the perspective of genome structure and nucleotide diversity, we compared the cp genomes of the four sections of *Quercus*, including 19 of section *Quercus*, 37 of section *Cyclobalanopsis*, three of section *Cerris*, and 26 of section *Ilex*.

The expansion and contraction of IR regions were analyzed respectively in the four sections (Fig. S3–6). The results showed that four genes (*rps19*, *ycf1*-JSB, *ycf1*-JSA, and *trnH*) were located at four junctions and affected the expansion and contraction of IR regions. The boundary positions of the four boundary genes were exactly the same of the three species in section *Cerris*, which indicated that there was no contraction and expansion of IR regions in section *Cerris* (Fig. S5). There were differences in the boundary positions of the four boundary genes in the other three sections, among which section *Ilex* had the highest number of expansion and contraction types (Fig. S7). The *rps19* and *trnH* genes in section *Quercus* and *Ilex* were contracted more significantly than section *Cyclobalanopsis*. While the *ycf1*_JSB and *ycf1*_JSA genes in section *Cyclobalanopsis* were expanded more significantly than the other three sections (Fig. 8).

Nucleotide diversity analysis was also performed for these four sections. After polymorphic site detection, it was found that section *Ilex* had the most singleton variable sites among the four sections, showing the highest level of sequence diversity. However, section *Cerris* was the most conservative and contained the fewest variation sites (Table 4). The sliding window analysis obtained the nucleotide diversity value (Pi) between the whole

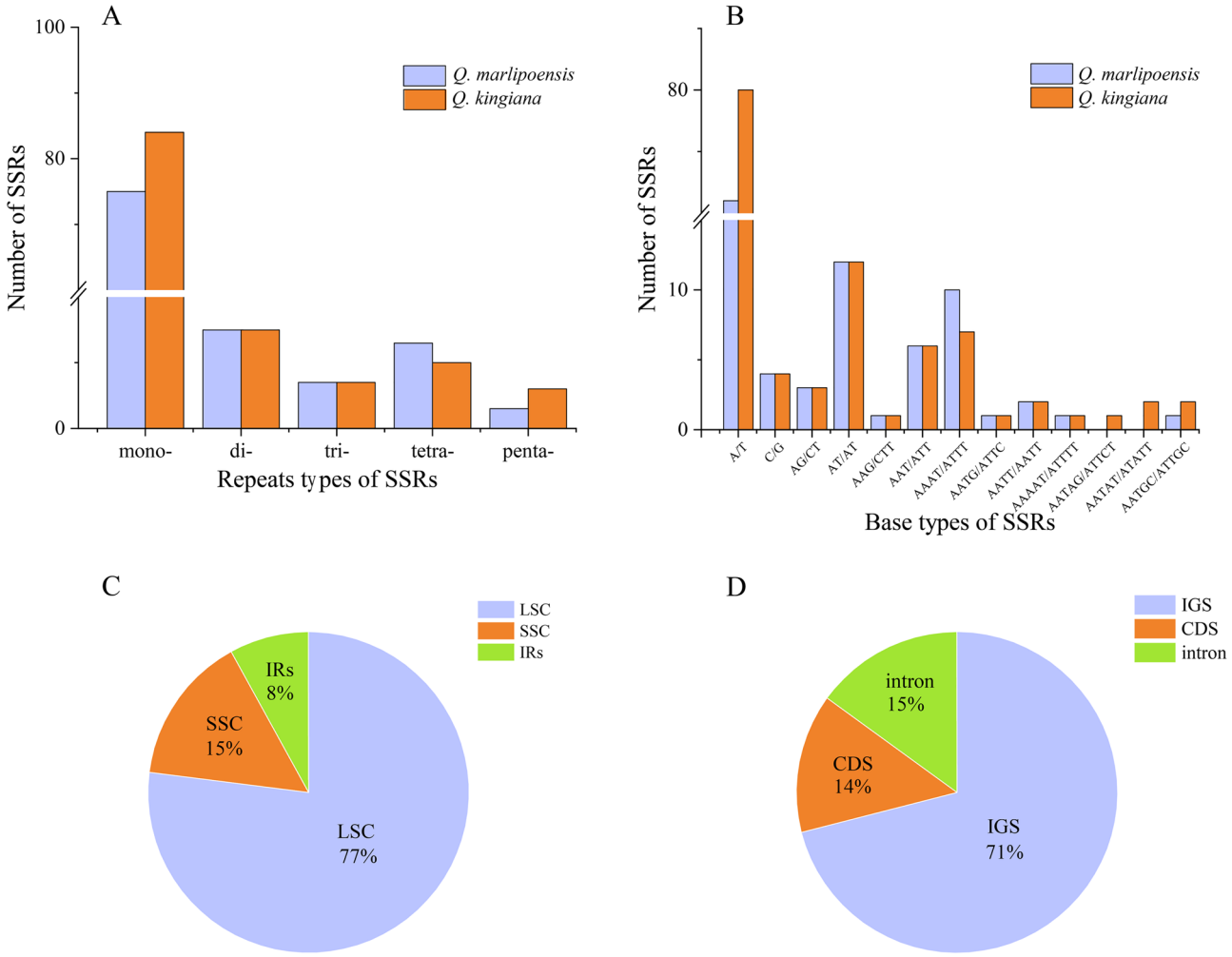

Fig. 2. Type, distribution, and numbers of SSRs in the cp genomes of *Q. marlipoensis* and *Q. kingiana*. (**A**) The number of SSRs with six repeated types (mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides); (**B**) the number of SSRs with different base types; (**C** and **D**) the total number of SSRs in different regions of the two cp genomes. *LSC* large single copy, *SSC* small single copy, *IRs* inverted repeats, *IGS* inter-genic spacer, *CDS* coding sequence.

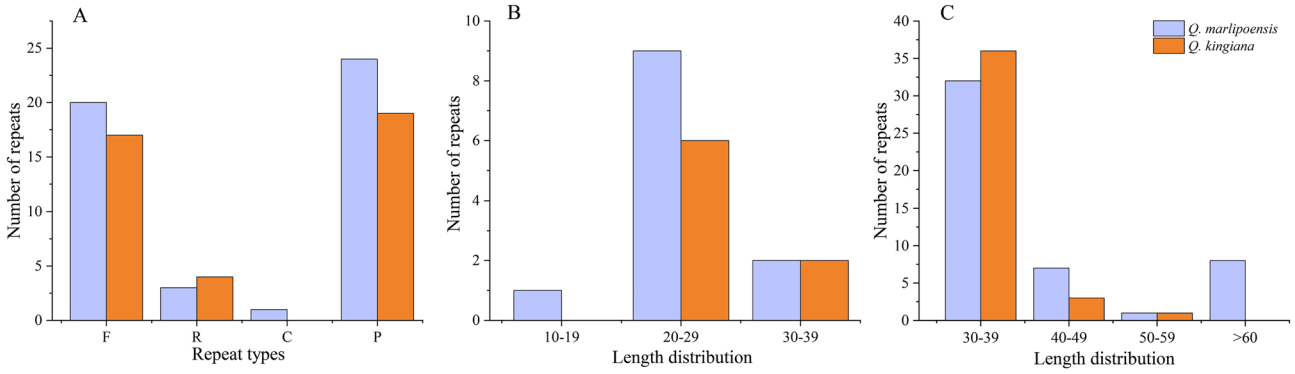

Fig. 3. The distribution of the number and length of tandem and scattered repeats in the cp genomes of *Q. marlipoensis* and *Q. kingiana*. (**A**) Numbers of scattered repeats in the two cp genomes; (**B**) The length distribution of tandem repeats; (**C**) The length distribution of scattered repeats. F: forward repeats; R: reverse repeats; C: complementary repeats; P: palindromic repeats.

