

# Comparative Complete Chloroplast Genome of Nine Species in *Litsea* (Lauraceae) and Uncover Evolutionary Dynamic Patterns

**Weicai Song**

Qingdao University of Science and Technology

**Zimeng Chen**

Qingdao University of Science and Technology

**Qi Feng**

Qingdao University of Science and Technology

**Chuxuan Ji**

Imperial College London

**Chengbo Wei**

Peking University Shenzhen Graduate School

**Chao Shi** (✉ [chaoshi@qust.edu.cn](mailto:chaoshi@qust.edu.cn))

Qingdao University of Science and Technology

**Michael S. Engel**

University of Kansas

**Shuo Wang**

Qingdao University of Science and Technology

---

## Research Article

**Keywords:** *Litsea*, chloroplast genome, structural variations, genetic relationship

**Posted Date:** November 12th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-1068132/v1>

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

[Read Full License](#)

---

# Abstract

**Background:** *Litsea*, Lauraceae, is a group of evergreen trees or shrubs that widely distributed in tropical and subtropical countries, such as Asia and America. Species in *Litsea* are spontaneously distributed at a maximum altitude of 2,700 m from sea level. Plants and its extractions from *Litsea* species cover a wide range of medicinal and industrial values. The aromatic oil extracted from *Litsea* is of great value with citral as its main component. At present, studies related to gene resources of *Litsea* are limited in the morphological analysis, while studies at the genetic level are insufficient. We therefore firstly assembled and annotated the complete chloroplast genome of nine species in *Litsea*, carried out a series of comparative analysis, and completed the construction of phylogenetic tree within genus *Litsea*.

**Results:** The genome length ranged from 152,051 to 152,717 bp. A total of 128 genes were identified, including 84 protein-coding genes, 36 rRNA genes and 8 tRNA genes. High consistency of codon bias, repeats, divergent analysis, single nucleotide polymorphisms (SNP) and insertions and deletions (InDels) revealed highly conserved chloroplast phenotypes in species within the genus *Litsea*. Changes in gene length and the present of pseudogene *ycf1<sup>ψ</sup>* that caused by IR contraction and expansion were reported. The non-coding regions, especially *atpF - atpH* and *ndhC - trnV-UAC* presented high gene divergence. *PsbJ - psbE* regions showed remarkably high nucleotide diversity (Pi) values. Furthermore, we constructed two phylogenetic trees, demonstrating two dominant clades within genus *Litsea*. And the differences between trees constructed by full chloroplast (cp) genome and protein-coding genes were revealed.

**Conclusion:** Overall, the evolutionary pattern of *Litsea* species regarding structural features, repeats sequences and variations presented high consistency. Valuable genomic resources and theoretical basis were also provided for further research of taxonomic discrepancies, molecular marker-assisted breeding and phylogenetic relationships of *Litsea* and other angiosperm species.

## Introduction

*Litsea* is an evergreen tree or shrub and is one of the most diverse genera (about 400 species) in the family Lauraceae (Mesangiospermae: Magnoliids: Laurales). It is widely distributed in tropical and subtropical Asia, North and South America [1,2], and 74 species of which are located in China, at a maximum elevation of 2,700 m above sea level [3]. Species of *Litsea* are utilized in a wide range of applications, covering medical, agricultural, industrial, and many other fields. *Litsea* can be used to treat a variety of conditions such as diarrhea, stomach pain, indigestion, the common cold, gastroenteritis, diabetes, edema, arthritis, asthma, pain, and trauma [2]. In addition, *Litsea* is also known for the highly effective properties of its essential oil against food-borne pathogens [4]. Its essential oils can also be resistant to several types of bacteria, has antioxidant, anti-parasitic, acute toxicity, genotoxic, and cytotoxic properties, and can even prevent several types of cancer [5–7]. Despite the pharmaceutical applications of *Litsea*, it is also widely used as feed for silkworm pupae, especially for muga silk worms (*Antheraea assama*) [5]. By comparison with ordinary silk produced from other food sources, muga silk produced from *Litsea* possesses a higher value and is considered of better quality, as reflected in its

creamy and lustrous appearance and texture. Some representative species of *Litsea* are industrially important and have been utilized extensively [6]. For instance, *Litsea cubebais* is a spice shrub of considerable economic importance. The essential oil prepared from citric acid extracted from the plant's body is a natural spice, with a wide number of potential applications. Moreover, it is also an important raw material for the synthesis of vital compounds, such as vitamin A [7].

Chloroplasts are organelles that occur in green plants and algae, taking the responsibility for photosynthesis and other housekeeping functions. Additionally, they are essential for nitrate and sulfate assimilation as well as the synthesis of amino acids, fatty acids, chlorophyll, and carotenoids [8]. In general, chloroplast (cp) genomes have a conservative genome structure, gene content, and gene order in most monocotyledon plants [8,9]. The complete chloroplast genome of angiosperms is usually composed of four parts: a large single-copy (LSC) region, a small single-copy (SSC) region, and two similar inverted repeat (IR) regions, with a highly conservative structure. The cp genome consists of 110 to 130 genes primarily involved in photosynthesis, transcription, and translation [8,9]. Contraction and expansion of IR regions and gene and intron loss events have also occurred commonly during evolution [10]. The sequences of cp genomes can provide information for genetic relationship, gene transfer, cloning, and species domestication. The cp genome of advanced plants is inherited from a single parent [11], which can be used as an effective barcode for species identification as well as the development of other potential identification markers [12]. Identification of cp genomes promote the sustainable development of plant species, their utilization in a more rigorous scientific manner, as well as for species conservation [13–15].

Currently, evolutionary studies on *Litsea* mostly focus on the exploration of phylogenetic relationships between *Litsea* and related genera [16,17], and lack a deep consideration of relationships within the genus itself [18]. Genetic resources for *Litsea* need also to be supplemented. Therefore, a detailed assembly and annotation of the complete cp genomes of various species within *Litsea* will greatly enrich the existing database, deepen the genetic recognition of the genus, and contribute to phylogenetic, evolutionary, developmental, conservation, and taxonomic investigations. Advancing our taxonomic knowledge for *Litsea* will enable us to refine conservation efforts and the utilization of natural resources, providing sufficient genetic resources for artificial breeding and drug development. In this study, we first sequenced and assembled the complete cp genomes of nine species of *Litsea*. A comparative analysis was performed, including gene types, GC content in different regions, codon usage, IR junction, types and location of simple sequence repeats (SSR), as well as nucleotide diversity (Pi). The results provide informative and valid data regarding genotype and suitable DNA markers. Moreover, using 21 species from *Litsea*, evolutionary relationships within the genus were analyzed using the complete cp genome as well as protein-coding sequences. Ultimately, this study provides a reliable resource for further utilization and conservation of genetic resources for *Litsea*.

## Results And Discussion

### 2.1. Chloroplast genome features of *Litsea*

The cp genome features of nine species were analyzed and the total length ranged from 152,051 bp to 152,747 bp (Fig. 1). 128 genes were found in these complete cp genomes, including 36 tRNA genes, 8 rRNA genes, and 84 protein-coding genes. These genes can be divided into three categories: self-replication related, photosynthesis related, and other genes. Large subunit of ribosomal proteins, small subunit of ribosomal proteins, DNA-dependent RNA polymerase, rRNA genes, and tRNA genes belong to the Self-replication category; Photosystem I, Photosystem II, NADH oxidoreductase, Cytochrome b6/f complex, ATP synthase, and Rubisco belong to the Photosynthesis category; while the remaining genes that have not been authorially classified yet were attributed to the other genes category (Table 1) [19].

Typical quadripartite and circular structures were discovered. These cp genomes contain a large single-copy (LSC) of 93,093 – 93,631 bp, a small single-copy region (SSC) of 18,813 – 18,902 bp, separated by two identical interspersed regions (IRs) of 20,014 – 20,117 bp. Among four types of regions, the LSC region contains the largest number of genes, including 66 protein-coding genes and 23 tRNA genes. The SSC region contained only 11 protein-coding genes and one tRNA gene, but its average gene length was the longest at 1,100 bp, far exceeding that of the LSC and IR regions, each with an average length of 625 and 639 bp, respectively. Two identical IR regions contained 5 protein-coding genes, 6 tRNA genes, along with 4 rRNA genes (Table 1). The genome features of *Litsea* are consistent with the basic structure of chloroplasts reported by other studies [20].

We also analyzed the GC content of the complete cp genome for the nine species of *Litsea*, as well as the values of each region (Table 2). We discovered that the average GC content of the full cp genome was 39.2% for all species except for *L. sericea*, which was 39.1%. In addition, the GC content of the IR region was firmly consistent at 44.4% and significantly higher than the other two regions, which was assumed to be related to the presence of many rRNA genes [21].

