Plant and Fungal Systematics 66(2): 155–165, 2021
DOI: https://doi.org/10.35535/pfsyst-2021-0014

ISSN 2544-7459 (print)
ISSN 2657-5000 (online)

Dedicated to the Professor Adam Boratyński

The complete chloroplast genome sequence of *Quercus ningangensis* and its phylogenetic implications

Tian-Rui Wang[1], Zheng-Wei Wang[2], Yi-Gang Song[1*] & Gregor Kozlowski[1,3,4*]

Article info

Received: 2 Mar. 2021
Revision received: 28 Apr. 2021
Accepted: 29 Jun. 2021
Published: 31 Dec. 2021

Associate Editor

Monika Dering

Abstract. *Quercus ningangensis* is an economically and ecologically important tree species belonging to the family *Fagaceae*. In this study, the complete chloroplast (cp) genome of *Q. ningangensis* was sequenced and assembled, and 18 published cp genomes of *Quercus* were retrieved for genomic analyses (including sequence divergence, repeat elements, and structure) and phylogenetic inference. With this study, we found that complete cp genomes in *Quercus* are conserved, and we discovered a codon composition bias, which may be related to genomic content and genetic characteristics. In addition, we detected considerable structural variations in the expansion and contraction of inverted repeat regions. Six regions with relatively high variable (*matK-rps16*, *psbC*, *ycf3* intron, *rbcL*, *petA-psbJ*, and *ycf1*) were detected by conducting a sliding window analysis, which has a high potential for developing effective genetic markers. Phylogenetic analysis based on Bayesian inference and maximum likelihood methods resulted in a robust phylogenetic tree of *Quercus* with high resolution for nearly all identified nodes. The phylogenetic relationships showed that the phylogenetic position of *Q. ningangensis* was located between *Q. sichourensis* and *Q. acuta*. The results of this study contribute to future research into the phylogenetic evolution of *Quercus* section *Cyclobalanopsis* (*Fagaceae*).

Key words: chloroplast genome, *Cyclobalanopsis ningangensis*, phylogeny, *Quercus*, repeat regions, structural variation

Introduction

Woody species have received much scientific and media attention over the past decade due to their very significant economic value, as well as their importance with respect to ecology and biodiversity (Häggström 2019; Carrero et al. 2020; Fazan et al. 2020). The genus *Quercus* (oaks, *Fagaceae*) is among the most widespread and species-rich tree genera in the Northern Hemisphere and has diversified in rainforests, montane cloud forests, and temperate broad-leaved forests (Nixon 1997; Huang et al. 1999; Song et al. 2019; Fazan et al. 2020; Kremer & Hipp 2020). Oaks arose an estimated 56 million years ago (Ma) and subsequently radiated and expanded across the Northern Hemisphere (Manos & Stanford 2001; Hipp et al. 2020). Today, there are estimated to be at least 435 species, which are assigned to eight different sections (Denk et al. 2017). The genus extends from the equator to the boreal regions at a latitude of 60°N in Europe, and from sea level to 4,500 m above sea level in the Tibetan Plateau (Huang et al. 1999; Kremer & Hipp 2020).

Quercus ningangensis belongs to the *Quercus* section *Cyclobalanopsis*, which diverged at the transition of the Oligocene and Miocene (~26 Ma) (Deng et al. 2018). The distribution of this species is very narrow in Guangxi, Hunan, and Jiangxi provinces, and its conservation status is uncertain due to data deficiency (Huang et al. 1999; Carrero et al. 2020). Morphologically, the single-celled trichome base supported *Q. ningangensis* as a member of the single-celled trichome bases (STB) lineage in the section *Cyclobalanopsis* (Deng et al. 2014). It has four trichome types, including fasciculate, simple stellate, appressed lateral attached, and uniseriate, which is one of the most diversified in this section (Deng et al. 2014). However, recent phylogenetic and phylogenomic studies have not included samples of this species (Deng et al. 2018; Hipp et al. 2020; Yang et al. 2020).

Chloroplast (cp) genomes have been widely used to infer plant phylogenetic relationships at all taxonomic levels in angiosperms because of their conserved structure,

[1] Eastern China Conservation Centre for Wild Endangered Plant Resources, Shanghai Chenshan Botanical Garden, Shanghai, 201602, China (Song, ORCID: 0000-0003-2584-2338; Kozlowski, ORCID: 0000-0003-4856-2005)
[2] Horticulture and Landscape Department, Shanghai Chenshan Botanical Garden, Shanghai, 201602, China
[3] Department of Biology and Botanic Garden, University of Fribourg, Fribourg, Switzerland
[4] Natural History Museum Fribourg, Fribourg, Switzerland
* Corresponding authors e-mail: cherish-faith@163.com (Y.G.S.), gregor.kozlowski@unifr.ch (G.K.)

uniparental inheritance, general lack of recombination, and small effective population size (Cosner et al. 2004; Wicke et al. 2011). With the rapid development of next-generation sequencing, cp genomes are increasingly being used for phylogenetic relationship reconstruction (Li et al. 2020a; Mu et al. 2020; Zhao et al. 2020; Uckele et al. 2021). Since the first oak cp genome for *Q. rubra* was published in 2014, up to 30 species have been sequenced (Alexander & Woeste 2014; Li et al. 2020b; Zhang et al. 2020). Although angiosperm cp genomes exhibit a remarkably conserved gene content and order (Jansen & Ruhlman 2012), complete cp genome data will provide more valuable information for understanding the phylogenetic relationships and intraspecific diversity of *Quercus* (Li et al. 2018).

There are seven complete cp genomes in the *Quercus* section *Cyclobalanopsis* that were also included in the phylogenetic reconstruction based on the data of restriction-site associated DNA sequencing. *Q. acuta*, *Q. ciliaris*, *Q. arbutifolia*, *Q. myrsinifolia*, and *Q. stewardiana* belong to the STB lineage, and *Q. gilva* and *Q. sichourensis* belong to the compound trichome base (CTB) lineage (Deng et al. 2018). In addition to the other two complete cp genomes of *Q. edithiae* and *Q. glauca*, we sought to provide a relatively stable systematic position for *Q. ningangensis*.

We first sequenced and described the complete cp genome of *Q. ningangensis* and performed a comparative analysis of the cp genomes of multiple *Quercus* species to (1) investigate the structural patterns of the whole cp genome of *Quercus* section *Cyclobalanopsis* species including the genome structure, gene order and gene content; (2) examine abundant simple sequence repeats (SSRs) and repeat structure in the whole cp genome of *Q. ningangensis* to provide markers for phylogenetic and genetic studies; and (3) construct a cp phylogeny for *Quercus*, especially section *Cyclobalanopsis* species to illuminate the phylogenetic position of *Q. ningangensis*.

Material and methods

Plant material and DNA extraction

In total, 19 complete cp genomes belonging to the key genera of *Quercus* were analyzed in this study, including one newly generated complete cp genome (*Q. ningangensis*) and 18 published complete cp genomes in *Quercus*. The collection and GenBank accession information for the analyzed taxa are listed in Table 1. A single *Q. ningangensis* individual was sampled from Da-Long town (114.084°E, 26.603°N; 750 m), Jing-Gang-Shan City, Jiangxi Province, China, in September 2019. The voucher specimen (accession no. SYG00053) was deposited at the Shanghai Chenshan Botanical Garden. Total genomic DNA was extracted from fresh leaves using the modified CTAB method (Doyle 1987).

Illumina sequencing, assembly, and annotation

Total genomic DNA was sequenced using an Illumina NovaSeq6000 platform (Illumina, San Diego, CA, USA) with PE150 based on the whole-genome shotgun strategy. Raw reads were cleaned using SOAPnuke Toolkit v.1.3.0 (Chen et al. 2018) with the default parameters set to remove low-quality reads (Patel & Jain 2012), and approximately 4 Gb of clean data were generated. Reference-guided assembly was then used to construct the plastid genomes using SPAdes v.3.13.0 (Bankevich et al. 2012). Principal contigs representing the cp genome were obtained after a BLAST search (NCBI BLAST v.2.2.30). In this process, the complete cp genome sequence of *Q. glauca* (GenBank accession number: NC_036930) was used as the reference genome. Gapcloser (v.1.12) was used to

Table 1. Information of the *Quercus* chloroplast genome used in this study.