Species	GC1 (%)	GC2 (%)	GC3 (%)	GC_all (%)	GC3s (%)	ENC	CAI	CBI	FOP	No. of codons
Q. marlipoensis	46.54	37.79	28.80	37.71	25.80	49.21	0.168	-0.098	0.355	17,070
Q. kingiana	46.62	37.82	28.89	37.78	25.90	49.26	0.168	-0.097	0.355	16,929

Table 3. Codon parameter characterization of the cp genomes of *Q. marlipoensis* and *Q. kingiana*.

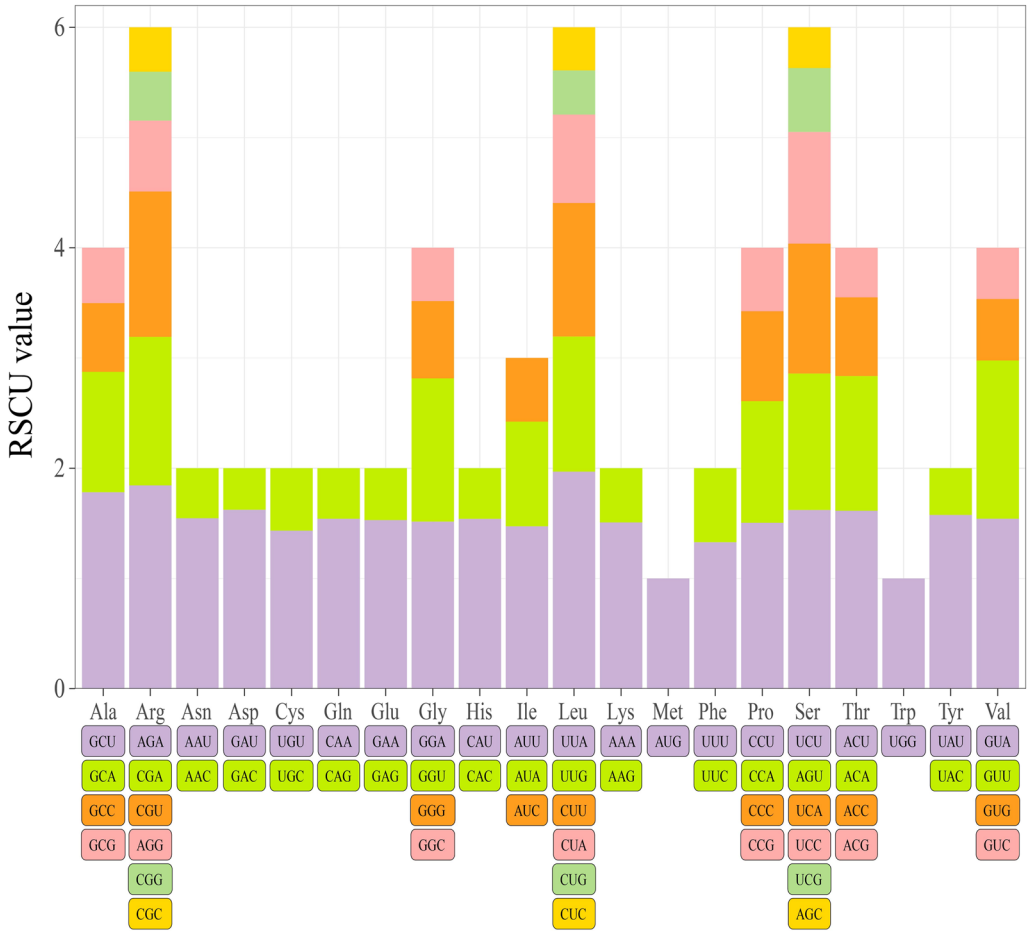

Fig. 4. Relative synonymous codon usage (RSCU) analysis of the cp genomes of *Q. marlipoensis* and *Q. kingiana*.

genome sequences of each section, and it was found that section *Ilex* had the highest mean value, followed by section *Quercus* (Fig. 9).

Phylogenetic analysis

Using the whole cp genome data, the phylogenetic relationship of the two newly sequenced cp genomes in this study and their closely related species in the *Quercus* genus was reconstructed. A total of 33 taxa were used to reconstruct the phylogenetic tree, and the majority of branches had high bootstrap support values (Fig. 10). First, two distinct groups were recognized among all *Quercus* species, with the subgenera *Quercus* and *Cerris*. In the subgenus *Cerris*, the first clade was composed of section *Cerris* and part of section *Ilex*, whereas the second clade comprised section *Cyclobalanopsis* and another part of section *Ilex*. *Quercus marlipoensis* and *Q. kingiana* were nested within the second clade. *Quercus marlipoensis* had a close relationship with species from Himalayan subalpine areas. However, *Quercus kingiana* was more closely related to section *Cyclobalanopsis*.

Selective pressure analysis

The selection pressure of common PCGs in the cp genomes of *Quercus* section *Ilex* was detected using the site model of the PAML v4.9j software. Using the ratio of dN (non-synonymous substitution rate) to dS (synonymous substitution rate), the evolution rates of the PCGs were calculated. We detected 27 and 36 PCGs with positive selection sites at M2 and M8 site models, respectively. Then, the likelihood ratio test (LRT) was performed for

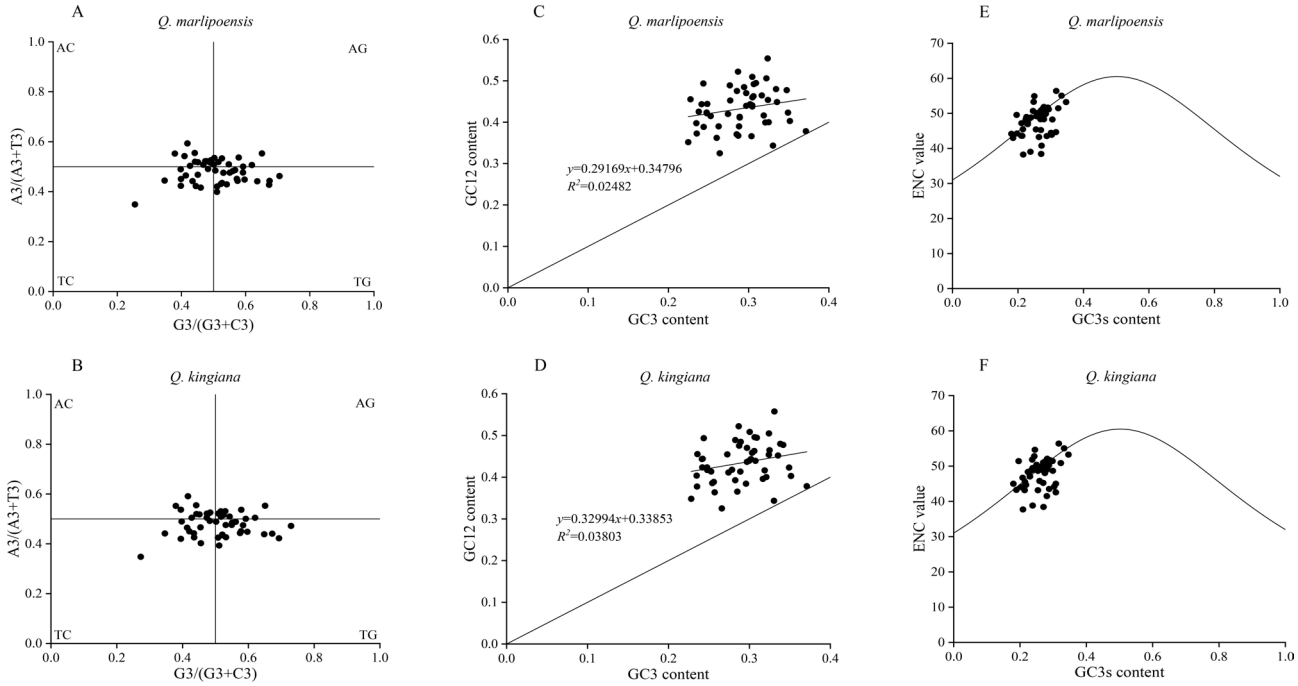

Fig. 5. PR2-bias-plot analysis (**A** and **B**); Neutrality-plot analysis (**C** and **D**); and ENC-plot analysis (**E** and **F**) of the cp genomes of *Q. marlipoensis* and *Q. kingiana*.

these 36 genes, where those with $P < 0.05$ were positively selected genes. A total of 11 PCGs were positively selected: *ccsA*, *ndhB*, *ndhD*, *ndhF*, *rbcL*, *rpl22*, *rpl32*, *rpoC2*, *rps12*, *ycf1*, and *ycf2* (Table 5). Subsequently, based on the Bayes empirical Bayes (BEB) algorithm, the 11 PCGs had 103 positive selection sites (Table S5).