## 2.2. Codon usage analysis

All organisms share a common codon table, reflective of the shared ancestry of all life, but through time disproportionate biases have evolved in various clades. Different species exhibit certain preferences for different synonymous codons, and even different proteins within the same species may show a preference for the same amino acid, a phenomenon called *codon bias*. A measurement called Relative Synonymous Codon Usage (RSCU) removes the effect of amino acid composition on codon usage [20]. The codon usage and RSCU value of coding sequences (CDSs) are reported for the species examined in this study (Table S1). The protein-coding genes in the complete chloroplast genome of *L. moupinensis* consist of 84 genes coded by 61 codons, which encode 20 amino acids. The results showed that Leu (UUA), Ala (GCU), and Arg (AGA) are the most frequently used amino acids, while Ser (AGC) and Arg (CGC) were the least abundant amino acids (Fig. 2). RSCU values greater than 1 mean that there is significant codon bias. This results in a different use of amino acids, which correlates with protein properties and functions [21]. Analysis of RSCU values of the codons encoding each amino acid revealed that most codons with RSCU > 1 contained either an A- or G-terminal. By contrast, RSCU values for

codons that ended with a C-terminal, such as CGC (Arg), UGC (Cys), CAC (His), and AGC (Ser), are relatively low. This result was consistent with previous studies [22].

### 2.3. Long repeat and SSR analysis

Simple sequence repeats (SSRs), also known as microsatellites, commonly exist throughout the cp genome, consisting of one to six nucleotide repeats (Cui et al., 2019). Due to its variability at the intraspecific level, SSRs are commonly used as markers in population genetic analyses [23,24]. In the cp genome of nine species, the total number of repeats ranged from 109 (*L. chuni*) to 119 (*L. auriculata*) (Table S2). 111 SSRs were detected from the cp genome of the representative species *L. moupinensis*, including 62 mononucleotide, 36 dinucleotide, 3 trinucleotide, 8 tetranucleotide, 1 pentanucleotide, and 1 hexanucleotide repeats. In general, the SSR number decreased along with the increase in nucleotide number. The percentage of tri-, tetra-, penta-, and hexa-nucleotide repeat sequences detected were remarkably lower than that of mono- and di-nucleotide repeats, a phenomenon reported previously [25]. Mono-nucleotide repeats, the largest class of SSRs and consisting of 56.97% of all repeats, are notably rich in A/T bases, causing the differences in terms of base content, which is quite similar to that of other angiosperm species [26]. We also analyzed the distribution of SSRs in LSC/SSC/IR regions. The number of SSR markers in the LSC region of nine species of *Litsea* ranged from 79 to 87, far exceeding that of SSC (19) and IR regions (12). In particular, IR region contains the lowest number of SSRs, which further demonstrates the high degree of conservatism of IR regions.

Some repeats larger than 30 bp in length are called long repeat sequences, which increase the rearrangement of the cp genome [27]. We detected interspersed repeated sequences (IRs) including four types of long repeat sequences: complement repeats (C), forward repeats (F), palindromic repeats (P), reverse repeats (R). Among all types of repeats detected, palindromic repeats (P, 16) were richest in most species, followed by forward repeats (F), with an average number of 12.5 and reverse repeats (R) at 5.8 (Table S3). Complement repeats (C) were notably rare among all species. However, in the cp genome of *L. ichangensis*, the number of forward repeats (F, 17) were slightly higher than that of palindromic repeats (P, 16). What more, in the cp genome of *L. auriculata*, the ratio of reverse repeats (R) was more than that of forward repeats (F), which is also different from the other eight species (Fig. 3B). We also measured the number of long repeat sequences with different lengths (Fig. 3C). It was found that long repetitive sequences of length of 20 – 21 bp were most common, while the remainder decreased in number with an increase in length, with two exceptions for 33 bp and 44 bp. Notably, the repeats with 29, 31, and 38 bp in length were almost absent, indicating that mutations may have occurred during evolution in the corresponding species.

### 2.4. IR contraction analysis and genome divergence between the *Litsea* species

The contraction and expansion of IR regions contribute greatly for to variations of cp genomes among different species, resulting in gene duplication, deletion, and the generation of pseudogenes. Studying the characteristic genes of the border region contributes to species identification and phylogenetic analyses [28]. In this study, we analyzed and visualized the genes located in the junction region of LSC and IRa

(JSa) as well as the junction of SSC and IRb (JSb) in the cp genome of the nine species of *Litsea* (Fig. 4). JLa represents the junction between LSC and IRa, and the same applies for JLb. In this study, we observed that genes located in the junction of four regions were highly conserved, with only a few variations. Most genes located at cp genome junctions in all nine species differed only in the distance to their corresponding boundaries, such as *ycf2*, *ndhF*, *trnH*, and *psbA*. To be more specific, the *ycf2* gene spans LSC/IRb and is distributed in both regions of similar length, with the LSC region being slightly longer. The *ndhF* gene existed among nine species, completely located in SSC and a short distance from IRb except for that of *L. sericea*, which was longer and closer to the JSb boundary. The *trnH* gene was located in the LSC region, adjacent to the IRa/LSC border, and was 21 – 22 bp in length. *PsbA* was located entirely in the LSC region. Yet, notable variations were found. The *ycf1* gene was absent in this junction, while the remaining eight species contain *ycf1* $\Psi$  (pseudocopy, 5' end missing) in JSb, which spans JSb with only 4 – 5 bp of length located in SSC. Apart from that, the contraction and expansion event located in the JSb was greater than that of the JLa boundary. This pattern is consistent with previous studies [29].

The whole sequence identity plot of nine species within *Litsea* was analyzed using mVISTA with *L. garretti* set as the reference sequence for comparison (Fig. 5). Genome sequences of the nine *Litsea* exhibited a high degree of concordance. In this study, we revealed that most of the variations in the cp genome of different species were distributed in CNS (non-coding sequences) region, and there were two distinct slips in the genome sequence alignment diagram. Notable high-divergent regions in CNS were *atpF - atpH* and *ndhC - trnV-UAC*, the divergent value of which exceeded 100%. Other variant regions include *rps16 - trnQ-UUG*, *ycf4 - cemA*, *rps8 - rpl14*, *rps12 - trnV-GAC*. Some of the coding genes, such as *ndhK*, *ndhF*, and *ycf1*, were found to be highly divergent. In general, the divergence in the IR region was significantly smaller than that in the LSC and SSC regions, a result comparable to the divergence analysis.

## 2.5. Nucleotide divergence and SNPs

The highly variable area in the cp genomes of the nine species was studied by a sliding window analysis and variation patterns are depicted in figure 6. The average Pi value in the SSC fragment was the highest, whereas the LSC region fluctuated at a little lower value. However, a spike containing *psbJ*, *psbL*, *psbF*, and *psbE* genes appeared in the LSC region, which is the peak with the largest Pi value in the entire cp genome. At the same time, the nucleotide diversity at the IR fragment was the lowest, indicating a high degree of conservatism, a conclusion supported by previous studies [30]. Moreover, the nucleotide divergent value of some genes was relatively high as well, such as *ycf1*, *ycf2*, *matK*, *ndhA*, *ndhF*, *rpoC2*, *trnG-UCC*, and *trnK-UUU*. These results were in line with the findings of previous studies [31,32]

To further explore the divergence of nucleotides, we compared and analyzed Single Nucleotide Polymorphisms (SNPs) and Insertions and Deletions (InDels) of nine species within *Litsea*. The polymorphism ratio of transition substitution (Ts) was higher than transversion substitutions (Tv) in the LSC region of nine cp genomes (Table 3). The most substitutions were located in the LSC region, while IR

regions contained the lowest rate of polymorphisms. This result is consistent with previous studies [33]. In terms of transition substitutions, the polymorphism ratios of A/G and C/T were almost the same, although the former took up a slightly larger proportion, with only three exceptions (*L. auriculata*, *L. chunii*, and *L. tsinlingensis*). As for transversion substitutions, the polymorphism ratios of A/T and C/G were greatly lower than that of A/C and G/C substitutions. The same pattern applied for InDels (Table 4). LSC presented the largest number of InDels in comparison with IR and SSC regions, while the average length of InDels in IR regions was the longest, with the longest variation length at 678 bp (*L. tsinlingensis*). It is worth mentioning, in the cp genome of *L. auriculata*, the average length of InDels located in the IR regions contained a considerable number of small InDels rather than only several long InDels, as was found in the other eight species, causing its average length to be three times shorter than others, indicating that *L. auriculata* may have experienced some degree of mutation during its evolution that differed from related species.