Species	Family/Genus/Section	Genbank No.	Size (bp)
Q. myrsinifolia	Fagaceae, *Quercus*, *Cyclobalanopsis*	MN199025	160, 803
Q. ciliaris	Fagaceae, *Quercus*, *Cyclobalanopsis*	MN199024	160, 842
Q. stewardiana	Fagaceae, *Quercus*, *Cyclobalanopsis*	MN199023	160, 842
Q. ningangensis	Fagaceae, *Quercus*, *Cyclobalanopsis*	MW628880	160, 736
Q. acuta	Fagaceae, *Quercus*, *Cyclobalanopsis*	MT742291	160, 533
Q. glauca	Fagaceae, *Quercus*, *Cyclobalanopsis*	NC_036930	160, 798
Q. arbutifolia	Fagaceae, *Quercus*, *Cyclobalanopsis*	NC_039972	160, 817
Q. edithiae	Fagaceae, *Quercus*, *Cyclobalanopsis*	KU382355	160, 988
Q. sichourensis	Fagaceae, *Quercus*, *Cyclobalanopsis*	NC_036941	160, 681
Q. gilva	Fagaceae, *Quercus*, *Cyclobalanopsis*	NC_049876	160, 742
Q. acutissima	Fagaceae, *Quercus*, *Cerris*	NC_039429	161, 127
Q. variabilis	Fagaceae, *Quercus*, *Cerris*	NC_031356	161, 077
Q. phillyraeoides	Fagaceae, *Quercus*, *Ilex*	MK105462	161, 132
Q. tarokoensis	Fagaceae, *Quercus*, *Ilex*	NC_036370	161, 355
Q. aliena	Fagaceae, *Quercus*, *Quercus*	NC_026790	160, 921
Q. fabri	Fagaceae, *Quercus*, *Quercus*	MK922346	161, 227
Q. mongolica	Fagaceae, *Quercus*, *Quercus*	NC_043858	161, 194
Q. robur	Fagaceae, *Quercus*, *Quercus*	NC_046388	161, 172
Q. rubra	Fagaceae, *Quercus*, *Quercus*	NC_020152	161, 304
Trigonobalanus doichangensis	Fagaceae	NC_023959	159, 938
Juglans mandshurica	Juglandaceae	NC_033892	159, 729

fill in the gaps (Luo et al. 2012), and the final average assembly coverage was 394×. The complete cp genome was annotated using the program CPGAVAS2 (Liu et al. 2012; http://47.96.249.172:16019/analyzer/annotate). Circular plastid genome maps were drawn using the online program Organellar GenomeDRAW (Lohse et al. 2013; http://ogdraw.mpimp-golm.mpg.de/). The final annotated cp genome of *Q. ningangensis* was deposited in GenBank with the accession number MW628880.

Intraspecific sequence divergence hotspot identification

The alignment of ten complete cp genome sequences of section *Cyclobalanopsis* was conducted using the Mauve Tool of HomBlocks software (Bi et al. 2018). To determine the nucleotide diversity of the ten complete cp genomes of section *Cyclobalanopsis*, sliding window analysis was conducted to generate nucleotide diversity (Pi) of the cp genomes using DnaSP (DNA Sequence Polymorphism v.5.10.01) software (Librado & Rozas 2009). The step size was set to 200 bp with an 800 bp window length (Zhang et al. 2020).

Simple sequence repeats elements and structure analysis

Simple sequence repeats (SSR) within the 21 complete cp genomes were detected using MISA (MIcroSAtellite; http://pgrc.ipk-gatersleben.de/misa) with a motif size of one to six nucleotides (Thiel et al. 2003). Thresholds for a minimum number of repeat units were established as follows: >10 for mono-nucleotide, >5 for di-nucleotide, >4 for tri-nucleotide, and >3 for tetra-nucleotide, penta-nucleotide, or hexa-nucleotide SSRs, respectively. All identified SSRs were manually verified to compare the variation type. To discover and visualize the interspecific variation among the ten complete cp genome sequences of section *Cyclobalanopsis*, we constructed multiple alignments of ten cp genome sequences using the mVISTA comparative genomics tool (Frazer et al. 2004) with the annotation of *Q. ningangensis* as a reference. The borders and gene rearrangement of IR regions among ten species (section *Cyclobalanopsis*) were ascertained using the IRscope Online tool (Amiryousefi et al. 2018; https://irscope.shinyapps.io/irapp/) to analyze the expansions and contractions, as well as the variation in junction regions.

Phylogenetic analysis

To reconstruct the phylogenetic relationship and verify the phylogenetic position of *Q. ningangensis* in the *Quercus* section *Cyclobalanopsis*, phylogenetic analysis was conducted based on 21 taxa, including one species in the current study, 18 other *Quercus* species, and two species (*Trigonobalanus doichangensis* and *Juglans mandshurica*) that were used as outgroups. Homologous sequences of 21 complete cp genomes, including conserved coding region genes and non-coding region sequences, were screened using Homblocks software (Bi et al. 2018). The phylogenetic tree was constructed using the aligned homologous sequences by the Mauve alignment tool in the Homblocks software (Bi et al. 2018). Modeltest v.3.7 (Brigham Young University, Provo, UT, USA; Posada & Crandall 1998) was used to determine the best-fitting model for homologous sequences based on the Akaike information criterion (AIC).

The maximum likelihood method and Bayesian inference were implemented using the IQtree (Nguyen et al. 2014) and MrBayes v.3.1.2 (Swedish Museum of Natural History, Stockholm, Sweden; Ronquist & Huelsenbeck 2003), respectively. The ML tree was constructed with 1,000 bootstrap replicates. Bayesian inference (BI) was performed using the following settings: Markov chain Monte Carlo (MCMC) algorithm for 2 million generations with two incrementally heated chains, starting from random trees, and sampling one out of every 100 generations. Convergence was determined by examining the average standard deviation of split frequencies (<0.01). The first 25% of the trees were discarded as burn-in (Meng et al. 2008; Ma et al. 2014) and the remaining trees were used to build a majority-rule consensus tree.

Results

General feature of the complete chloroplast genome in *Q. ningangensis*

The complete assembled cp genome of *Q. ningangensis* was 160,736 bp in length with a typical quadripartite structure. Gene content and order were similar to those of other published cp genomes in *Fagaceae* (Dane et al. 2015; Yang et al. 2018; Zhang et al. 2020). The base composition of the cp genome was asymmetric (31.15% A, 18.78 C, 18.11% G, 31.96% T) with an overall guanine-cytosine (GC) content of 36.89% and the corresponding values of the inverted repeat (IR), large single copy (LSC), and small single copy (SSC) regions were 42.76%, 34.74%, and 31.09%, respectively. In particular, the cp genome consists of a pair of IR regions of 25,825 bp, a LSC region of 90,182 bp, and a SSC region of 18,904 bp (Fig. 1, Table 2). In total, this cp genome contains 127 genes including 86 protein-coding genes, 37 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes. Most of the genes occurred as single copies in the LSC and SSC, but 18 genes were duplicated in the IR regions, including seven protein-coding genes (*ndhB*, *rpl2*, *rpl23*, *rps12*, *rps7*, *ycf1*, and *ycf2*), seven tRNA genes (*trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*), and four rRNA genes (*16S*, *23S*, *4.5S*, and *5S*). Among these genes, 15 genes (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps16*, *trnA-UGC*, *trnG-GCC*, *trnK-UUU*, *trnL-UAA*, *trnI-GAU*,

Table 2. General features of *Q. ningangensis* chloroplast genomes compared in this study.