Discussion

Chloroplast (Cp) genome characterization of newly sequenced species

High-throughput sequencing technology has accelerated chloroplast genomics research. However, less than 20% of *Quercus* species have completed chloroplast genome studies, and the data is still insufficient. In the current study, we assembled the complete cp genomes of *Q. marlipoensis* and *Q. kingiana*. Similar to other *Quercus* species, these two cp genomes exhibited a quadripartite circular structure[18–20]. In addition, their cp genome content and composition were similar to those of previously sequenced cp genomes, indicating high conservation across *Quercus* cp genomes[21,22]. However, the length of cp genomes differed between the two species. Such differences are predominantly caused by IR contraction and expansion, a common evolutionary phenomenon noticed across the cp genomes of angiosperms[23–25].

Repeat sequences exist widely in plant genomes and have likely contributed to plant evolution through repeatability of IR regions[26,27]. We detected 235 SSRs in both species, which were largely composed of A/T repeat units. This observation is consistent with previous reports on the AT richness of cp genomes. Indeed, SSRs serve as valuable molecular markers and have been extensively studied across various research fields, including ecology, evolution, population genetics, and polymorphism studies at both species and population levels[28–30]. Our observation of most SSRs being distributed in the IGS and LSC regions is consistent with similar findings in other angiosperm lineages[21,31].

GC content is a major factor in codon usage bias and may play an important role in the evolution of genome structure[32]. In our study, 30 out of the 59 synonymous codons exhibited RSCU values greater than 1. The third base of the codons was biased towards A/U, a phenomenon that was commonly noticed in angiosperms[33–35]. Codon usage bias is a natural phenomenon that is caused by mutations, selection, genomic composition, and other related factors[36].

Sequence differences and evolution of *Quercus*

IR regions are crucial for stabilizing cp genome structure and are highly conserved across angiosperms. Variation in angiosperm cp genome length is primarily attributed to the expansion and contraction of IRs[37–39], highlighting the importance of their analysis in evolutionary studies[40]. By comparing the four boundary regions of the cp genomes across 19 species of *Quercus* section *Ilex*, we found a consistent gene distribution and size, indicating strong uniformity in this group. In addition, the cp genomes in the *Quercus* section *Ilex* displayed good collinearity in the current study.

Both mVISTA sequence variation and nucleotide diversity analyses showed differences in the degree of variation among different regions of the cp genomes. Specifically, the SC regions exhibited higher variability compared to that in the IR regions. This variability may be caused by the presence of more conserved rRNA genes in the IR regions. Further, the IGS regions showed higher variations than that in the coding regions, a

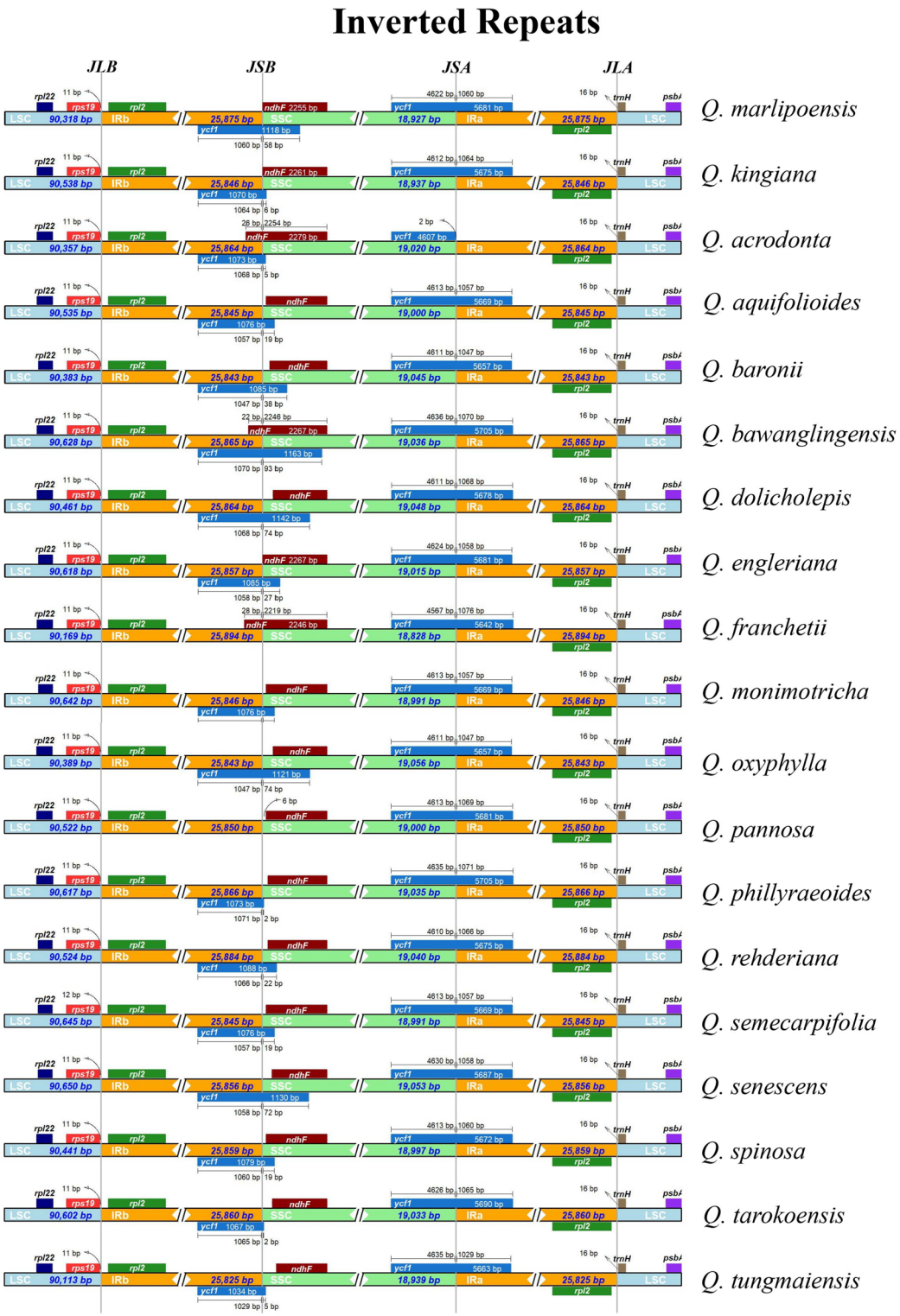

Fig. 6. Comparison of the junction regions (JLA, JLB, JSB, JSA) among 19 cp genomes of *Quercus* section *Ilex*. The boundary genes are denoted with colored boxes. The number above the gene boxes indicates the distance between the end of the gene and the junction regions (JLA, JLB, JSB, JSA).

pattern consistent with observations in other *Quercus* species[18,41]. Importantly, these highly variable regions can be used to develop DNA barcodes for species identification and systematic classification[42]. The *ycf1* gene and two IGS regions, *trnH*/*psbA* and *trnK*/*rps16*, have previously been identified as practical barcodes for plants[31,43,44].