## 2.6. Phylogenetic analysis

The expanding cp genome database provides an important basis for determining evolutionary relationships [30,34]. Phylogenetic trees based on different data had slightly varied topologies, with trees based on the whole cp genome and CDS data having the same topology, and being more credible than trees based on the IR area and introns [35–38]. We found two similar topological structures with few changes based on the full cp genome and the protein-coding sequences of 23 selected species, with *Neolitsea sericea* and *Actinodaphne obovate* as outgroup species (Table S3, Fig 7).

In general, the entire phylogenetic tree was divided into three main branches, with the two outgroup species representing two distinct branches, each with high bootstrap values. The first subclade consists of 11 species: *L. moupinensis*, *L. rubescens*, *L. populifolia*, *L. veitchiana*, *L. pungens*, *L. sericea*, *L. ichangensis*, *L. chunii*, *L. tsinlingensis*, *L. acutivena*, *L. glutivena*, and *L. auriculata*. Among them, the clade of *L. chunii* and *L. tsinlingensis*, and the clade of *L. acutivena* and *L. glutinosa* form sister pairs, respectively. Notably, *L. pungens* switched phylogenetic positions with *L. sericea*, with relatively low bootstrap values in both trees. Another clade included 10 species: *L. cubeba*, *L. mollis*, *L. dilleniifolia*, *L. szemaois*, *L. auriculata*, *L. coreana*, *L. monpinensis*, *L. garrettii*, *L. elongata*, and *L. japonica*. Among them, *L. cubeba* and *L. mollis* were grouped as sisters and clustered with eight other species. It is worth noting that in topology based on the complete cp genome, *L. coreana* and *L. monopetala* were sisters with low support (only 57). However, in the CDS-based tree, *L. dilleniifolia* and *L. szemaois* split into a clade that aggregated with the remaining four species (*L. monpinensis*, *L. garrettii*, *L. elongata*, and *L. japonica*), and merged with *L. coreana* to converge as a single branch. In other words, in the two different analyses, the clade consisting of *L. dilleniifolia* and *L. szemaois* switched its position with *L. coreana*.

We assumed that the variations presented by the different trees may be rooted in the changes occurring in the IGS (intergenic spacer) regions, clarifying the importance of a greater knowledge of non-coding regions. Despite minor differences, the phylogenetic relationships of most species in the two topologies

were consistent, showing similar genetic affinities in the topology, and which aligned nicely with the elevational distribution of the species [39–41].

## Conclusion

We sequence and report complete cp genome sequences from nine species of *Litsea*, revealing typical quadripartite and circular structures. The cp genome size of nine *Litsea* ranged from 152,051 bp to 152,747 bp. We performed long repeat and SSR analyses of the complete cp genomes of nine species of *Litsea* to better study practical gene markers. *Litsea auriculata* contained the most SSR sequences, while *L. chunii* had the least. Codon bias was observed in the protein-coding region. In the comparative analysis, differences between species occurred mainly in non-coding sequences, except for a few highly divergent coding genes, such as *ycf1*. We also observed the contraction and expansion of IR boundaries, which caused gene loss, changes in gene length, and the occurrence of pseudogenes, resulting in differences between the species. In terms of nucleotide differences, the LSC region had the largest number of nucleotide variants, but the variation frequency was lower than that of the SSC region. The IR regions showed a high degree of conservation. Phylogenetic relationships within the genus were explored using two sets of data from the complete cp genome and another from 84 protein-coding genes for 21 species of *Litsea* and two outgroup species. Essentially the same conclusions were obtained: *L. moupinensis* and *L. rubescens*, *L. chunii* and *L. tsinlingensis* were sisters in the phylogenies and showed similar genetic relationships consistent with their elevational distributions. This study provides aid to taxonomic studies for *Litsea*, providing specific genetic markers for taxon identification and for inferring evolutionary relationships among the species. These data may also contribute to future conservation efforts as well as the practical use of these species.

## Materials And Methods

### 4.1. Sample collection, DNA extraction, and sequencing

In this study, the nine species of *Litsea* were collected from Plant Germplasm and Genomics Center, Kunming Institute of Botany, the Chinese Academy of Sciences, and was approved by Kunming Institute of Botany and local policy. Fresh leaf tissue was collected without apparent disease symptoms and preserved in silica gel. Total genomic DNA was extracted from fresh leaves using modified CTAB [42], and the quantity and quality of the extracted DNA was assessed by spectrophotometry while the integrity was evaluated using a 1% (w/v) agarose gel electrophoresis [18]. The Illumina TruSeq Library Preparation Kit (Illumina, San Diego, CA, USA) was used to prepare approximately 500 bp of paired-end libraries for DNA inserts, according to the manufacturer's protocol. These libraries were sequenced on the Illumina HiSeq 4000 platform in Novogene (Beijing, China), generating raw data of 150 bp paired-end reads. About 22.5 Gb high quality, 2 × 150 bp pair-end raw reads were obtained and were used to assemble the complete cp genome of *Litsea*.

### 4.2. Chloroplast genome *de novo* assembly and annotation



The raw data were preprocessed using Trimmomatic 0.39 software [43], including removal of adapter sequences and other sequences introduced during sequencing, removal of low-quality and over-N-base reads, etc. The quality of newly produced clean short reads was assessed using FASTQC v0.11.9 [44] and MULTIQC software [45], and high-quality data with Phred scores averaging above 35 were screened out. According to the reference sequence (*Litsea glutinosa*), the chloroplast-like reads were isolated from clean reads by BLAST [46]. Short reads were *de novo* assembled into long contigs with SOAPdenovo 2.04 [47] by setting kmer values of as 35, 44, 71, and 101. Finally, the long-contigs complete sequence expansion and gap filling using Geneious ver 8.1 [48], which forms the complete cp genome. The complete cp genome was further validated and calibrated by using *de novo* splicing software NOVOplsty 4.2 [49]. GeSeq [50] was used to annotate the assembled genomes, and tRNAscanSE ver 1.21 [51] was applied to detect tRNA genes with default settings, and RNAmmer [52] was used to validate rRNA genes with default settings. As a final check, we compared the results with the reference sequence and corrected misannotated genes by GB2Sequin [53] in an artificial manner. The circular map of the genomes was drawn by using Organellar Genome DRAW (OGDRAW) [54]. The nine newly assembled *Litsea* cp genomes were deposited in GenBank with the accession numbers **MW802253–MW802261**.

### 4.3. Analysis of chloroplast genome characteristics

Information regarding GC content, genome length, and number of each region in cp genomes was obtained using Geneious Prime software [48]. RSCU (Relative synonymous codon usage) was calculated by Computer Codon Usage Bias function in MEGA X [55]. SSRs were identified using MISA [56], with a setting of ten repeats for mononucleotide SSRs, four for dinucleotide and trinucleotide SSRs, and three for tetranucleotide, pentanucleotide, and hexanucleotide SSRs. REPuter [57] was used to identify four types of repeats with the minimum repetition unit set as 20 bp and the maximum as 300, and the remaining options set to default parameters.

### 4.4. Whole genome alignment

To compare the gene differences among the nine species, *L. garrettii* was selected as the reference species and the online comparison tool mVISTA [58] was used for sequence alignment. IRscope [39] was used to detect and visualize the contraction and expansion of IRs boundaries. We also used DNAsp 6 software [60] to calculate the nucleotide diversity ( $\Pi$ ) of the cp genome with 1000 bp window length and 50 bp step size.

### 4.5. Phylogenetic analysis

We downloaded 12 further chloroplast genomes of *Litsea* from NCBI (National Center for Biotechnology Information). Two species from Lauraceae but in different genera, *Actinodaphne obovate* and *Neolitsea sericea*, were selected as outgroups to root our phylogenetic networks. A total of 23 species were compared for phylogenetic evaluation. MAFFT v7 was used to perform multiple genome alignment [61], and we used the complete cp genome sequence data as well as a separate dataset of 84 protein-coding genes to construct individual maximum likelihood (ML) topologies. The MPL analyses were performed

using MEGA X [55], and bootstrap tests were performed with 1000 replicates with tree bisection-reconnection branch swapping.