Region	Length	A	T	C	G	AT	GC
LSC	90182	31.92	33.34	17.79	16.94	65.26	34.74
SSC	18904	34.36	34.55	16.37	14.72	68.91	31.09
IR	51650	28.62	28.62	21.38	21.38	57.24	42.76
Total	160736	31.15	31.96	18.78	18.11	63.11	36.89

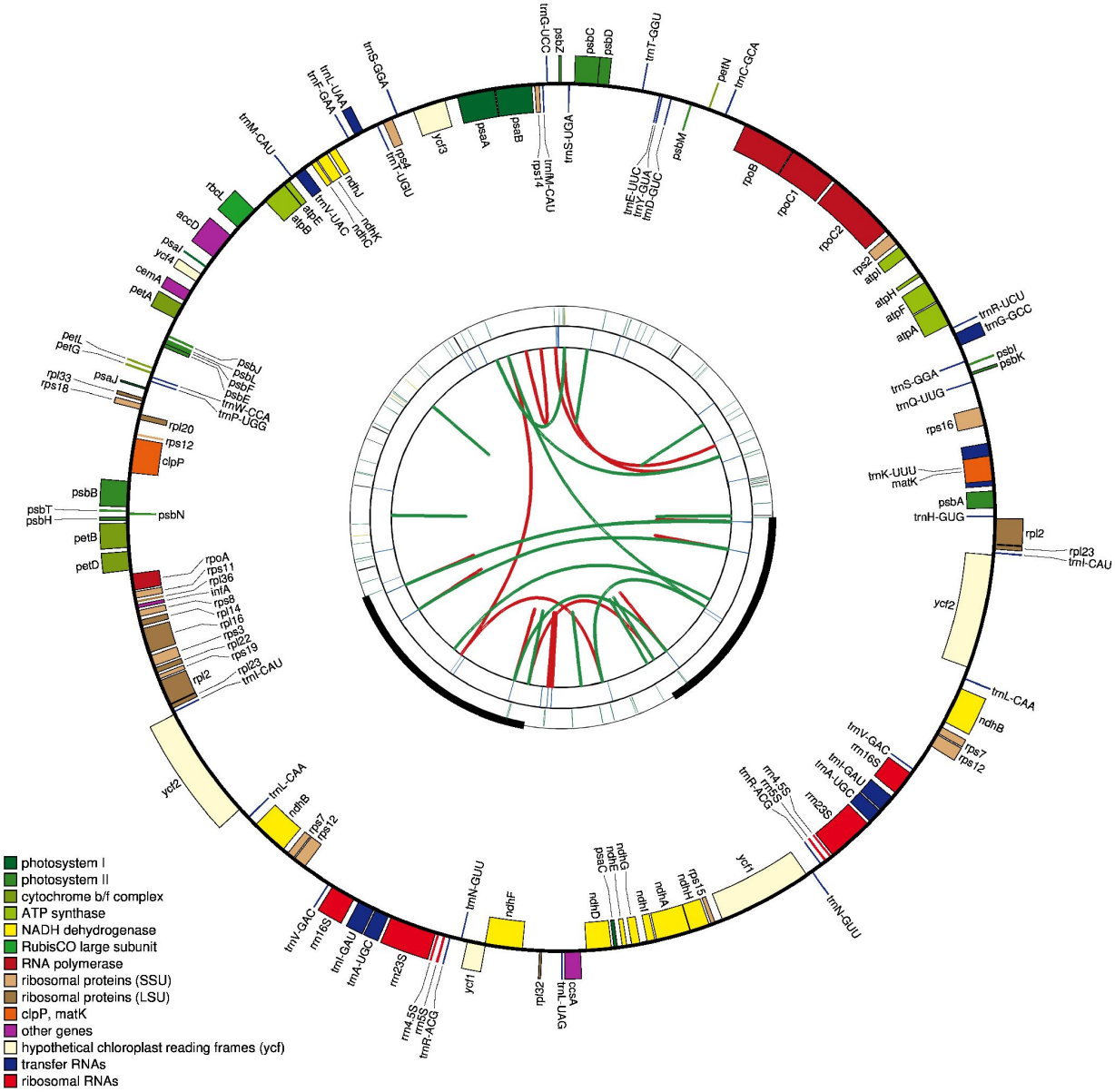

Figure 1. Gene maps of chloroplast genomes of *Quercus ningangensis*. The genes shown outside of the circle are transcribed clockwise,while those inside are transcribed counterclockwise. Genes belonging to different functional groups are color-coded. Dashed area in the inner circle indicates the guanine-cytosine (GC) content of the chloroplast genome.

and *trnV-UAC*) contained a single intron, and the other three genes (*ycf3*, *rps12*, and *clpP*) harbored two introns.

Intraspecific sequence divergence and microsatellites (SSR)

Based on the cp genome sequence alignment of the section *Cyclobalanopsis* taxa, we performed a sliding window analysis to detect the highly variable regions in the *Q. ningangensis* cp genomes (Fig. 2). The average value of nucleotide diversity (Pi) within *Q. ningangensis* was 0.00057. Six mutation hotspots were identified with higher Pi values (0.003) in the *Q. ningangensis* cp genomes including two intergenic regions (*matK-rps16* and *petA-psbJ*), three gene regions (*psbC*, *rbcL*, and *ycf1*), and one gene intron region (*ycf3*). The most variable region identified was *petA-psbJ* (IGS) with a Pi value of 0.00826. Only one hotspot, *ycf1*, was located in the SSC region.

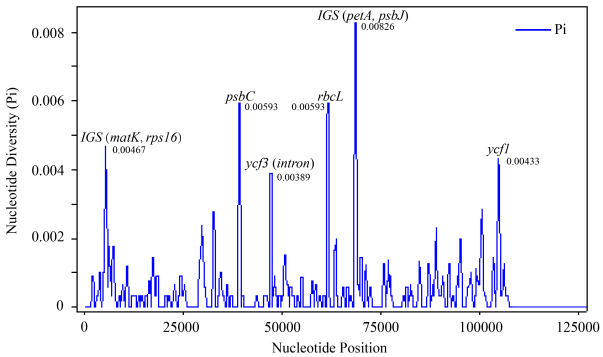

Figure 2. Sliding window analysis of ten chloroplast genomes of the section *Cyclobalanopsis*. Windowlength: 800 bp; step size: 200 bp. X-axis: position of the middle point of the window. Y-axis: nucleotide diversity per window. The genes and numbers on the curve represent the six hotspots and the corresponding Pi values with higher Pi values.

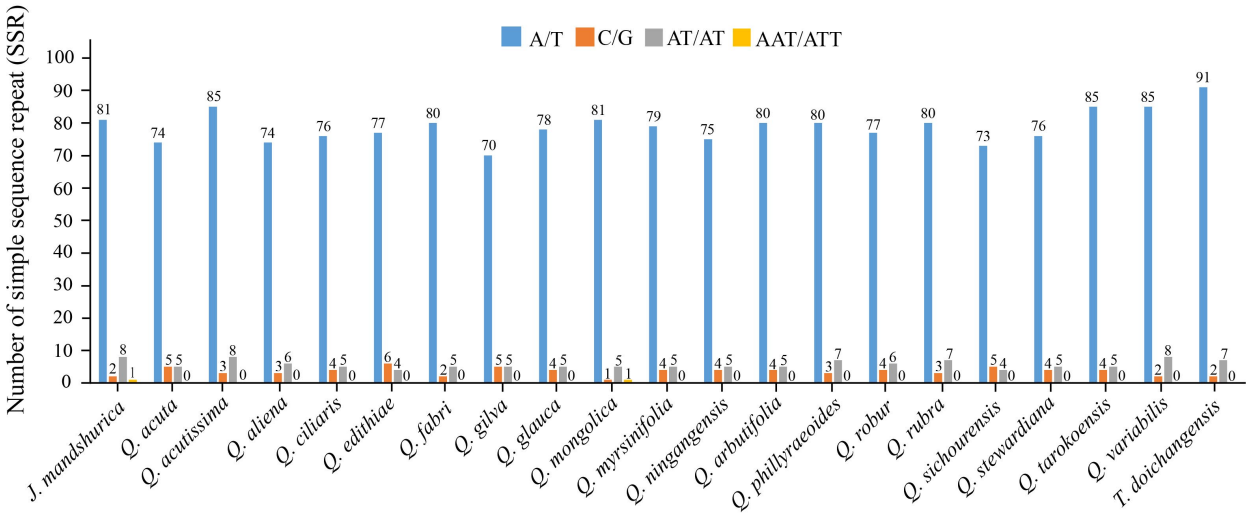

Figure 3. Simple sequence repeat (SSR) analysis in different repeat type classes based on 21 complete chloroplast genome sequences.