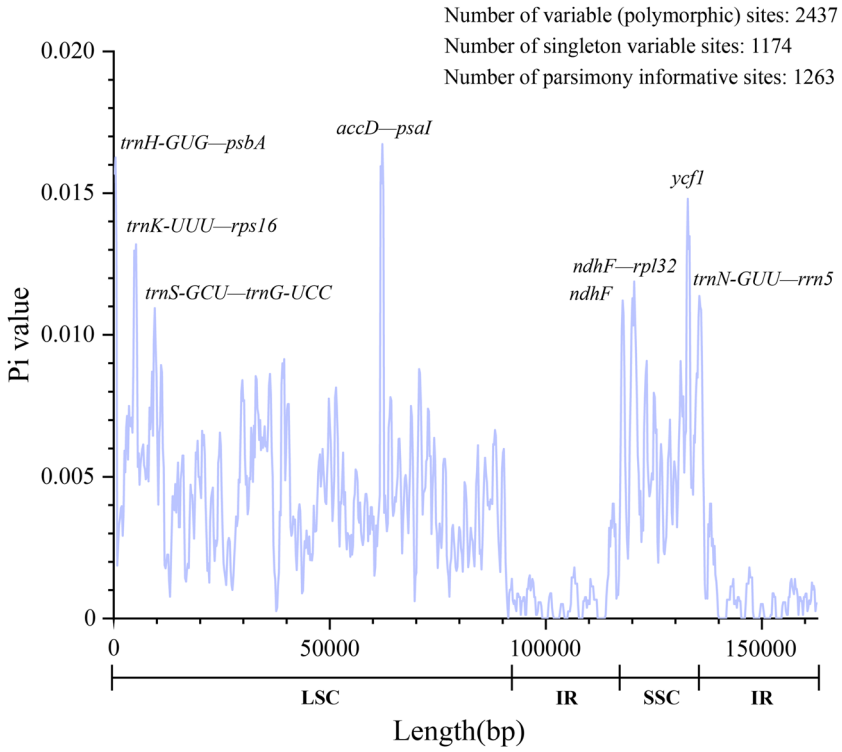

Fig. 7. Sliding window analysis of 19 cp genomes of *Quercus* section *Ilex* (window length: 800 bp; step size: 200 bp). The X-axis represents the nucleotide position of the window's mid-point, and the Y-axis represents the nucleotide diversity (Pi) value per window.

The cp genome structure of the four sections of *Quercus* was compared, and it was found that there were conservation and differences within and among the four sections. Within section *Cerris*, cp genomes exhibited high conservation, most likely due to their close phylogenetic relationships[45]. However, the differences within other sections may be the result of adaptations between species to the local environment. Of course, a small percentage of the differences may be due to personal errors encountered in assembly and annotation, which require us to proofread manually. There were also significant differences in cp genome structure among different sections. The expansion and contraction of boundary genes of IR regions in section *Cyclobalanopsis* were mostly significantly different from those in other sections. This may be due to the fact that this section contained more species, so the structural differences between sections still need to be further explored by expanding the number of species. Nucleotide diversity analysis showed that section *Ilex* had the highest nucleotide diversity, which was consistent with other studies in chloroplast markers[17,46].

Phylogeny of *Quercus* section *Ilex*

Due to complex evolutionary problems such as convergent evolution and hybridization introgressions, studies on the phylogenetic relationships of *Quercus* face great challenges[47–49]. The cp genomes are valuable tools for understanding phylogenetic relationships in angiosperms[50]. In our study, we analyzed cp genomes from 32 *Quercus* species to reconstruct their phylogenetic relationships, which were inconsistent with classification systems based on nuclear markers[1,51]. The different evolutionary patterns revealed by the oak chloroplast and nuclear genomes are likely due to historical introgression of ancestral lineages and recent or ongoing gene flow between closely related species[46]. Notably, sections *Cyclobalanopsis* and *Cerris* were found within the section *Ilex.* This finding suggests possible introgression among ancestral taxa or chloroplast capture. Similar phenomena have been observed in white oaks[52] and South American *Nothofagus*[53], attributed to cytoplasmic nuclear inconsistencies. A similar unresolved complex phylogenic relationship within subgenus *Cerris* was also found in other studies using cpDNA markers[46]. The chloroplast capture events within a genus are mostly due to hybridization[53,54]. However, the hybridization between the extant sections in oaks is extremely rare in the wild[55]. Although the phylogenetic relationships of the *Quercus* section *Ilex* are complex, the development of sequencing techniques will allow for the inclusion of more taxa and samples in future studies. Such progress will facilitate further exploration of the interspecific relationships and phylogenomics of the *Quercus* section *Ilex*. These results are likely to further our understanding of taxonomy, phylogenetic evolution, and conservation efforts related to Chinese *Quercus* species.

Adaptive evolution of *Quercus* section *Ilex*

The *Quercus* section *Ilex* is widely distributed in habitats across humid to semi-arid regions in the Eurasian tropics and subtropics, making it an important regional vegetation group[56,57]. While the genetic homogeneity