## Abbreviations

AGE, agarose gel electrophoresis; BI, Bayesian inference; CDS, Coding sequences; Cp, chloroplast; CTAB, Cetyltrimethylammonium bromide; DOGMA, Dual Organellar Genome Annotator; IGS, Intergenic spacer; InDels, insertions/deletions; IR, inverted repeat; Ka, Nonsynonymous; Ks, Synonymous; LSC, large single-copy; ML, Maximum likelihood; NJ, neighbor joining; PCGs, Protein-coding genes; Pi, Nucleotide variance; RSCU, relative synonymous codon usage; SBS, Sequencing By Synthesis; SNP, single nucleotide polymorphisms; SSC, small single-copy; SSR, simple sequence repeat;

## Declarations

### Data accessibility

The data that support the findings of this study are openly available in the Genbank database at <https://www.ncbi.nlm.nih.gov/>, under accession number [MW802253–MW802261].

### Credit authorship contribution statement

Weicai Song: Conceptualization, Investigation, Writing - review & editing. Zimeng Chen: Data curation, Writing - original draft. Qi Feng: Conceptualization, Investigation. Chuxuan Ji, Chengbo Wei: Writing - review & editing. Michael S. Engel: Writing - review & editing. Chao Shi, Shuo Wang: Conceptualization, Data curation, Resources.

### Declaration of Competing Interest

The authors declare no conflict of interest associated with the work described in this manuscript.

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (NO. 31801022 and NO. 31701090) and Shandong Province Natural Science Foundation of China (NO. ZR2019BC094). We are thankful to Beijing-based Novogene for their NGS service that was instrumental to the execution of the project.

## References

[1] Kong DG, Zhao Y, Li GH, Chen BJ, Wang XN, Zhou HL, et al. The genus *Litsea* in traditional Chinese medicine: an ethnomedical, phytochemical and pharmacological review. *Journal of Ethnopharmacology*. 2015;164:256–264. <https://doi.org/10.1016/J.JEP.2015.02.020>.

- [2] Wang YS, Wen ZQ, Li BT, Zhang HB, Yang JH. Ethnobotany, phytochemistry, and pharmacology of the genus *Litsea*: An update. *Journal of Ethnopharmacology*. 2016;181:66–107. <https://doi.org/10.1016/J.JEP.2016.01.032>.
- [3] Wu ZY, Raven PH, Missouri BG. *Flora of China* 1994.
- [4] Tyagi AK, Malik A. Antimicrobial potential and chemical composition of *Eucalyptus globulus* oil in liquid and vapour phase against food spoilage microorganisms. *Food Chemistry*. 2011;126:228–235. <https://doi.org/10.1016/J.FOODCHEM.2010.11.002>.
- [5] Choudhury S, Ahmed R, Barthel A, Leclercq PA. Composition of the stem, flower and fruit oils of *litseacubeba* Pers. From two locations of Assam, India. *Journal of Essential Oil Research*. 1998;10:381–386. <https://doi.org/10.1080/10412905.1998.9700927>.
- [6] Kajaria DK, Gangwar M, Kumar D, Sharma AK, Tilak R, Nath G, et al. Evaluation of antimicrobial activity and bronchodilator effect of a polyherbal drug-Shrishadi. *Asian Pacific Journal of Tropical Biomedicine*. 2012;2:905. [https://doi.org/10.1016/S2221-1691\(12\)60251-2](https://doi.org/10.1016/S2221-1691(12)60251-2).
- [7] Kamle M, Mahato DK, Lee KE, Bajpai VK, Gajurel PR, Gu KS, et al. Ethnopharmacological properties and medicinal uses of *Litsea cubeba*. *Plants*. 2019;8:150. <https://doi.org/10.3390/PLANTS8060150>.
- [8] Deng YW, Luo YY, He Y, Qin XS, Li CG, Deng XM. Complete chloroplast genome of *Michelia Shiluensis* and a comparative analysis with four magnoliaceae species. *Forests*. 2020;11:267. <https://doi.org/10.3390/F11030267>.
- [9] Yang Z, Zhao TT, Ma QH, Liang L, Wang GX. Comparative genomics and phylogenetic analysis revealed the chloroplast genome variation and interspecific relationships of *Corylus* (Betulaceae) species. *Frontiers in Plant Science*. 2018;9: 1–13. <https://doi.org/10.3389/FPLS.2018.00927>.
- [10] Kirsten K. Piecing together the puzzle of parasitic plant plastome evolution. *Planta*. 2011;234:647–656. <https://doi.org/10.1007/S00425-011-1494-9>.
- [11] Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology*. 2016;17:1–29. <https://doi.org/10.1186/S13059-016-1004-2>.
- [12] Vu HT, Tran N, Nguyen TD, Vu QL, Bui MH, Le MT, et al. Complete chloroplast genome of *Paphiopedilum delenatii* and phylogenetic relationships among Orchidaceae. *Plants*. 2020;9:61. <https://doi.org/10.3390/PLANTS9010061>.
- [13] Guo S, Guo LL, Zhao W, Xu J, Li YY, Zhang XY, et al. Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. *Molecules*. 2018;23:1–14. <https://doi.org/10.3390/MOLECULES23020246>.

- [14] Niu ZT, Xue QY, Wang H, Xie XZ, Zhu SY, Liu W, et al. Mutational biases and gc-biased gene conversion affect gc content in the plastomes of *Dendrobium* genus. *International Journal of Molecular Sciences*. 2017;18:2307. <https://doi.org/10.3390/IJMS18112307>.
- [15] Tian N, Han L, Chen C, Wang Z. The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. *PloS One*. 2018;13:e0192956. <https://doi.org/10.1371/JOURNAL.PONE.0192956>.
- [16] Fijridiyanto IA, Murakami N. Phylogeny of *Litsea* and related genera (Laureae-Lauraceae) based on analysis of *rpb2* gene sequences. *Journal of Plant Research*. 2009;122:283–98. <https://doi.org/10.1007/S10265-009-0218-8>.
- [17] Li J, Conran JG, Christophel DC, Li ZM, Li L, Li HW. Phylogenetic relationships of the *Litsea* complex and core *laureae* (Lauraceae) using its and ets sequences and morphology. *Annals of the Missouri Botanical Garden*. 2008;95:580–99. <https://doi.org/10.3417/2006125.9504>.
- [18] Zhang YY, Tian YJ, Tng DYP, Li PF, Wang ZS, Zhou JB, et al. Comparative chloroplast genomics of *Litsea* Lam. (Lauraceae) and its phylogenetic implications. *Forests*. 2021;12:744. <https://doi.org/10.3390/F12060744>.
- [19] Sasaki C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, et al. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Molecular Biology*. 2005;59:309–322. <https://doi.org/10.1007/S11103-005-8882-0>.
- [20] Cui YX, Nie LP, Sun W, Xu ZC, Wang Y, Yu J, et al. Comparative and phylogenetic analyses of ginger (*Zingiber officinale*) in the family Zingiberaceae based on the complete chloroplast genome. *Plants*. 2019;8:283. <https://doi.org/10.3390/PLANTS8080283>.
- [21] Mcinerney JO. Gcua: general codon usage analysis. *Bioinformatics*. 1998;14:372–373. <https://doi.org/10.1093/BIOINFORMATICS/14.4.372>.
- [22] Zuo LH, Shang AQ, Zhang S, Yu XY, Ren YC, Yang MS, et al. The first complete chloroplast genome sequences of *Ulmus* species by *de novo* sequencing: genome comparative and taxonomic position analysis. *PloS One*. 2017;12:e0171264. <https://doi.org/10.1371/JOURNAL.PONE.0171264>.
- [23] Suo ZL, Li WY, Jin XB, Zhang HJ. A new nuclear DNA marker revealing both microsatellite variations and single nucleotide polymorphic loci: a case study on classification of cultivars in *Lagerstroemia indica*. *Article in Journal of Microbial and Biochemical Technology*. 2016;8:266–271. <https://doi.org/10.4172/1948-5948.1000296>.
- [24] Zhang Y, Du LW, Liu A, Chen JJ, Wu L, Hu WM, et al. The complete chloroplast genome sequences of five *Epimedium* Species: lights into phylogenetic and taxonomic analyses. *Frontiers in Plant Science*. 2016;7:306. <https://doi.org/10.3389/FPLS.2016.00306>.