The IR regions exhibited lower variability than the LSC and SSC regions.

Overall, the numbers and distributions of all repeat elements were similar and conserved across these 21 species (Fig. 3). Among these elements, there were more single nucleotide repeats than double nucleotide repeats, most of the single nucleotide repeat types were composed of A/T, and trinucleotide repeats were very rare across the cp genomes.

Comparison of border regions and sequence identity

The border regions of the section *Cyclobalanopsis* cp genomes were compared with those of *Q. ningangensis* to analyze the expansion and contraction variation in

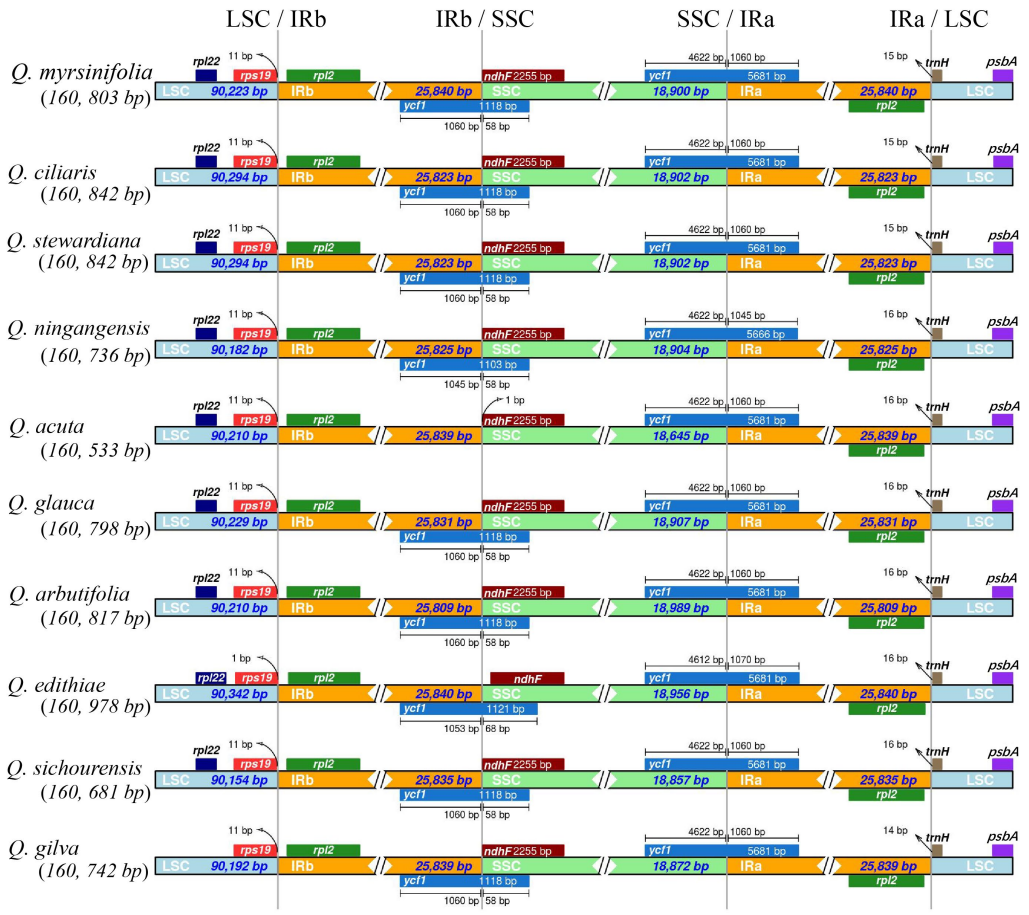

Figure 4. Comparision of the large single copy (LSC), inverted repeat (IR) and small single copy (SSC) border regions among *Quercus* section *Cyclobalanopsis* chloroplast genomes. Boxes above or below the main line represent the genes at the IR/SC borders. Numbers above the gene features represent the distance from the end of gene to the boundary region.

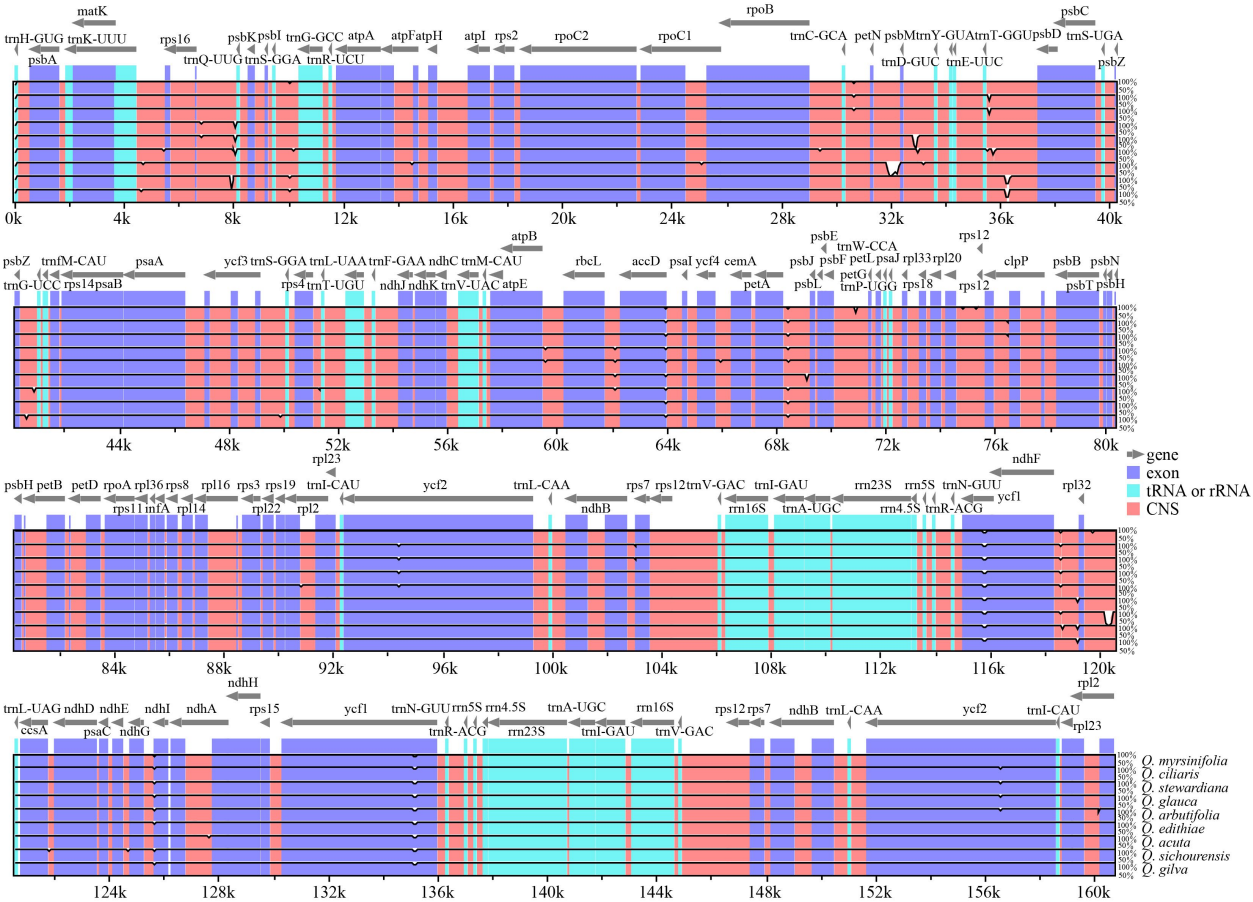

Figure 5. Sequence identity plot comparing the ten section *Cyclobalanopsis* chloroplast genomes with *Q. ningangensis* as a reference by using mVISTA. The top line shows annotated genes in order (transcriptional direction indicated by arrows). *Quercus ningangensis* is shown as horizontal bars indicating the average percent identity percentage ranging from 50 to 100% (shown on the y-axis of the graph). The x-axis corresponds to the coordinates within the chloroplast genome. Protein-coding (exon), tRNA or rRNA and conserved non-coding sequences (CNS) are marked in purple, blue and pink, respectively.