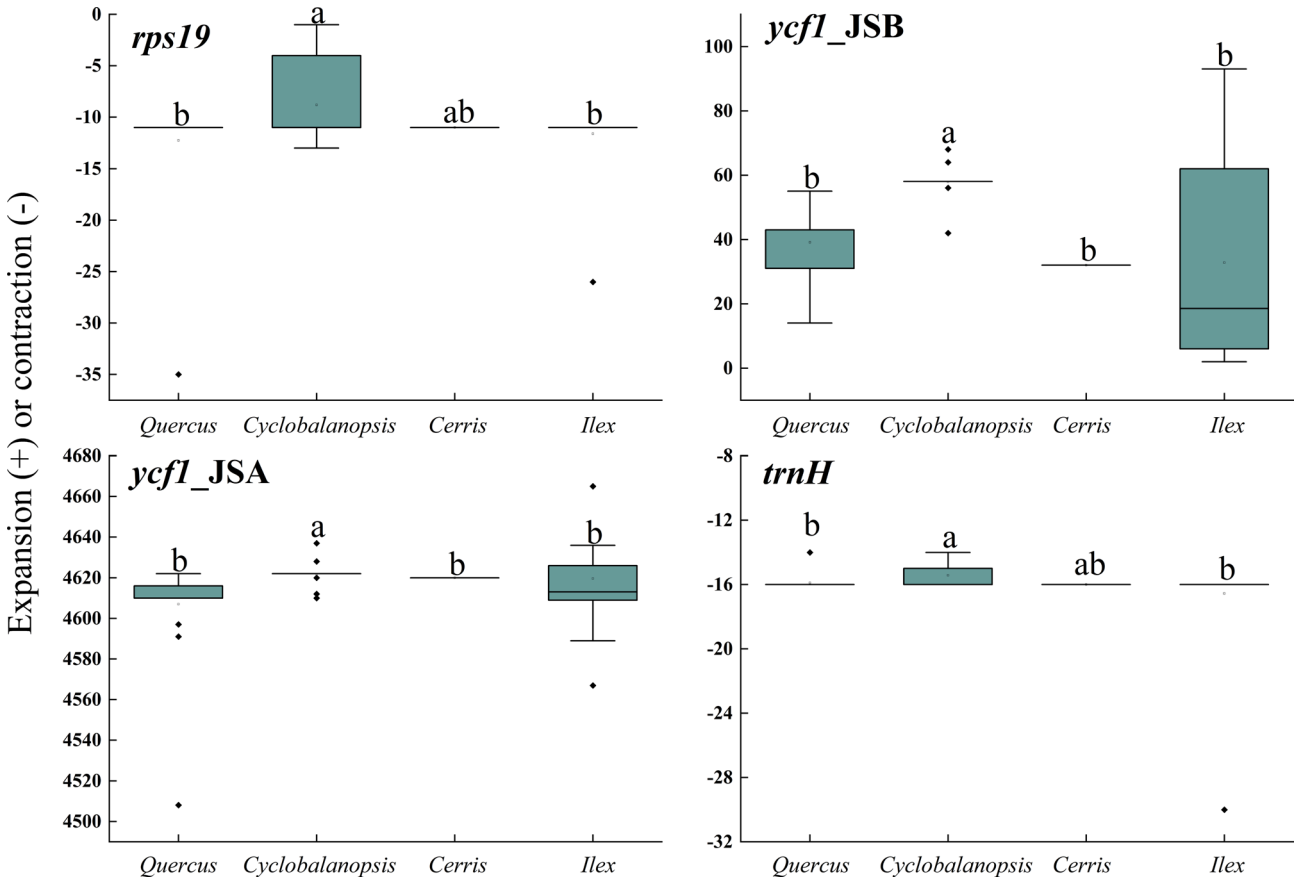

Fig. 8. The box plots of expansion and contraction of the four boundary genes in the four sections of *Quercus* (section *Quercus*, *Cyclobalanopsis*, *Cerris*, and *Ilex*).

Polymorphic sites	*Quercus*	*Cyclobalanopsis*	*Cerris*	*Ilex*
Numbers of species	19	37	3	26
Numbers of sites	165,567	162,455	161,176	163,647
Numbers of singleton variable sites	602	709	70	1150
Numbers of parsimony informative sites	7853	215	0	1526

Table 4. The distribution of singleton variable sites and parsimony informative sites in the four sections of *Quercus*.

of section *Cyclobalanopsis* implicates selection as a driving force in its plastid evolution, section *Ilex* exhibits the hallmarks of geographic isolation[17]. At the cp genome level, 11 protein-coding genes (PCGs) were positively selected during evolution, which may be related to environmental suitability. Four of these are photosynthesis-related genes, including three NADH dehydrogenase genes (*ndhB*, *ndhD*, and *ndhF*) and the RubisCO large subunit (*rbcL*). These genes are known to be adept to light environmental stress and are commonly positively selected across many species[58–60]. Further, four genes are involved in the transcription and translation processes: *rpl22*, *rpl32*, *rps12*, and *rpoC2*. The first three encode ribosomal proteins and control protein synthesis, while *rpoC2* encodes RNA polymerase and is essential for its function. The *ccsA* is responsible for C-type cytochrome synthesis, is thought to be linked to the binding of the C-type cytochrome and heme, potentially driving adaptive evolution[61]. Finally, *ycf1* and *ycf2* are the two largest hypothetical chloroplast reading frames within the cp genome. While their encoded products are believed to be crucial for chloroplast function[62], their specific roles and their evolutionary significance remain unclear. These genes serve as a valuable gene set for further investigation into the mechanisms of adaptive evolution in Chinese *Quercus* species[63].

Conclusions

In this study, the cp genomes of two threatened section *Ilex* species were described and compared with those of other *Quercus* species in NCBI. The results showed that the *Quercus* cp genomes were similar in quadripartite structure, GC content, codon usage features and gene order. Despite significant conserved overall cp genome

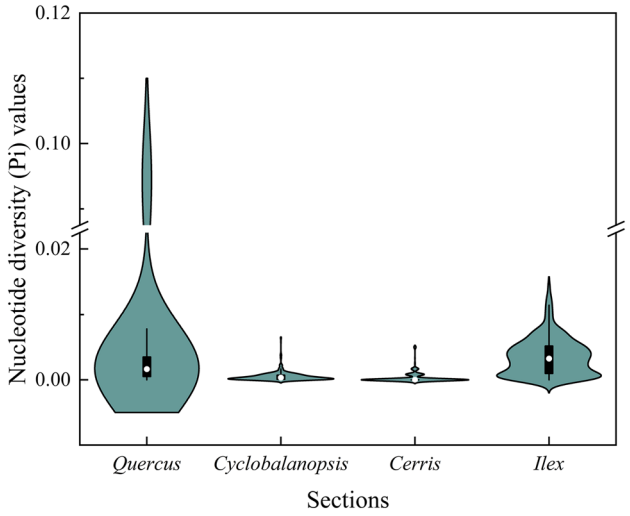

Fig. 9. Nucleotide diversity values of cp genomes in the four sections of *Quercus* (section *Quercus*, *Cyclobalanopsis*, *Cerris*, and *Ilex*).

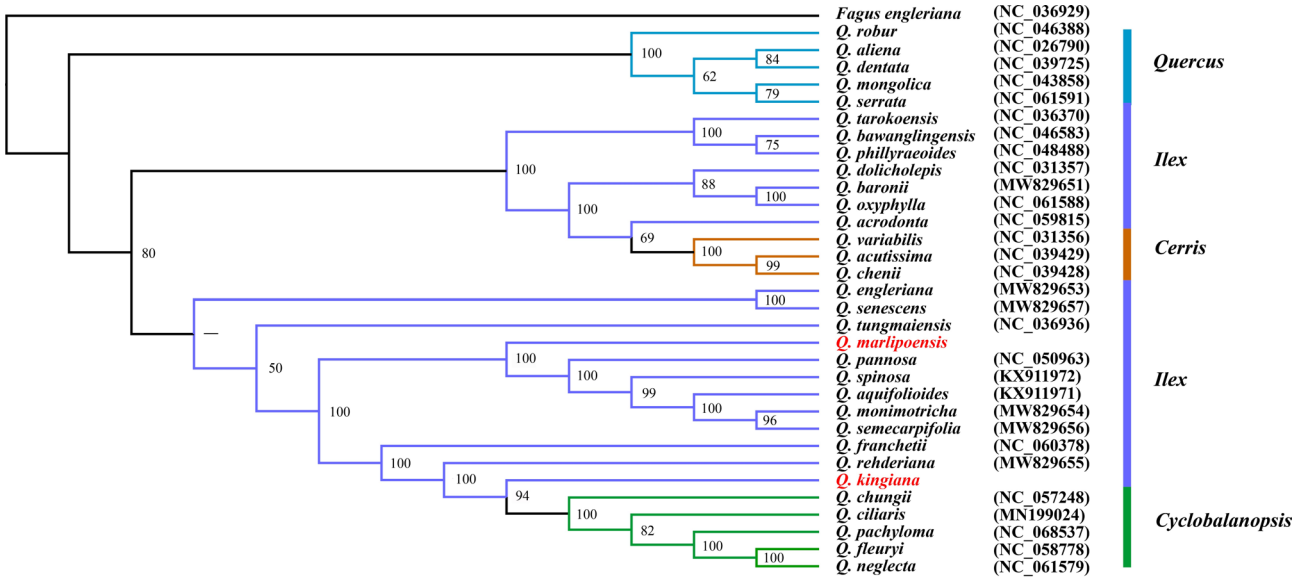

Fig. 10. Maximum Likelihood (ML) phylogenetic tree among 32 cp genomes of *Quercus* species and an outgroup *Fagus engleriana*. Values above the branch represent bootstrap support, where BS less than 50% is represented by a "—".

structure and gene content, significant sequence differences were found in alternating regions of these genomes. Seven highly divergent regions (*rpoC2*, *rps19*, *ycf1*, *rps16/psbK*, *psbM/trnD-GUC*, *psbZ/trnG-UCC*, and *rpl32/trnL-UAG*) might be used as phylogenetic molecular markers. The nine positively selected sites identified in this analysis included photosynthesis-related genes (*ndhB*, *ndhD*, *ndhF*, and *rbcL*), transcription and transcript processing genes (*rpl22*, *rpl32*, *rps12*, and *rpoC2*) and could facilitate understanding of the adaptive evolution of *Quercus* genus. Overall, the data obtained will contribute to further studies on the diversity, ecology, taxonomy, phylogenetic evolution and conservation of *Quercus* genus.