- [25] Dong WL, Wang RN, Zhang NY, Fan WB, Fang MF, Liu ZH. Molecular evolution of chloroplast genomes of *Orchid* Species: insights into phylogenetic relationship and adaptive evolution. *International Journal of Molecular Sciences*. 2018;19:3. <https://doi.org/10.3390/IJMS19030716>.
- [26] Chen JH, Hao ZD, Xu HB, Yang LM, Liu GX, Sheng Y, et al. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. *Frontiers in Plant Science*. 2015;6:447. <https://doi.org/10.3389/FPLS.2015.00447>.
- [27] Park I, Yang SY, Choi G, Kim WJ, Moon BC. The complete chloroplast genome sequences of *Aconitum pseudolaeve* and *Aconitum longecassidatum*, and development of molecular markers for distinguishing species in the *Aconitum* Subgenus *Lycoctonum*. *Molecules*. 2017;22:2012. <https://doi.org/10.3390/molecules22112012>.
- [28] Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology*. 2008;8:36. <https://doi.org/10.1186/1471-2148-8-36>.
- [29] Huo YM, Gao LM, Liu BJ, Yang YY, Kong SP, Sun YQ, et al. Complete chloroplast genome sequences of four *Allium* species: comparative and phylogenetic analyses. *Scientific Reports*. 2019;9:1–14. <https://doi.org/10.1038/s41598-019-48708-x>.
- [30] Kim SC, Lee JW, Choi BK. Seven complete chloroplast genomes from *Symplocos*: genome organization and comparative analysis. *Forests*. 2021;12:608. <https://doi.org/10.3390/f12050608>.
- [31] Li B, Zheng YQ. Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. *Scientific Reports*. 2018;8:1–11. <https://doi.org/10.1038/s41598-018-27453-7>.
- [32] Zong D, Gan PH, Zhou AP, Zhang Y, Zou X, L Duan AA, et al. Plastome sequences help to resolve deep-level relationships of *Populus* in the family Salicaceae. *Frontiers in Plant Science*. 2019;10:5. <https://doi.org/10.3389/fpls.2019.00005>.
- [33] Muraguri S, Xu W, Chapman M, Muchugi A, Oluwaniyi A, Oyebanji O, et al. Intraspecific variation within castor bean (*Ricinus communis* L.) based on chloroplast genomes. *Industrial Crops and Products*. 2020;155:112779. <https://doi.org/10.1016/j.indcrop.2020.112779>.
- [34] Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology*. 2014;14:1–17. <https://doi.org/10.1186/1471-2148-14-151>.
- [35] Meng KK, Chen SF, Xu KW, Zhou RC, Li MW, Dhamala MK, et al. Phylogenomic analyses based on genome-skimming data reveal cyto-nuclear discordance in the evolutionary history of *Cotoneaster* (Rosaceae). *Molecular Phylogenetics and Evolution*. 2021;158:107083. <https://doi.org/10.1016/J.YMPEV.2021.107083>.

- [36] Li P, Lu RS, Xu WQ, Ohi-Toma T, Cai MQ, Qiu YX, et al. Comparative genomics and phylogenomics of East Asian Tulips (*Amansia*, Liliaceae). *Frontiers in Plant Science* 2017;8:451. <https://doi.org/10.3389/FPLS.2017.00451>.
- [37] Wang X, Zhou T, Bai G, Zhao Y. Complete chloroplast genome sequence of *Fagopyrum dibotrys*: genome features, comparative analysis and phylogenetic relationships. *Scientific Reports*. 2018;8:12379. <https://doi.org/10.1038/S41598-018-30398-6>.
- [38] Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology*. 2015;64:1000. <https://doi.org/10.1093/sysbio/syv045>.
- [39] Ma J, Clemants S. A history and overview of the flora reipublicae popularis sinicae (FRPS, Flora of China, Chinese edition, 1959–2004). *Taxon*. 2006;55:451–460. <https://doi.org/10.2307/25065592>.
- [40] Li J, Christophel DC, Conran JG, Li HW. Phylogenetic relationships within the 'core' Laureae (*Litsea* complex, Lauraceae) inferred from sequences of the chloroplast gene *matK* and nuclear ribosomal DNA its regions. *Plant Systematics and Evolution*. 2004;246:19–34. <https://doi.org/10.1007/S00606-003-0113-Z>.
- [41] Chen YC, Li Z, Zhao YX, Gao M, Wang JY, Liu KW, et al. The *Litsea* genome and the evolution of the laurel family. *Nature Communications*. 2020;11:1. <https://doi.org/10.1038/S41467-020-15493-5>.
- [42] Liu XH, Xu XZ, Zhao JX. A new generalized  $p$ -value approach for testing equality of coefficients of variation in  $k$  normal populations. *Journal of Statistical Computation and Simulation*. 2011;81:1121–1130. <https://doi.org/10.1080/00949651003724790>.
- [43] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- [44] Andrews S. FastQC: A quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed September 12, 2021).
- [45] Ewels P, Magnusson M, Lundin S, Källner M. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>.
- [46] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*. 1990;215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- [47] Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan JY, et al. Soapdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*. 2012;1:18. <https://doi.org/10.1186/2047-217X-1-18>.

- [48] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–1649. <https://doi.org/10.1093/BIOINFORMATICS/BTS199>.
- [49] Dierckxsens N, Mardulyn P, Smits G. Novoplasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*. 2017;45:e18. <https://doi.org/10.1093/nar/gkw955>.
- [50] Tillich M, Lehwarck P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. Geseq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*. 2017;45:W6–W11. <https://doi.org/10.1093/nar/gkx391>.
- [51] Lowe TM, Eddy SR. Trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Research*. 1997;25:955–964. <https://doi.org/10.1093/nar/25.5.955>.
- [52] Lagesen K, Hallin P, Rodland EA, Starfeldt HH, Rognes T, Ussery DW. Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Research*. 2007;35:3100–3108. <https://doi.org/10.1093/nar/gkm160>.
- [53] Lehwarck P, Greiner S. Gb2sequin - a file converter preparing custom genbank files for database submission. *Genomics*. 2019;111:759–761. <https://doi.org/10.1016/j.ygeno.2018.05.003>.
- [54] Lohse M, Drechsel O, Bock R. Organellargenomedraw (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics*. 2007;52:267–274. <https://doi.org/10.1007/s00294-007-0161-y>.
- [55] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. Mega x: molecular evolutionary genetics analysis across computing *Molecular Biology and Evolution*. 2018;35:1547–1549. <https://doi.org/10.1093/molbev/msy096>.
- [56] Beier S, Thiel T, Münch T, Scholz U, Mascher M. Misa-web: a web server for microsatellite prediction. *Bioinformatics*. 2017;33:2583–2585. <https://doi.org/10.1093/bioinformatics/btx198>.
- [57] Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*. 2001;29:4633–4642. <https://doi.org/10.1093/nar/29.22.4633>.
- [58] Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. Vista: computational tools for comparative genomics. *Nucleic Acids Research*. 2004;32:W273–W279. <https://doi.org/10.1093/nar/gkh458>.
- [59] Amiryousefi A, Hyvönen J, Poczai P. Irscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*. 2018;34:3030–3031. <https://doi.org/10.1093/bioinformatics/bty220>.

[60] Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. Dnasp 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*. 2017;34:3299–3302. <https://doi.org/10.1093/molbev/msx248>.

[61] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 2013;30:772–780. <https://doi.org/10.1093/molbev/mst010>.

## Tables

Table 1. Gene content of the *L. moupinensis* chloroplast genome.

Group of Genes	Gene Names	Amount
Photosystem I	<i>psaA, psab, psac, psal, psaj</i>	5
Photosystem II	<i>psbA, psbK, psbl, psbM, psbD, psbC, psbZ, psbG, psbL, psbF, psbE, psbB, psbT, psbN, psbH</i>	15
Cytochrome	<i>petA, petG, petL, petN, petB, petD</i>	6
ATP synthase	<i>atpA, atpF, atpH, atpI, atpE, atpB</i>	6
NADH dehydrogenase	<i>ndhJ, ndhB*, ndhK, ndhC, ndhD, ndhF, ndhE, ndhG, ndhI, ndhA, ndhH</i>	12
RubisCO large subunit	<i>rbcL</i>	1
RNA polymerase	<i>RpoCl, rpoC2, rpoB, rpoA</i>	4
Ribosomal proteins (SSU)	<i>rps16, rps12*, rps2, rps14, rps4, rps18, rps7*, rps11, rps8, rps3, rps19, rps15</i>	14
Ribosomal proteins (LSU)	<i>rpB3, rpL20, rpB6, rpB4, rpB6, rpL22, rpL2, rpL23, rpB2</i>	9
Transfer RNAs	<i>trnH-GUG, trnK-UUU, trnQ-UUG, trnS-GCU, trnG-UCC, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-UCC, trnI-GGU, trnS-UGA, trnG-UCQ, trnM-CAU, trnS-GGA, trnI-UGU, trnL-UAA, trnF-GUU, trnV-UAC, trnM-CAU, trnW-CCA, trnP-UGG, trnI-CAU, trnA-UGC, trnR-ACG, trnL-UAG, trnN-GUU, trnR-GUG, trnA-UGC, trnI-GAU, trnL-CAA</i>	34
Ribosomal RNAs	<i>rna4.5*, rna5*, rna16*, rna23*</i>	8
clpPjnak	<i>matK</i>	1
Hypothetical chloroplast reading frames (ycf)	<i>ycf1, ycf2, ycf3, ycf4</i>	4
Other genes	<i>accD, clpP, ccsA, cemA, infA, rpoA</i>	6

\* Gene with two copies.