junction regions. The length of the LSC regions varied from 90,154 bp to 90,342 bp, and the IR regions of ten cp genomes ranged from 25,809 bp (*Q. arbutifolia*) to 25,840 bp (*Q. myrsinifolia*) in size, of which *rps19*, *ycf1*, *ndhF*, *rpl2*, and *trnH* genes were present at the junctions of the LSC/IR and SSC/IR borders (Fig. 4). Considerable variation was observed in the expansion and contraction of the IR regions. Although the genomic structure and size were highly conserved in the ten cp genomes, the IR/SC boundary regions still showed slight differences. For the LSC/IR borders, the gene rps19 in the LSC of section *Cyclobalanopsis* species extended 1–11 bp into the IRb region. The gene *trnH* in the LSC region contracted 14–16 bp from the junction region of IRa/LSC. In contrast, the SSC/IR boundary regions are relatively stable. The gene *ycf1* in the SSC region exhibited an interesting astride at the border of SSC/IRa. *Q. ningangensis* extended 1,045 bp into the IRa region, whereas other species of section *Cyclobalanopsis* extended 1,060 bp.

The sequence identity of the section *Cyclobalanopsis* cp genomes was performed using mVISTA with *Q. ningangensis* as a reference (Fig. 5). Overall, sequence divergence was highly conserved and similar across the ten cp genomes. Most of the significant divergence was found in conserved non-coding sequences (CNS), such as the high degree of divergence was found in *psbK-trnQ* (*UUG*), *trnD* (*GUC*)-*trnY* (*GUA*), and *psbD-trnT* (*GGU*). As expected, the IR and coding regions exhibited higher conservation than the SC and noncoding regions.

Phylogenetic analysis

Aligned homologous sequences of 21 complete cp genomes were used to construct the phylogenetic trees using maximum likelihood and Bayesian inference analyses (Fig. 6). The best-fit models used in the ML and BI analyses were General Time Reversible (GTR) with gamma distribution and invariable (GTR+G+I) (Table S1). All nodes of the phylogenetic trees were strongly supported by 0.90–1.00 Bayesian posterior probabilities in BI analysis and 57%–100% bootstrap values in ML analysis (Fig. 7). ML tree reconstructions using the complete cp genomes showed identical topologies and well-resolved nodes for the BI tree generated by MrBayes (Fig. 7). Three clades were recognized as section *Quercus*, section *Cerris*/section *Ilex*, and section *Cyclobalanopsis*. Within the section *Cyclobalanopsis*, three lineages were inferred based on the taxa we used: (1) *Q. myrsinifolia* was sister to the other species; (2) *Q. ciliaris* and *Q. stewardiana* were sister to the other seven species in this lineage; and (3) the last seven species comprised the core lineage of

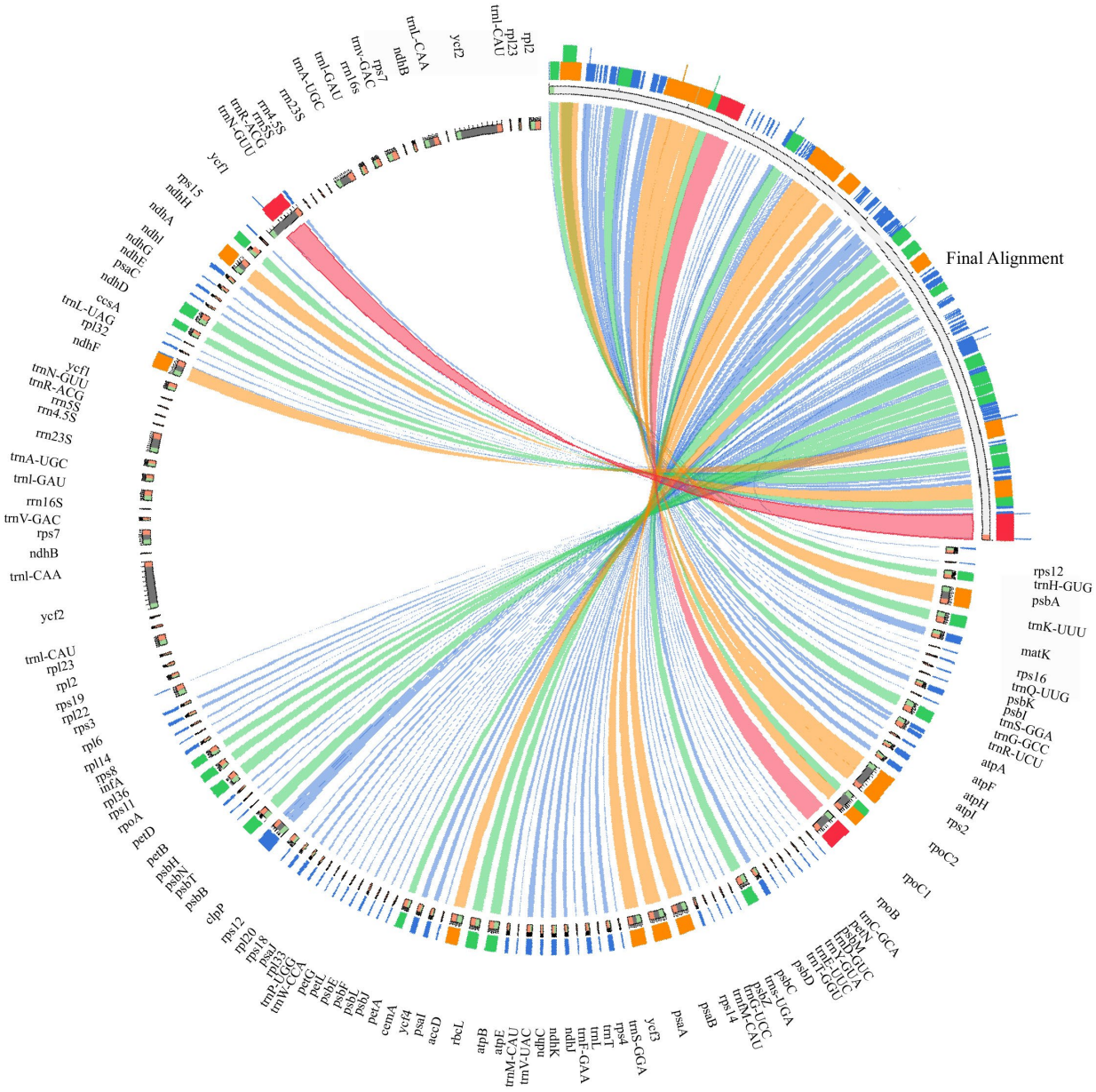

Figure 6. Homologous sequences visualization of 21 complete chloroplast genomes. The small ring on the right represents the resulting alignment sequence, and the large ring on the left represents the corresponding gene. Lines show the relative positions of genes. The color of the line indicates the similarity between the resulting alignment sequence and the original sequence (Blue ≤0.5, red ≥0.99).

section *Cyclobalanopsis* and was significantly favored by the BI tree. *Q. ningangensis* was nested in this lineage, confirming its placement (Fig. 7).