Methods

Plant material, DNA extraction, and sequencing

Tender, healthy leaf samples of *Q. marlipoensis* (98°93′E, 18°79′N; Altitude, 2250 m) and *Q. kingiana* (104°83′E, 23°15′N; Altitude, 1789 m) were harvested from Malipo county and Simao city in Yunnan province, China. Silica gel was used to desiccate the collected material. Voucher specimens were deposited in the herbarium of the Shanghai Chenshan Botanical Garden. High-quality DNA was extracted using a modified cetyltrimethyl ammonium bromide (CTAB) protocol[64] and its purity, concentration, and integrity were determined. Double-terminal sequencing was performed using a high-throughput sequencing platform DNBSEQ (http://www.bgite

Gene	Model comparison	df	ΔlnL	2ΔlnL	LRT(P-value)	No. of positively selected sites (BEB: *: P > 95%; **: P > 99%)
ccsA	M0 versus M3	4	6.16163	12.32326	1.51E−02	
	M1 versus M2	2	4.516695	9.03339	1.09E−02	1
	M7 versus M8	2	4.557689	9.115378	1.05E−02	9
ndhB	M0 versus M3	4	5.524202	11.048404	2.60E−02	
	M1 versus M2	2	4.02318	8.04636	1.79E−02	1
	M7 versus M8	2	4.027433	8.054866	1.78E−02	1
ndhD	M0 versus M3	4	17.294943	34.589886	5.64E−07	
	M1 versus M2	2	12.610907	25.221814	3.34E−06	2
	M7 versus M8	2	12.621978	25.243956	3.30E−06	8
ndhF	M0 versus M3	4	41.467949	82.935898	4.16E−17	
	M1 versus M2	2	17.266444	34.532888	3.17E−08	15
	M7 versus M8	2	17.265698	34.531396	3.17E−08	17
rbcL	M0 versus M3	4	7.689562	15.379124	3.98E−03	
	M1 versus M2	2	3.062625	6.12525	4.68E−02	2
	M7 versus M8	2	3.254789	6.509578	3.86E−02	2
rpl22	M0 versus M3	4	10.642221	21.284442	2.78E−04	
	M1 versus M2	2	5.863119	11.726238	2.84E−03	1
	M7 versus M8	2	6.034404	12.068808	2.39E−03	1
rpl32	M0 versus M3	4	6.187839	12.375678	1.48E−02	
	M1 versus M2	2	5.065357	10.130714	6.31E−03	3
	M7 versus M8	2	5.107086	10.214172	6.05E−03	5
rpoC2	M0 versus M3	4	9.620859	19.241718	7.04E−04	
	M1 versus M2	2	7.288661	14.577322	6.83E−04	2
	M7 versus M8	2	7.310929	14.621858	6.68E−04	2
rps12	M0 versus M3	4	215.362127	430.724254	6.38E−92	
	M1 versus M2	2	139.496594	278.993188	2.61E−61	5
	M7 versus M8	2	172.218683	344.437366	1.61E−75	5
ycf1	M0 versus M3	4	84.010444	168.020888	2.78E−35	
	M1 versus M2	2	50.859144	101.718288	8.17E−23	15
	M7 versus M8	2	50.876717	101.753434	8.03E−23	39
ycf2	M0 versus M3	4	17.77844	35.55688	3.57E−07	
	M1 versus M2	2	3.828956	7.657912	2.17E−02	13
	M7 versus M8	2	11.908646	23.817292	6.73E−06	13

Table 5. Likelihood ratio test (LRT) and positive selection sites under different site models of PCGs in the *Quercus* section *Ilex*.

chsolutions.com/), with a read length of 150 bp. The raw data were filtered using SOAPnuke v1.3.0 to obtain 20 GB of clean data[65].

Chloroplast genome assembly and annotation

The cp genomes were assembled de novo into rings using GetOrganelle version 1.7.6.1 in a Linux system[66]. The assembled cp genomes were annotated using the GeSeq online site (https://chlorobox.mpimp-golm.mpg.de/geseq.html)[67]. Finally, cp genome graphs were drawn using the online program Organellar GenomeDRAW (OGDRAW) version 1.3.1 (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html)[68]. We used Geneious R9.0.2 software and cloud platform tools (http://cloud.genepioneer.com:9929) to perform statistics on the basic characteristics of the cp genomes. The newly cp genome sequences were uploaded and saved to the NCBI database under GenBank accession numbers OR966903 and OR966904.

Repeated sequences and codon usage *bias*

Simple sequence repeats (SSRs) were detected via batch analysis using MISA Perl scripts (https://webblast.ipk-gatersleben.de/misa/)[69]. The minimum number of repeats of mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides were set to ten, five, four, three, three, and three, respectively. We then used Tandem Repeats Finder (TRF) v4.09 software (https://tandem.bu.edu/trf/trf.html) to detect tandem repeats[70]. The alignment parameters match, mismatch, and indels were set to two, seven, and seven. respectively, and the minimum alignment score to report repeats was set to 80. In addition, we also used the website REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer) to predict scattered repeats within each cp genome[71]. The scattered repeats included forward/direct repeats (F), reverse repeats (R), complement repeats

(C), and palindromic repeats (P). Three and 30 were set as the hamming distance and minimum repeat size, respectively.

CodonW version 1.4.2, a common tool for codon bias analysis, was used to perform codon bias analysis and statistics on the coding sequences (CDSs) in these two cp genomes[72]. We screened CDSs with ≥ 300 bp for subsequent analysis. The GC content at the first, second, and third codon sites (GC1, GC2, and GC3) and the overall mean value (GC_all) of the two cp genomes were calculated. The effective number of codon (ENC), codon bias index (CBI), codon adaptation index (CAI), frequency of optical codons (FOP), and codon numbers were calculated using CodonW version 1.4.2. RSCU values were calculated and presented graphically using the R package (ggplot2). Neutrality, ENC-plot and PR2-bias-plot analyses were performed with different data to analyze the influencing factors and degree of codon usage bias.

Comparative genomic analyses of *Quercus* section *Ilex*

The currently published 17 cp genomes of *Quercus* section *Ilex* species, were downloaded from the National Center of Biotechnology Information (NCBI) database for comparative genomic analyses along with the two species in this study. The four boundary regions of each cp genome were visualized using IRscope (https://irscope.shinyapps.io/irapp/). The expansion and contraction of the IR regions were compared to identify differences in the cp genomes of *Quercus* section *Ilex*[73]. We used the Mauve plugin in Geneious R9.0.2 software to analyze structural changes in the cp genomes of *Quercus* section *Ilex*[74]. To evaluate the similarities and differences among cp genomes, the cp genome of *Q. bawanglingensis* was used as the reference sequence, and the online analysis tool mVISTA (http://genome.lbl.gov/vista/mvista/submit.shtml) was used to visualize the cp genomes of other species of *Quercus* section *Ilex*. DnaSP v6.12.03 software was used to detect polymorphic sites and nucleotide diversity (Pi values). A sliding window analysis with a step size of 200 bp and a window length of 800 bp was performed to identify regions with high variability in the cp genomes.

Comparison of chloroplast genome among different sections of *Quercus*

We downloaded all cp genomes of the four sections of *Quercus* from the NCBI database, including 19 of section *Quercus*, 37 of section *Cyclobalanopsis*, three of section *Cerris*, and 24 of section *Ilex* (available until July 2024). Combined with the two newly sequenced species of section *Ilex* in our study, the expansion and contraction of the IR regions and nucleotide diversity analysis of each section were carried out in the four sections, respectively. Then we performed statistics and significance analysis to explore the consistency and differences among the sections in the software Origin.