Table 2. Chloroplast genome features of nine species of *Litsea*.



<i>L. auriculata</i>	<i>L. chunii</i>	<i>L. ichangensis</i>	<i>L. moupinensis</i>	<i>L. populifolia</i>	<i>L. rubescens</i>	<i>L. sericea</i>	<i>L. tsinlingensis</i>	<i>L. veitchiana</i>
152,377	152,081	152,747	152,588	152,619	152,581	152,717	152,051	152,578
93,535	93,138	93,631	93,552	93,569	93,550	93,583	93,093	93,540
18,814	18,813	18,902	18,824	18,838	18,819	18,900	18,828	18,826
20,014	20,065	20,107	20,106	20,106	20,106	20,117	20,065	20,106
39.2%	39.2%	39.2%	39.2%	39.2%	39.2%	39.1%	39.2%	39.2%
37.9%	37.9%	38.0%	37.9%	38.0%	37.9%	37.9%	37.9%	38.0%
33.9%	33.9%	33.9%	33.9%	33.9%	33.9%	33.9%	34.0%	34.0%
44.4%	44.4%	44.4%	44.4%	44.4%	44.4%	44.4%	44.4%	44.4%
128 [113]	128 [113]	128 [113]	128 [113]	128 [113]	128 [113]	128 [113]	128 [113]	128 [113]
84 [79]	84 [79]	84 [79]	84 [79]	84 [79]	84 [79]	84 [79]	84 [79]	84 [79]
9 [4]	8 [4]	8 [4]	8 [4]	8 [4]	8 [4]	8 [4]	8 [4]	8 [4]
36 [30]	36 [30]	36 [30]	36 [30]	36 [30]	36 [30]	36 [30]	36 [30]	36 [30]
MW802253	MW802254	MW802255	MW802256	MW802257	MW802258	MW802259	MW802260	MW802261
49.12%	49.35%	49.06%	49.12%	49.11%	49.12%	49.15%	49.37%	49.12%
8.51%	7.54%	8.81%	8.82%	8.82%	8.95%	8.81%	8.85%	8.82%
30.19%	29.92%	30.93%	30.83%	30.86%	1.36%	30.83%	29.90%	30.85%
13	13	13	13	13	13	13	13	13

Table 3. Comparative analyses of the number and average length of InDel sites in LSC, SSC, and IR regions in the complete chloroplast genomes of nine species of *Litsea*.

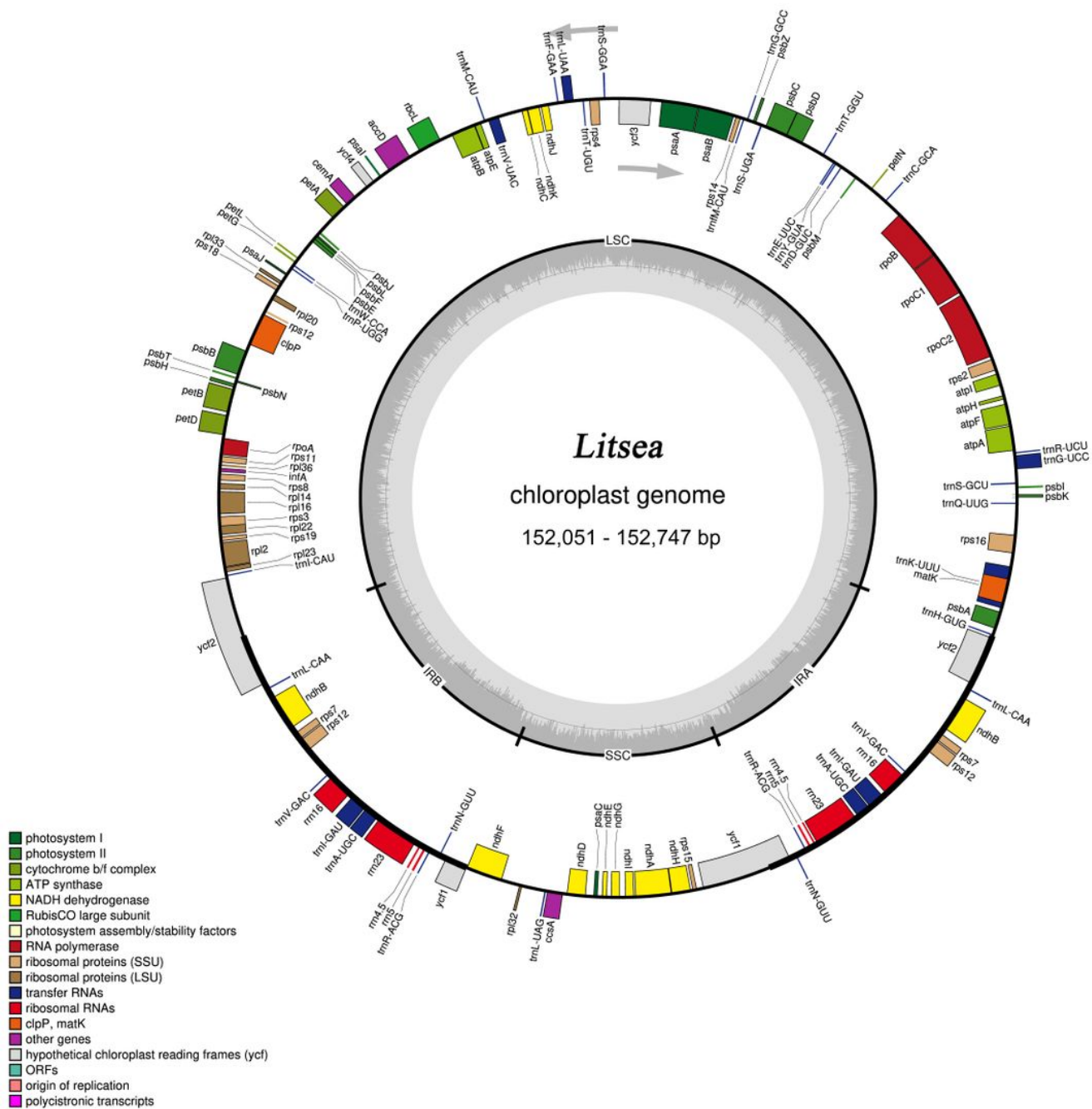
Comparative analyses of InDel sites						
Species	Large single copy		Inverted repeat		Small single copy	
	No's of InDels	InDels' average length	No's of InDels	InDels' average length	No's of InDels	InDels' average length
<i>L. auriculata</i>	86	4.4 (bp)	16	103.3 (bp)	18	1.3 (bp)
<i>L. chunii</i>	89	8.4 (bp)	3	458.7 (bp)	20	1.4 (bp)
<i>L. ichangensis</i>	99	3.7 (bp)	5	276.0 (bp)	19	1.6 (bp)
<i>L. moupinensis</i>	88	3.7 (bp)	4	339.0 (bp)	16	1.9 (bp)
<i>L. populifolia</i>	86	3.8 (bp)	4	339.0 (bp)	17	1.5 (bp)
<i>L. rubescens</i>	88	3.9 (bp)	5	272.4 (bp)	16	1.9 (bp)
<i>L. sericea</i>	86	3.7 (bp)	4	339.0 (bp)	19	2.0 (bp)
<i>L. tsinlingensis</i>	87	9.1 (bp)	2	678.0 (bp)	19	1.4 (bp)
<i>L. veitchiana</i>	88	3.7 (bp)	4	339.0 (bp)	18	1.8 (bp)
<b>Average</b>	88.6	4.9 (bp)	5.2	349.4 (bp)	18.0	1.6 (bp)

Table 4. The number of SNP types in LSC, IR and SSC regions of nine *Litsea* chloroplast genomes.