Discussion

Chloroplast sequence variation and evolution

Understanding nucleotide diversity is of fundamental importance in molecular evolution (Muse & Gaut 1994). Sliding window analysis identified that the most variable region was *petA-psbJ* (IGS) with a Pi value of 0.00826 (Fig. 2). The variation region of *petA-psbJ* has been found in previous studies to be highly variable (Timme et al. 2007; Yu et al. 2020) and the mutation hotspots identified with *matK* are recommended regions for DNA barcoding in plants (CBOL Plant Working Group 2009; Hollingsworth et al. 2011). Furthermore, most divergent hotspot loci are located in the LSC region, which allows for the proper design of genetic markers for classification and revealing the genetic divergence of the *Quercus* taxa. The IR regions exhibited lower variability than the LSC and SSC regions. This result was similar to other cp genomes (Han et al. 2016; He et al. 2017; Zhang et al. 2020).

Previous research has shown that repeated sequences may play an important role in the rearrangement of cp genome sequence divergences (Timme et al. 2007; Weng et al. 2013) and GC content is significant in shaping codon usage bias and evolution of genomic structure (Sueoka & Kawanishi 2000; Bellgard et al. 2001). Therefore, we investigated the numbers and distributions of all repeat elements across these 21 species. We found a high adenine-thymine (AT) content level in *Quercus* and there was a strong bias toward AT content among these species (Fig. 3), which is consistent with studies of previous cp

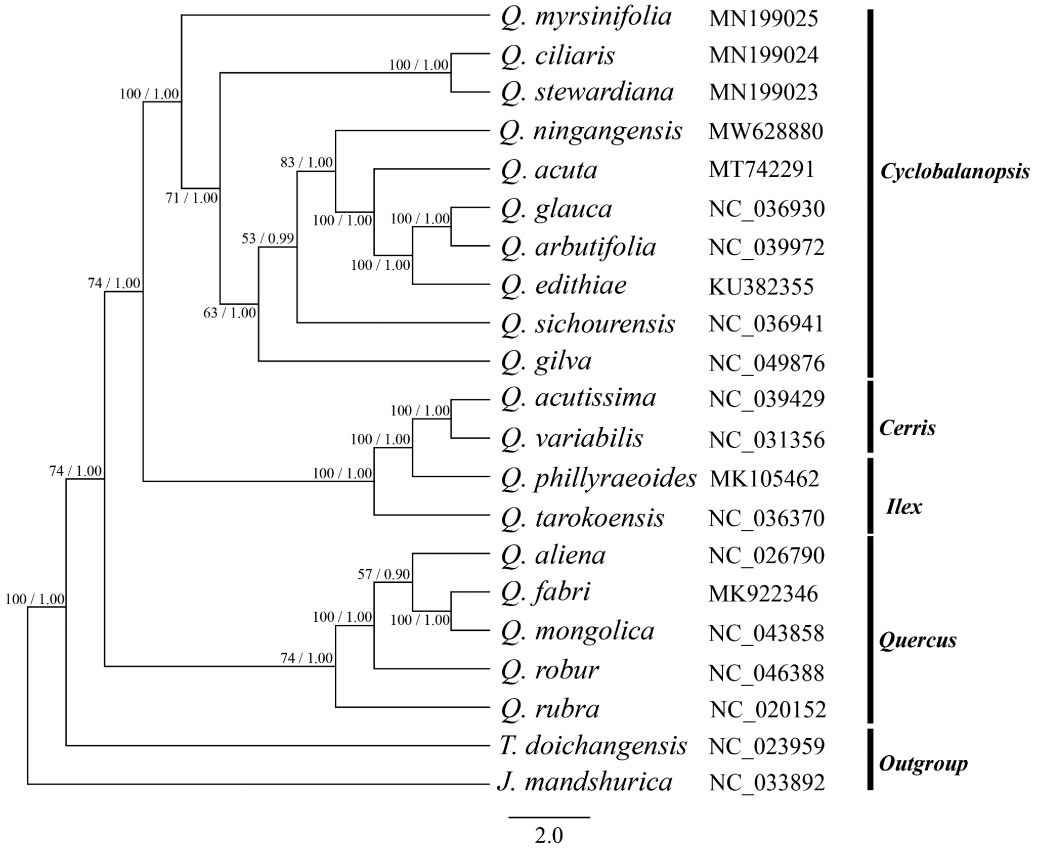

Figure 7. The phylogenetic relationships among 21 complete chloroplast genome homologous sequences based on Maximum likelihood (ML) and Bayesian Inference (BI) phylogenetic tree. Values besides the branch represented ML bootstrap and Bayesian posterior probabilities (BSP). GenBank accession numbers were listed following species name. *Trigonobalanus doichangensis* and *Juglans mandshurica* were used as the outgroups.

genomes (Morton & Clegg 1995; Morton et al. 1997). This bias may also be related to the genomic content and genetic characteristics of the codons (Yang & Yoder 1999; Morton 2003).

Although cp genomes are highly conserved in terms of genomic structure and order (Terakami et al. 2012; Wang et al. 2018), contraction and expansion of the IR regions are a common evolutionary phenomenon (Kim & Lee 2004; Huang et al. 2014), which may change the junction position of these regions. We found that the IR and coding regions exhibited higher conservation than SC and noncoding regions. In addition, the SSC/IR boundary regions were relatively stable instead of differences in the IR/SC boundary regions. For the conservation of IR regions, the substitution rates in SC regions have been found to be several times higher than those in IR regions among diverse plants (Zhu et al. 2015). A copy-dependent repair mechanism has been proposed to explain the lower substitution rate in IR regions (Perry & Wolfe 2002), which may be important for the evolution of cp genomes (Huang et al. 2014; Curci et al. 2015; Guo et al. 2017).

Phylogenetic relationship and taxonomy of *Q. ningangensis* and its close relatives

Chloroplast genomes have been proven effective for resolving different phylogenetic relationships in various land plants (Ma et al. 2014; Carbonell-Caballero et al. 2015). Previous phylogenetic studies have shown that there are two subgenera (*Cerris* and *Quercus*) in *Quercus* (Hipp et al. 2020). The phylogenetic tree of the cp genomes in this study supports the previous conclusion. Within the subgenus *Cerris*, it was also well resolved that section *Cyclobalanopsis* was sister to section *Cerris*/section *Ilex*, as in a previous study (Hipp et al. 2020). However, due to limited sampling, the relationship between section *Cerris* and section *Ilex* was not supported in this study.

Within the section *Cyclobalanopsis*, *Q. gilva* and *Q. sichourensis* did not form a CTB lineage sister to the STB lineage, as demonstrated in previous studies (Deng et al. 2018; Hipp et al. 2020; Yang et al. 2020). Our results showed that the phylogenetic position of *Q. ningangensis* was located between *Q. sichourensis* and *Q. acuta*. Based on the current study, it was difficult to provide an accurate relationship between *Q. ningangensis* and its close taxa. Thus, in order to explore the relationship among taxa in the section *Cyclobalanopsis*, cp genomes covering more taxa in this section need to be sequenced in the future.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (No. 31901217) and the Shanghai Municipal Administration of Forestation and City Appearances (G192422). We would like to thank Editage (www.editage.cn) for English language editing.