Phylogenetic analysis of *Quercus*

The phylogenetic relationships of *Quercus* were inferred using the maximum likelihood (ML) method based on the complete cp genomes. *Fagus engleriana* (NC_036929) was selected as the outgroup for tree rooting purposes. The whole cp genomes of 33 species were aligned using the multi-sequence alignment program MAFFT v7.450[75]. The best nucleotide substitution model was determined, and a phylogenetic tree was built with 1000 ultrafast bootstrap replicates using IQ-tree v2.1.3 software[76]. The resulting phylogenetic tree was constructed and saved using the FigTree v.1.4.4 software (http://tree.bio.ed.ac.uk/software/figtree/).

Selection pressure analysis

The Codeml program in the PAML v4.9j software was used for positive selection analysis, aiming to explore the adaptive evolution of protein-coding genes (PCGs) in the cp genomes of *Quercus* section *Ilex*[77]. The non-repeating PCGs were aligned using MEGA-X v10.2.6 software, followed by the removal of their stop codons. Subsequently, the aligned sequences were converted into a recognizable PAML format using DAMBE v7.3.32 software[78]. A phylogenetic tree was then constructed based on PCGs using IQ-tree v2.1.3 software. The selection pressure of 77 common PCGs was identified using six site models (seqtype = 1, model = 0, NSsites = 0, 1, 2, 3, 7, and 8). The posterior probability of each site was calculated using Bayes empirical Bayes (BEB) to identify positively selected sites ($P > 0.95$). For genes exhibiting positive selection sites, a likelihood ratio test (LRT) was carried out through pairwise comparison of M0 (single ratio) vs. M3 (discrete), M1 (near neutral) vs. M2 (positive selection), and M7 (beta) vs. M8 (beta & ω). The chi-square test (chi^2) was used to detect statistical significance, with genes exhibiting $P < 0.05$ considered as positively selected genes[79].

Data availability

The data that support the finding of this study are openly available in the GenBank of NCBI at https://www.ncbi.nlm.nih.gov (accessed on 20 December 2024), reference number (OR966903 and OR966904).

Received: 20 March 2024; Accepted: 30 August 2024
Published online: 04 September 2024

References

1. Denk, T. *et al.* An updated infrageneric classification of the oaks: Review of previous taxonomic schemes and synthesis of evolutionary patterns. In Oaks Physiological Ecology exploring the functional diversity of genus *Quercus* L. *Tree Physiol.* **66**, 13–38 (2017).
2. Carrero, C. *et al.* The Red List of Oaks 2020. *The Morton Arboretum: Lisle.* (2020).
3. Cavender-Bares, J. Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. *New Phytol.* **221**, 669–692 (2019).
4. Sancho-Knapik, D. *et al.* Deciduous and evergreen oaks show contrasting adaptive response in leaf mass per area across environments. *New Phytol.* **230**, 521–534 (2020).

5. Sun, X. Q., Song, Y. G., Ge, B. J., Dai, X. L. & Kozlowski, G. Intermediate epicotyl physiological dormancy in the recalcitrant seed of *Quercus chungii* F.P.Metcalf with the elongated cotyledonary petiole. *Forests.* **12**, 263 (2021).
6. Jin, D. M., Yuan, Q., Dai, X. L., Kozlowski, G. & Song, Y. G. Enhanced precipitation has driven the evolution of subtropical evergreen broad-leaved forests in eastern China since the early Miocene: Evidence from ring-cupped oaks. *J. Syst. Evol.* https://doi.org/10.1111/jse.13022 (2023).
7. Fazan, L., Song, Y. G. & Kozlowski, G. The woody plant: From past triumph to manmade decline. *Plants* **9**, 1593 (2020).
8. Hipp, A. L. *et al.* Genomic landscape of the global oak phylogeny. *New Phytol.* **226**, 1198–1212 (2020).
9. Allendorf, F. W., Hohenloge, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709 (2010).
10. Kardos, M. *et al.* The crucial role of genome-wide genetic variation in conservation. *PNAS* **118**, 48 (2021).
11. Jing, Z. Y., Cheng, K. G., Shu, H., Ma, Y. P. & Liu, P. L. Whole genome resequencing approach for conservation biology of endangered plants. *Biodiv. Sci.* **31**, 22679 (2023).
12. Wang, J. *et al.* Plant organellar genomes: Much done, much more to do. *Trends Plant Sci.* https://doi.org/10.1016/j.tplants.2023.12.014 (2023).
13. Solties, D. E. Genome-scale data, angiosperm relationships, and "ending incongruence", a cautionary tale in phylogenetics. *Trends Plant Sci.* **9**, 477–483 (2004).
14. Xia, M. *et al.* Out of the Himalaya-Hengduan mountains: phylogenomics, biogeography and diversification of *Polygonatum* Mill. (Asparagaceae) in the northern Hemisphere. *Mol. Phylogenet.* **169**, 107431 (2022).
15. Wang, N. J. *et al.* The complete chloroplast genomes of three Hamamelidaceae species: comparative and phylogenetic analyses. *Ecol. Evol.* **12**, 2 (2022).
16. Jiang, X. L. *et al.* East Asian origins of European holly oaks (*Quercus* Section *Ilex* Loudon) via the Tibet-Himalaya. *J. Biogeogr.* **46**, 2203–2214 (2019).
17. Yan, M. *et al.* Ancient events and climate adaptive capacity shaped distinct chloroplast genetic structure in the oak lineages. *BMC Evol. Biol.* **19**, 202 (2019).
18. Li, X., Li, Y., Zang, M., Li, M. & Fang, Y. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *Int. J. Mol. Sci.* **19**, 2443 (2018).
19. Wang, T., Wang, Z., Song, Y. & Kozlowski, G. The complete chloroplast genome sequence of *Quercus ningangensis* and its phylogenetic implication. *Plant Fungal Syst.* **66**, 155–165 (2021).
20. Chen, S. *et al.* The complete chloroplast genome sequence of *Quercus sessilifolia* Blume (Fagaceae). *Mitochondrial DNA* **7**, 182–184 (2022).
21. Yang, Y. *et al.* Comparative analysis of the complete chloroplast genomes of five *Quercus* species. *Front. Plant Sci.* **7**, 959 (2016).
22. Hu, H. L. *et al.* The complete chloroplast genome of the daimyo oak, *Quercus dentata Thunb*. *Conserv. Genet. Resour.* **66**, 1–3 (2018).
23. Kim, K. & Lee, H. 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).
24. Lu, R. S., Li, P. & Qiu, Y. X. The complete chloroplast genomes of three *Cardiocrinum* (Liliaceae) species: Comparative genomic and phylogenetic analyses. *Front. Plant Sci.* **7**, 2054 (2017).
25. Zhang, S. D. *et al.* Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol.* **214**, 1355–1367 (2017).
26. Timme, R. E. *et al.* A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. *Am. J. Bot.* **94**, 302–312 (2007).
27. Weng, M. L. *et al.* Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol. Biol. Evol.* **31**, 645–659 (2014).
28. Roullier, C., Rossel, G., Tay, D., McKey, D. & Lebot, V. Combining chloroplast and nuclear microsatellites to investigate origin and dispersal of New World sweet potato landraces. *Mol. Ecol.* **20**, 3963–3977 (2011).
29. Yang, A. H., Zhang, J. J., Yao, X. H. & Huang, H. W. Chloroplast microsatellite markers in *Liriodendron tulipifera* (Magnoliaceae) and cross-species amplifcation in *L. chinense*. *Am. J. Bot.* **98**, 123–126 (2011).
30. Jiao, Y. *et al.* Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). *BMC Genom.* **13**, 201 (2012).
31. Yang, J. *et al.* Development of chloroplast and nuclear DNA markers for Chinese oaks (*Quercus* subgenus *Quercus*) and assessment of their utility as DNA barcodes. *Front. Plant Sci.* **8**, 816 (2017).
32. Yang, Y. *et al.* Plastid genome comparative and phylogenetic analyses of the key genera in Fagaceae: Highlighting the effect of codon composition bias in phylogenetic inference. *Front. Plant Sci.* **9**, 82 (2018).
33. Chi, X., Zhang, F., Dong, Q. & Chen, S. Insights into comparative genomics, codon usage bias, and phylogenetic relationship of species from Biebersteiniaceae and Nitrariaceae based on complete chloroplast genomes. *Plants* **9**, 1605 (2020).
34. Ren, T. *et al.* Plastomes of eight *Ligusticum* Species: Characterization, genome evolution, and phylogenetic relationships. *BMC Plant Biol.* **20**, 519 (2020).
35. Delannoy, E., Fujii, S., Colas Des Francs-Small, C., Brundrett, M. & Small, I. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Mol. Biol. Evol.* **28**, 2077–2086 (2011).
36. Xu, C. *et al.* Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium gower* Ramsey. *Evol. Bioinform.* **7**, 271–278 (2011).
37. Wang, R. J. *et al.* Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.* **8**, 36 (2008).
38. Raubeson, L. A. *et al.* Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genom.* **8**, 174 (2007).
39. Kode, V., Mudd, E. A., Iamtham, S. & Day, A. The tobacco plastid accD gene is essential and is required for leaf development. *Plant J.* **44**, 237–244 (2005).
40. Cai, Z. *et al.* Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: Implications for the phylogenetic relationships of Magnoliids. *BMC Evol. Biol.* **6**, 77 (2006).
41. Liu, X. *et al.* Complete chloroplast genome sequence and phylogenetic analysis of *Quercus bawanglingensis* Huang, Li Et Xing, a vulnerable oak tree in China. *Forests* **10**, 587 (2019).
42. Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* **7**, e35071 (2012).
43. Dong, W. *et al. Ycf1*, the most promising plastid Dna barcode of land plants. *Sci. Rep.* **5**, 8348 (2015).
44. Zecca, G. *et al.* The timing and the mode of evolution of wild grapes (*Vitis*). *Mol. Phylogenet. Evol.* **62**, 736–747 (2012).
45. Yang, Y., Hu, Y., Ren, T., Sun, J. & Zhao, G. Remarkably conserved plastid genomes of *Quercus* group *Cerris* in China: Comparative and phylogenetic analyses. *Nord. J. Bot.* **36**, e01921 (2018).
46. Yan, M., Xiong, Y., Liu, R., Deng, M. & Song, J. The application and limitation of universal chloroplast markers in discriminating East Asian Evergreen Oaks. *Front. Plant. Sci.* **9**, 569 (2018).
47. Yang, Y., Zhou, T., Qian, Z. & Zhao, G. Phylogenetic relationships in Chinese oaks (Fagaceae, *Quercus*): Evidence from plastid genome using low-coverage whole genome sequencing. *Genomics* **113**, 1438–1447 (2021).