Comparative analyses of different types of substitutions in *litsea* species

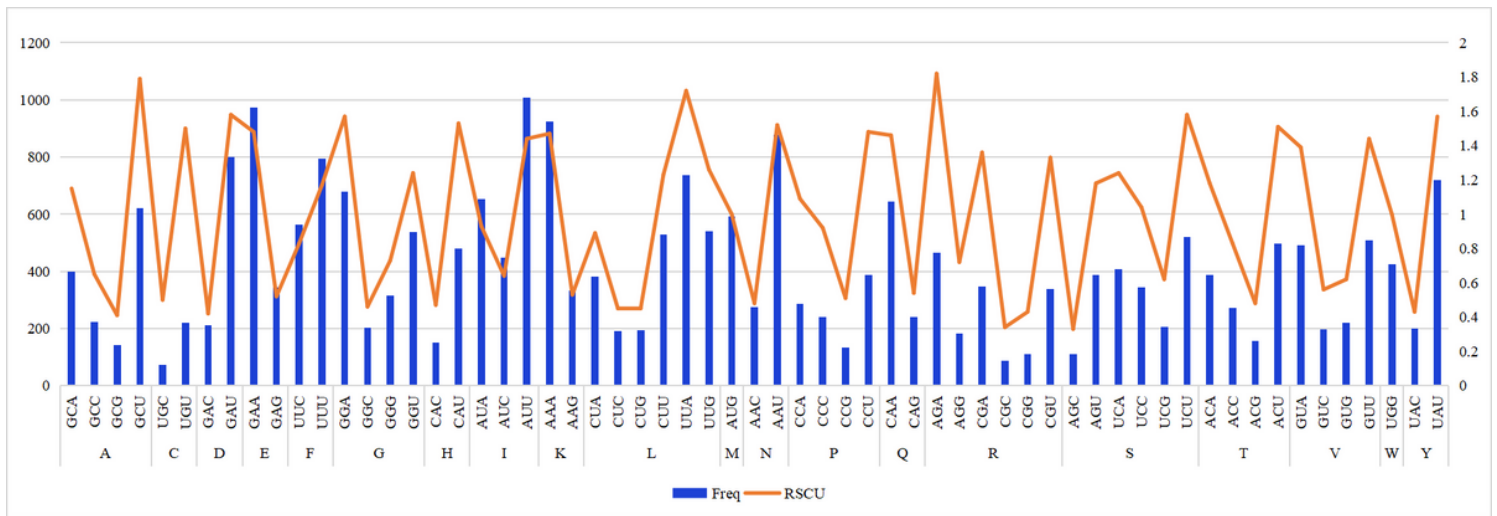
Species	Region	Transition substitutions		Transversion substitutions				
		A/G	C/T	A/T	A/C	C/G	G/T	
<i>L. auriculata</i>	Large	109	106	25	46	6	63	
<i>L. chinü</i>		137	129	20	52	11	80	
<i>L. ichangensis</i>		129	139	26	58	11	75	
<i>L. moupinensis</i>		single	134	139	21	58	10	78
<i>L. populifolia</i>		copy	129	138	22	60	10	80
<i>L. rubescens</i>			134	140	22	58	10	78
<i>L. sericea</i>			123	129	23	55	10	78
<i>L. tsinlingensis</i>			136	127	19	56	11	73
<i>L. veitchiana</i>			127	128	21	59	10	75
<i>L. auriculata</i>	Inverted repeat	3	5	0	2	2	11	
<i>L. chinü</i>		4	8	2	12	1	15	
<i>L. ichangensis</i>		4	5	1	3	1	3	
<i>L. moupinensis</i>		3	8	2	12	1	12	
<i>L. populifolia</i>		3	6	2	12	1	11	
<i>L. rubescens</i>		3	8	2	12	1	11	
<i>L. sericea</i>		3	6	3	12	1	12	
<i>L. tsinlingensis</i>		2	9	2	12	1	14	
<i>L. veitchiana</i>		3	7	2	12	1	12	
<i>L. auriculata</i>	Small	43	47	5	21	3	16	
<i>L. chinü</i>		42	37	10	19	4	17	
<i>L. ichangensis</i>		42	45	5	24	5	21	
<i>L. moupinensis</i>		38	41	5	24	6	21	
<i>L. populifolia</i>		single	37	43	4	19	5	21
<i>L. rubescens</i>		38	41	5	10	6	21	
<i>L. sericea</i>		copy	38	39	6	19	5	24
<i>L. tsinlingensis</i>		42	37	5	20	4	18	
<i>L. veitchiana</i>		37	43	4	18	5	22	

## Figures



**Figure 1**

Complete genome map of the chloroplast genome for representative species *L. moupinensis*. The inner grey ring is divided into four areas, Clockwise, and they are: SSC, IRb, LSC, and IRa. The genes in the outer ring region are transcribed clockwise, while those in the inner ring are transcribed counterclockwise. In addition, this figure also reflects the GC content, the inner ring dark gray indicates the GC content, the light gray reaction AT content. In the lower left is a legend that classifies chloroplast genes according to their functions.



**Figure 2**

The frequency (Freq) and preference of codon usage (RSCU) in *L. moupinensis* protein-coding region. Abscissae indicate each amino acid and its abbreviation as well as the respective codon, the blue bar in the figure is the frequency of codon usage, and the orange line represents codon preference.



**Figure 3**

Comparison of microsatellites and oligonucleotide repeats in the chloroplast genomes of *Litsea* species. (A) the number of SSR markers in the LSC/SSC/IR region for nine *Litsea* species. (B) number of four long repeat sequences in nine species: complement repeats. F represents forward repeats, P represents palindromic repeats, R represents reverse repeats, C represents complement repeats; (C) number of long repeat sequences with different lengths in nine species. Different colors in the figure represent different long repeat sequence types. Species from left to right are *L. auriculata*, *L. chunii*, *L. ichangensis*, *L. moupinensis*, *L. populifolia*, *L. rubescens*, *L. sericea*, *L. tsinlingensis*, *L. veitchiana*.

### Inverted Repeats

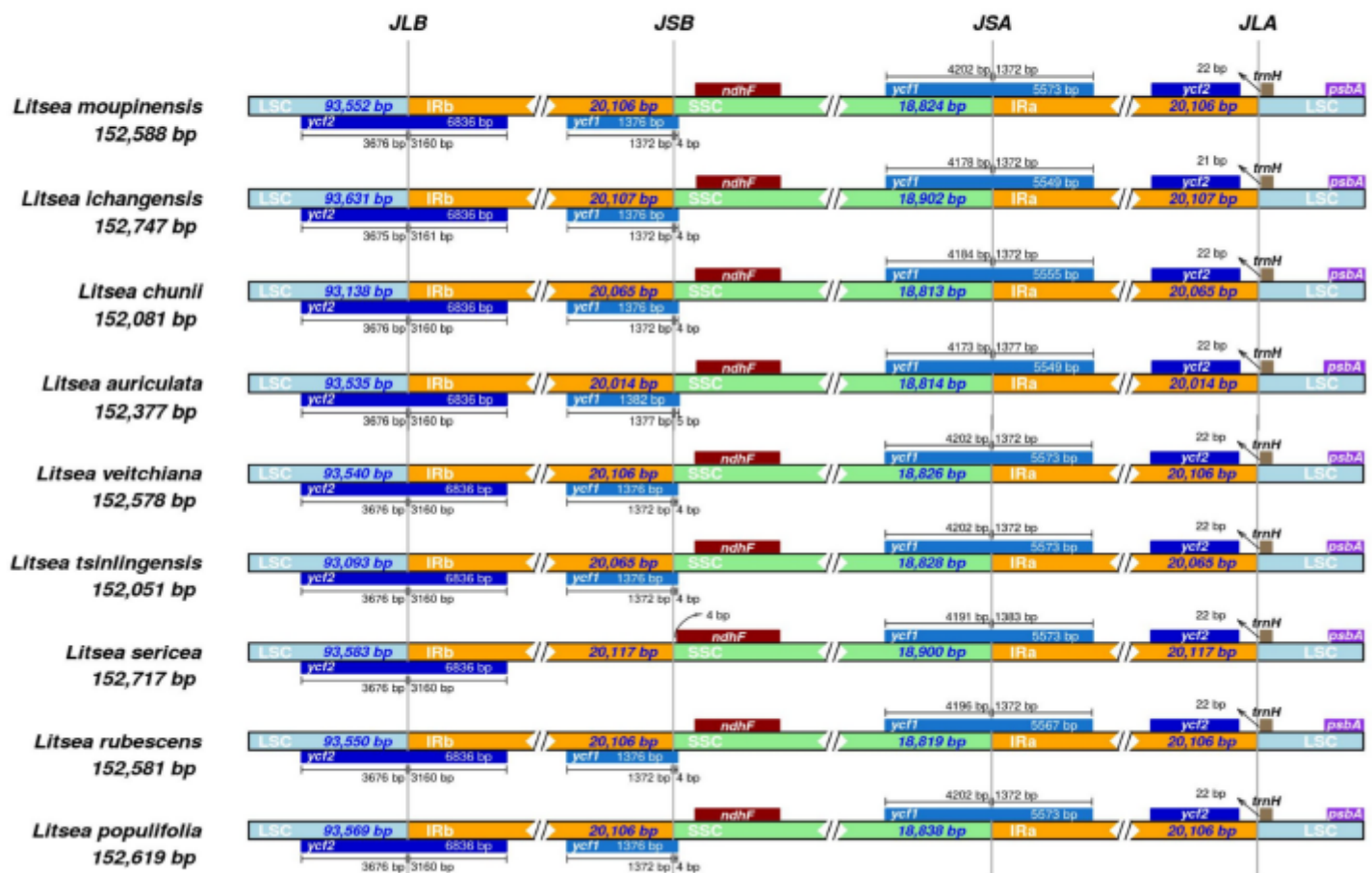


Figure 4

Comparison of SSC, LSC, IRa, and IRb boundary regions in the chloroplast genomes of nine species of *Litsea*.