Supplementary electronic material

Table S1. Maximum likelihood and Bayesian inference fits of 24 different nucleotide substitution models. Download file

References

Alexander, L. W. & Woeste, K. E. 2014. Pyrosequencing of the northern red oak (*Quercus rubra* L.) chloroplast genome reveals high quality polymorphisms for population management. *Tree Genetics & Genomes* 10: 803–812. https://doi.org/10.1007/s11295-013-0681-1

Amiryousefi, A., Hyvönen, J. & Poczai, P. 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34: 3030–3031. https://doi.org/10.1093/bioinformatics/bty220

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477. https://doi.org.fjny.80599.net/10.1089/cmb.2012.0021

Bellgard, M., Schibeci, D., Trifonov, E. & Gojobori, T. 2001. Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains. *Journal of Molecular Evolution* 53: 465–468. https://doi.org/10.1007/s002390010236

Bi, G. Q., Mao, Y. X., Xing, Q. K. & Cao, M. 2018. HomBlocks: a multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics* 110: 18–22. https://doi.org/10.1016/j.ygeno.2017.08.001

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M. & Dopazo, J. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. *Molecular Biology & Evolution* 32: 2015–2035. https://doi.org/10.1093/molbev/msv082

Carrero, C., Jerome, D., Beckman, E., Byrne, A., Coombes, A. J., Deng, M., Rodriguez, A. G., Sam, H. V., Khoo, E., Nguyen, N., Robiansyah, I., Correa, H. R., Sang, J., Song, Y. G., Strijk, J., Sugau, J., Sun, W., Valencia-Avalos, S. & Westwood, M. 2020. *The red list of oaks 2020*. The Morton Arboretum: Lisle, IL.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* 106: 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chen, Y. X., Chen, Y. S., Shi, C. M., Huang, Z. B., Zhang, Y., Li, S. K., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X. Q., Wang, J., Yang, H. M., Fang, L. & Chen, Q. 2018. SOAPnuke: A mapreduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* 7: 1–6. https://doi.org/10.1093/gigascience/gix120

Cosner, M. E., Raubeson, L. A. & Jansen, R. K. 2004. Chloroplast DNA rearrangements in *Campanulaceae*: phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biology* 4: 27. https://doi.org/10.1186/1471-2148-4-27

Curci, P. L., Paola, D. D., Danzi, D., Vendramin, G. G. & Sonnante, G. 2015. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other *Asteraceae*. *PLoS ONE* 10: e0120589. https://doi.org/10.1371/journal.pone.0120589

Dane, F., Wang, Z. & Goertzen, L. 2015. Analysis of the complete chloroplast genome of *Castanea pumila* var. *pumila*, the Alleghany chinkapin. *Tree Genetics & Genomes* 11: 14. https://doi.org/10.1007/s11295-015-0840-7

Deng, M., Jiang, X. L., Hipp, A. L., Manos, P. S. & Hahn, M. 2018. Phylogeny and biogeography of East Asian evergreen oaks (*Quercus* section *Cyclobalanopsis*; *Fagaceae*): Insight into the Cenozoic history of evergreen broad-leaved forests in subtropical Asia. *Molecular Phylogenetics & Evolution* 119: 170–181. https://doi.org/10.1016/j.ympev.2017.11.003

Deng, M., Hipp, A., Song, Y. G., Li, Q. S., Coombes, A. & Cotton, A. 2014. Leaf epidermal features of *Quercus* subgenus *Cyclobalanopsis* (*Fagaceae*) and their systematic significance. *Botanical Journal of the Linnean Society* 176: 224–259. https://doi.org/10.1111/boj.12207

Denk, T., Grimm, G. W., Manos, P. S., Deng, M. & Hipp, A. L. 2017. An Updated Infrageneric Classification of the Oaks: Review of Previous Taxonomic Schemes and Synthesis of Evolutionary Patterns. In: Gil-Pelegrin E., Peguero-Pina J., Sancho-Knapik D. (eds) *Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L.* Tree Physiology, vol 7. Springer, Cham. https://doi.org/10.1007/978-3-319-69099-5_2

Doyle, J. J. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Fazan, L., Song, Y. G. & Kozlowski, G. 2020. The woody planet: from past triumph to manmade decline. *Plants* 9: 1593. https://doi.org/10.3390/plants9111593

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32: W273–W279. https://doi.org/10.1093/nar/gkh458

Guo, H. J., Liu, J. S., Luo, L., Wei, X. P., Zhang, J., Qi, Y. D., Zhang, B. G., Liu, H. T. & Xiao, P. G. 2017. Complete chloroplast genome sequences of *Schisandra chinensis*: Genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. *Science China Life Science* 60: 1286–1290. https://doi.org/10.1007/s11427-017-9098-5

Häggström, M. 2019. Being in the forest – A matter of cultural connections with a natural environment. *Plants People Planet* 1: 221–232. https://doi.org/10.1002/ppp3.10056

Han, Y. W., Duan, D., Ma, X. F., Jia, Y., Liu, Z. L., Zhao, G. F. & Li, Z. H. 2016. Efficient identification of the forest tree species in *Aceraceae* using DNA barcodes. *Frontiers in Plant Science* 7: 1707. https://doi.org/10.3389/fpls.2016.01707

He, L., Qian, J., Li, X. W., Sun, Z. Y., Xu, X. L. & Chen, S. L. 2017. Complete chloroplast genome of medicinal plant *Lonicera japonica*: Genome rearrangement, intron gain and loss, and implications for phylogenetic studies. *Molecules* 22: 249. https://doi.org/10.3390/molecules22020249

Hipp, A. L., Manos, P. S., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., Crowl, A. A., Deng, M., Denk, T., Fitz-Gibbon, S., Gailing, O., González-Elizondo, M. S., González-Rodríguez, A., Grimm, G. W., Jiang, X. L., Kremer, A., Lesur, I., McVay, J. D., Plomion, C., Rodríguez-Correa, H., Schulze, E. D., Simeone, M. C., Sork, V. L. & Valencia-Avalos, S. 2020. Genomic landscape of the global oak phylogeny. *New Phytologist* 226: 1198–1212. https://doi.org/10.1111/nph.16162

Hollingsworth, P. M., Graham, S. W. & Little, D. P. 2011. Choosing and using a plant DNA barcode. *PLoS ONE* 6: e19254. https://doi.org/10.1371/journal.pone.0019254

Huang, C. C., Chang, Y. T. & Bartholomew, B. 1999. *Fagaceae*. In: *Flora of China*, English version. Vol. 4, pp. 380–400. Science Press and Missouri Botanical Garden Press: Beijing, China and St. Louis, MO, USA.

Huang, H., Shi, C., Liu, Y., Mao, S. Y. & Gao, L. Z. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology* 14: 151. https://doi.org/10.1186/1471-2148-14-151

Jansen, R. K. & Ruhlman, T. A. 2012. *Plastid genomes of seed plants*. Springer Press, Berlin.

Kim, K. J. & Lee, H. L. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research* 11: 247–261. https://doi.org/10.1093/dnares/11.4.247

Kremer, A. & Hipp, A. L. 2020. Oaks: an evolutionary success story. *New Phytologist* 226: 987–1011. https://doi.org/10.1111/nph.16274

Li, X., Li, Y. F., Zang, M. Y., Li, M. Z. & Fang, Y. M. 2018. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *International Journal of Molecular Sciences* 19: 2443. https://doi.org/10.3390/ijms19082443

Li, Q. J., Su, N., Zhang, L., Tong, R. C., Zhang, X. H., Wang, J. R., Chang, Z. Y., Zhao, L. & Potter, D. 2020a. Chloroplast genomes elucidate diversity, phylogeny, and taxonomy of *Pulsatilla* (*Ranunculaceae*). *Scientific Reports* 10: 19781. https://doi.org/10.1038/s41598-020-76699-7

Li, Y., Wang, L., Liu, Q. L. & Fang, Y. M. 2020b. The complete plastid genome sequence of *Quercus ciliaris* (*Fagaceae*). *Mitochondrial DNA Part B* 5: 1954–1955. https://doi.org/10.1080/23802359.2020.1756955

Librado, P. & Rozas, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452. https://doi.org/10.1093/bioinformatics/btp187

Liu, C., Shi, L. C., Zhu, Y. J., Chen, H. M., Zhang, J. H., Lin, X. H. & Guan, X. J. 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* 13: 715. https://doi.org/10.1186/1471-2164-13-715

Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. 2013. OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* 41: W575–W581. https://doi.org/10.1093/nar/gkt289