48. Manos, P. S., Doyle, J. J. & Nixon, K. C. Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Mol. Phylogenet. Evol.* **12**, 333–349 (1999).
49. Curtu, A. L., Gailing, O. & Finkeldey, R. Evidence for hybridization and introgression within a species-rich oak (*Quercus* Spp.) Community. *BMC Evol. Biol.* **7**, 218 (2007).
50. Li, H. *et al.* Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* **19**, 232 (2021).
51. Zhou, B. F. *et al.* Phylogenomic analyses highlight innovation and introgression in the continental radiations of Fagaceae across the Northern Hemisphere. *Nat. Commun.* **13**, 1320 (2022).
52. Petit, R. J. *et al.* Hybridization as a mechanism of invasion in oaks. *New Phytol.* **161**, 151–164 (2004).
53. Acosta, M. C. & Premoli, A. C. Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Mol. Phylogenet. Evol.* **54**, 235–242 (2010).
54. Fehrer, J., Gemeinholzer, B., Chrtek, J. & Bräutigam, S. Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in Pilosella hawkweeds (Hieracium, Cichorieae, Asteraceae). *Mol. Phylogenet. Evol.* **42**, 347–361 (2007).
55. Costello, L. R., Hagen, B. W. & Katherine, S. J. *Oaks in the Urban Landscape, Selection, Care and Preservation* (California University of California, 2011).
56. Hubert, F. O. *et al.* Multiple nuclear genes stabilize the phylogenetic backbone of the genus *Quercus*. *Syst. Biodiv.* **12**, 405–423 (2014).
57. Simeone, M. C. *et al.* Plastome data reveal multiple geographic origins of *Quercus* Group *Ilex*. *PeerJ.* **4**, e1897 (2016).
58. Zhao, D. N., Ren, Y. & Zhang, J. Q. Conservation and innovation: Plastome evolution during rapid radiation of *Rhodiola* on the Qinghai-Tibetan Plateau. *Mol. Phylogenet. Evol.* **144**, 106713 (2020).
59. Azarin, K. *et al.* Comparative analysis of chloroplast genomes of seven perennial *Helianthus* species. *Gene* **774**, 145418 (2021).
60. Kapralov, M. V. & Filatov, D. A. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol. Biol.* **7**, 73 (2007).
61. Xie, Z. & Merchant, S. The plastid-encoded *ccsA* gene is required for heme attachment to chloroplast c-type cytochromes. *Biol. Chem.* **271**, 4632–4639 (1996).
62. Drescher, A. *et al.* The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* **22**, 97–104 (2000).
63. Liu, X. *et al.* Comparative analysis of the complete chloroplast genomes of six white oaks with high ecological amplitude in China. *J. For. Res.* **32**, 2203–2218 (2021).
64. Doyle, J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
65. Chen, Y. *et al.* SOAPnuke: A MapReduce acceleration supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120 (2018).
66. Jin, J. *et al.* Getorganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
67. Michael, T. *et al.* GeSeq–versatile and accurate annotation of organelle genomes. *Nucl. Acids Res.* **45**, 6–11 (2017).
68. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 131: Expanded toolkit for the graphical visualization of organellar genomes. *Nucl. Acids Res.* **47**, 59–64 (2019).
69. Beier, S., Thiel, T., Munch, T., Scholz, U. & Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
70. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucl. Acids Res.* **27**, 573–580 (1999).
71. Kurtz, S. *et al.* REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucl. Acids Res.* **29**, 4633–4642 (2001).
72. Sharp, P. M. & Li, W. H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.* **15**, 1281–1295 (1987).
73. Amiryousefi, A., Hyvonen, J. & Poczai, P. IRscope: An online program to visualize the junction sites of cp genomes. *Bioinformatics* **34**, 3030–3031 (2018).
74. Darling, A. C. E. *et al.* Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
77. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
78. Xia, X. H. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **35**, 1550–1552 (2018).
79. Yang, Z. H. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).

Acknowledgements

This work was supported by grants from: the Special Fund for Scientific Research of Shanghai Landscaping & City Appearance Administrative Bureau (G242414, G242416).

Author contributions

Y.R.Z. and Y.L. conducted data analysis and draft manuscript. L.H.Y. conducted plant material preparation. Y.G.S. conducted plant identification. M.H.L. and S.S.Z. designed the article and experimental guidance. G.K., L.T.Y., M.H.L., S.S.Z., and Y.G.S. revised the manuscript. All authors reviewed the manuscript, and approved the final manuscript.

Funding

This work was supported by grants from: the Special Fund for Scientific Research of Shanghai Landscaping & City Appearance Adminis-trative Bureau (G242414, G242416).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-71838-w.

Correspondence and requests for materials should be addressed to M.-H.L. or S.-S.Z.