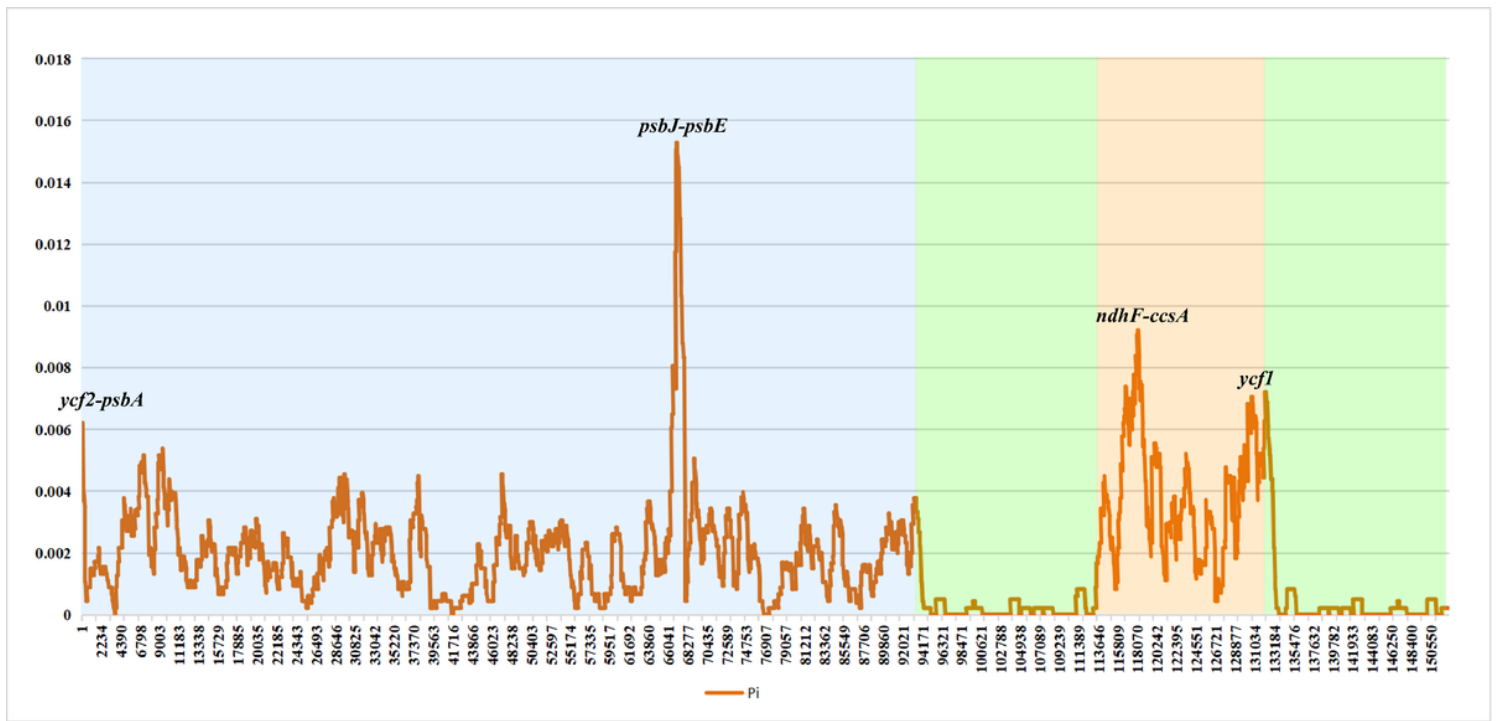


Figure 6

Nucleotide diversity in chloroplast genomes of nine species of *Litsea*. The abscissa represents the position, and the red line represents the average of the nucleotide variations of the nine species (Pi).

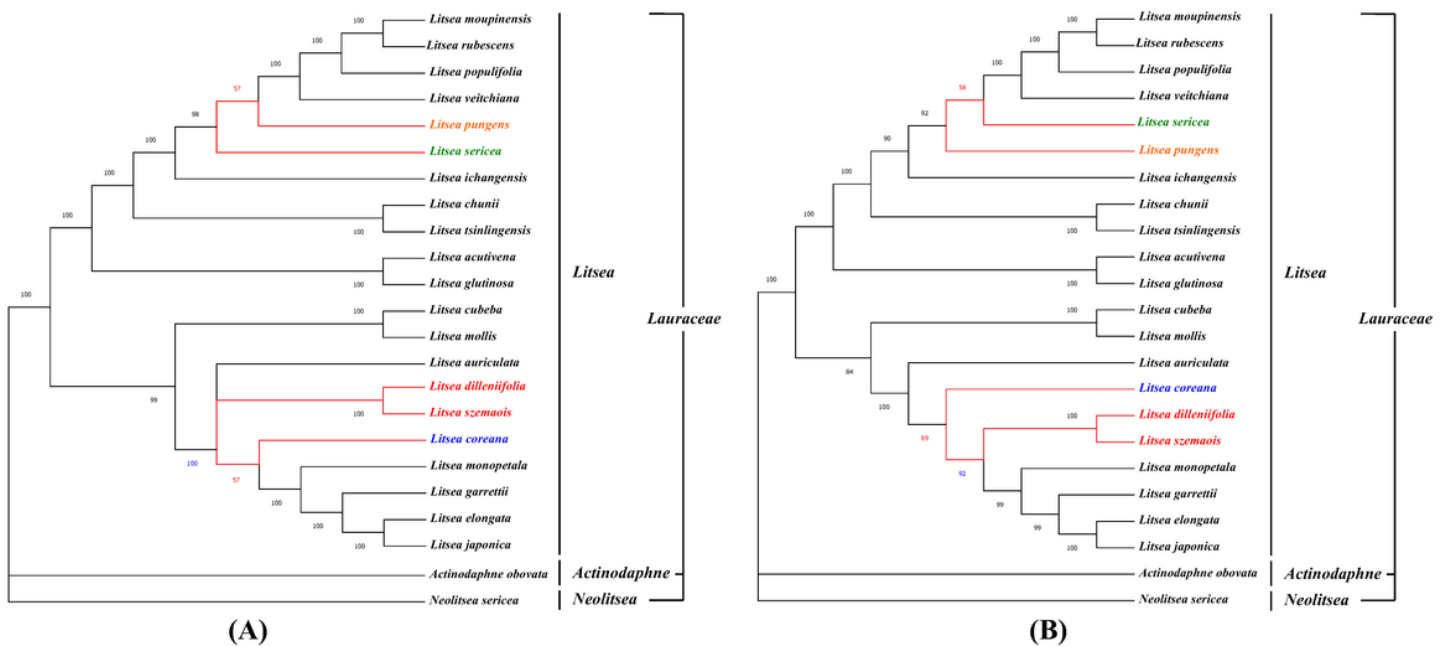


Figure 7

Phylogenetic analysis. (A) phylogenetic tree based on the complete chloroplast genome. (B) phylogenetic tree based on 64 sets of protein-coding genes. *Neolitsea sericea* and *Actinodaphne obovata* served as out groups. Numbers at branch nodes are bootstrap values.

## Supplementary Files

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

- [Table1genename.xlsx](#)
- [Table2genomefeatures.xlsx](#)
- [Table4SNP.xlsx](#)
- [Table3InDels.xlsx](#)
- [TableS1RSCU.xlsx](#)
- [TableS2SSR.xlsx](#)
- [TableS3Repeats.xlsx](#)
- [TableS4AssensionNumbers.xlsx](#)