Luo, R. B., Liu, B. H., Xie, Y. L., Li, Z. Y., Huang, W. H., Yuan, J. Y., He, G. Z., Chen, Y. X., Pan, Q., Liu, Y. J., Tang, J. B., Wu, G. X., Zhang, H., Shi, Y. J., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C. L., Cheung, D. W., Yiu, S. M., Peng, S. L., Zhu, X. Q., Liu, G. M., Liao, X. K., Li, Y. R., Yang, H. M., Wang, J., Lam, T. W. & Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1: 18. https://doi.org/10.1186/s13742-015-0069-2

Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H. & Li, D. Z. 2014. Chloroplast phylogenomic analysis resolve deep-level relationships of an intractable bamboo tribe *Arundinarieae* (*Poaceae*). *Systematic Biology* 63: 933–950. https://doi.org/10.1093/sysbio/syu054

Manos, P. S. & Stanford, A. M. 2001. The historical biogeography of *Fagaceae*: tracking the tertiary history of temperate and subtropical forests of the Northern Hemisphere. *International Journal of Plant Sciences* 162: S77–S93. https://doi.org/10.1086/323280

Meng, Y., Wen, J., Nie, Z. L., Sun, H. & Yang, Y. P. 2008. Phylogeny and biogeographic diversification of *Maianthemum* (*Ruscaceae*: Polygonatae). *Molecular Phylogenetics & Evolution* 49: 424–434. https://doi.org/10.1016/j.ympev.2008.07.017

Morton, B. R. 2003. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *Journal of Molecular Evolution* 56: 616–629. https://doi.org/10.1007/s00239-002-2430-1

Morton, B. R. & Clegg, M. T. 1995. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *Journal of Molecular Evolution* 41: 597–603. https://doi.org/10.1007/bf00175818

Morton, B. R., Oberholzer, V. M. & Clegg, M. T. 1997. The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *Journal of Molecular Evolution* 45: 227–231. https://doi.org/10.1007/pl00006224

Mu, X. Y., Tong, L., Sun, M., Zhu, Y. X., Wen, J., Lin, Q. W. & Liu, B. 2020. Phylogeny and divergence time estimation of the walnut family (*Juglandaceae*) based on nuclear RAD-Seq and chloroplast genome data. *Molecular Phylogenetics & Evolution* 147: 106802. https://doi.org/10.1016/j.ympev.2020.106802

Muse, S. V. & Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology & Evolution* 11: 715–724. https://doi.org/10.1093/oxfordjournals.molbev.a040152

Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. *Molecular Biology & Evolution* 32: 268–274. https://doi.org/10.1093/molbev/msu300

Nixon, K. C. 1997. *Quercus.* In: *Editorial Committee ed. Flora of North America North of Mexico*. pp. 445–447. New York, NY, USA: Oxford University Press.

Patel, R. K. & Jain, M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7: e30619. https://doi.org/10.1371/journal.pone.0030619

Perry, A. S. & Wolfe, K. H. 2002. Nucleotide substitution rates in Legume chloroplast DNA depend on the presence of the inverted repeat. *Journal of Molecular Evolution* 55: 501–508. https://doi.org/10.1007/PL00020998

Posada, D. & Crandall, K. A. 1998. Modeltest: testing the model DNA substitution. *Bioinformatics* 14: 817–818. https://doi.org/10.1093/bioinformatics/14.9.817

Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Song, Y. G., Petitpierre, B., Deng, M., Wu, J. P. & Kozlowski, G. 2019. Predicting climate change impacts on the threatened *Quercus arbutifolia* in montane cloud forests in southern China and Vietnam: Conservation implications. *Forest Ecology and Management* 444: 269–279. https://doi.org/10.1016/j.foreco.2019.04.028

Sueoka, N. & Kawanishi, Y. 2000. DNA G+C content of the third codon position and codon usage biases of human genes. *Gene* 261: 53–62. https://doi.org/10.1016/S0378-1119(00)00480-7

Terakami, S., Matsumura, Y., Kurita, K., Kanamori, H., Katayose, Y., Yamamoto, T. & Katayama, H. 2012. Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. *Tree Genetics & Genomes* 8: 841–854. https://doi.org/10.1007/s11295-012-0469-8

Thiel, T., Michalek, W., Varshney, R. K. & Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical & Applied Genetics* 106: 411–422. https://doi.org/10.1007/s00122-002-1031-0

Timme, R. E., Kuehl, J. V., Boore, J. L. & Jansen, R. K. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (*Asteraceae*) plastid genomes: Identification of divergent regions and categorization of shared repeats. *American Journal of Botany* 94: 302–312. https://doi.org/10.3732/ajb.94.3.302

Uckele, K. A., Adams, R. P., Schwarzbach, A. E. & Parchman, T. L. 2021. Genome-wide RAD sequencing resolves the evolutionary history of serrate leaf Juniperus and reveals discordance with chloroplast phylogeny. *Molecular Phylogenetics & Evolution* 156: 107022. https://doi.org/10.1016/j.ympev.2020.107022

Wang, Y. H., Wicke, S., Wang, H., Jin, J. J., Chen, S. Y., Zhang, S. D., Li, D. Z. & Yi, T. S. 2018. Plastid genome evolution in the early-diverging legume subfamily *Cercidoideae* (*Fabaceae*). *Frontiers in Plant Science* 9: 138. https://doi.org/10.3389/fpls.2018.00138

Weng, M. L., Blazier, J. C., Govindu, M. & Jansen, R. K. 2013. Reconstruction of the ancestral plastid genome in *Geraniaceae* reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. *Molecular Biology & Evolution* 31: 645–659. https://doi.org/10.1093/molbev/mst257

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. 2011. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Molecular Biology* 76: 273–297. https://doi.org/10.1007/s11103-011-9762-4

Yang, J., Guo, Y. F., Chen, X. D., Zhang, X., Ju, M. M., Bai, G. Q., Liu, Z. L. & Zhao, G. F. 2020. Framework phylogeny, evolution and complex diversification of Chinese oaks. *Plants* 9: 1024. https://doi.org/10.3390/plants9081024

Yang, Y. C., Zhu, J., Feng, L., Zhou, T., Bai, G. Q., Yang, J. & Zhao, G. F. 2018. Plastid genome comparative and phylogenetic analyses of the key genera in *Fagaceae*: highlighting the effect of codon composition bias in phylogenetic inference. *Frontiers in Plant Science* 9: 82. https://doi.org/10.3389/fpls.2018.00082

Yang, Z. H. & Yoder, A. D. 1999. Estimation of the transition/transversion rate bias and species sampling. *Journal of Molecular Evolution* 48: 274–283. https://doi.org/10.1007/pl00006470

Yu, T., Gao, J., Huang, B. H., Dayananda, B., Zhang, Y. Y., Liao, P. C. & Li, J. Q. 2020. Comparative plastome analyses and phylogenetic applications of the *Acer* section *Platanoidea*. *Forests* 11: 462. https://doi.org/10.3390/f11040462

Zhang, R. S., Yang, J., Hu, H. L., Xia, R. X., Li, Y. P., Su, J. F., Li, Q., Liu, Y. Q. & Qin, L. 2020. A high level of chloroplast genome sequence variability in the Sawtooth oak *Quercus acutissima*. *International Journal of Biological Macromolecules* 152: 340–348. https://doi.org/10.1016/j.ijbiomac.2020.02.201

Zhao, F., Chen, Y. P., Salmaki, Y., Drew, B. T., Wilson, T. C., Scheen, A. C., Celep, F., Brauchler, C., Bendiksby, M., Wang, Q., Min, D. Z., Peng, H., Olmstead, R. G., Li, B. & Xiang, C. L. 2020. An updated tribal classification of *Lamiaceae* based on plastome phylogenomics. *BMC Biology* 19: 2. https://doi.org/10.1186/s12915-020-00931-z

Zhu, A. D., Guo, W. H., Gupta, S., Fan, W. S. & Mower, J. P. 2015. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist* 209: 1747–1756. https://doi.org/10.1111/nph.13743