

# Complete Chloroplast Genome of *Castanopsis sclerophylla* (Lindl.) Schott: Genome Structures, Comparative and Phylogenetic Analysis

Xuemin Ye, Dongnan Hu, Yangping Guo, and Rongxi Sun\*

Jiangxi Provincial Key Laboratory of Silviculture, College of Forestry, Jiangxi Agricultural University, Nanchang 330045, China.

\*rongxisun@163.com

Received: date; Accepted: date; Published: date

## Abstract

*Castanopsis sclerophylla* (Lindl.) Schott is an important species of evergreen broad-leaved forest in subtropical area and has important ecological and economic value. However, there are little studies on its chloroplast genome. In this study, the complete chloroplast genome sequences of *C. sclerophylla* was reported based on the Illumina Hiseq 2500 platform. The complete chloroplast genome of *C. sclerophylla* was 160,497bp, including a pair of inverted repeated (IRs) regions (25,675bp) that were separated by a large single copy (LSC) region of 90,255bp, and a small single copy (SSC) region of 18,892bp. The overall GC content of chloroplast genome was 36.82%. A total of 131 genes were found, of these 111 genes were unique and annotated, including 79 protein-coding genes, 27 transfer RNA genes (tRNAs), and four ribosomal RNA genes (rRNAs). Twenty-one genes were found to be duplicated in the IR regions. Comparative analysis indicated that IR contraction might be the reason for the relatively smaller chloroplast genome size of *C. sclerophylla* compared with other three congeneric species. Sequence analysis detected that the LSC and SSC regions were more divergent than the IR regions within the *Castanopsis*, furthermore, a higher divergence was found in non-coding regions than in coding regions. The maximum likelihood (ML) phylogenetic analysis showed that these four species of the genus *Castanopsis* formed a monophyletic clade and that *C. sclerophylla* is closely related to *Castanopsis hainanensis* with strong bootstrap values. These results not only provide basic knowledge about characteristics of *C. sclerophylla* and also enhance our understanding of *Castanopsis* species evolution within the Fagaceae family. Meanwhile, these findings will contribute to the exploration, utilization and conservation genetics of *C. sclerophylla*.

**Keywords:** *Castanopsis sclerophylla*; chloroplast genome; comparative analysis; phylogenetic analysis

## Introduction

*Castanopsis sclerophylla* (Lindl.) Schott. is a monoecious, and broad-leaved tree of the genus *Castanopsis* belongs to the Fagaceae family. The genus contains about 120 known species, of these species 58 are native to China, and 30 are endemic. However, this species is widely native to East and South Asia, and had been introduced to North America[1, 2]. In China, it is a canopy tree widely distributed in subtropical evergreen forests[3]. Its fruit and wood are valuable, and are regarded as landscape and ornamental tree because of glossy evergreen leaves and white flowers cover the tree [4]. In Jiangxi province, its fruit has been used to make special foods such as sheet jelly, bean curd, and bean vermicelli[5]. The previous researches mainly focused on natural regeneration[6], biomass[7], morphology[8], chemotaxonomical[9], genetic diversity[1, 10]. Currently, because of the tree's economic value increasing, natural plants have been seriously destroyed by humans, and the number

of natural trees is decreasing due to the slowly grow rate, this species is severely fragmented and constantly threatened, and need urgent conservation and restoration [11]. Phylogenetic evolution and population genomics knowledge are very vital to make effective conservation and management strategies. With the rapid development of next-generation sequencing technology, compared with the Sanger method, the chloroplast genome assembly has become cheaper and easier based on Illumina sequencing technology, comparative analysis of the complete chloroplast genome among closely related species has been proven to be a valid and most effective method for the study of evolutionary history, species conservation, and phylogenetic relationships [12-14].

Chloroplasts are essential organelles in plant cells which play very important roles for photosynthesis, carbon fixation, and pigments, starch, fatty acids, and amino acids synthesizing [15, 16]. Typically, the chloroplast genome of angiosperms is a highly conserved circular DNA ranging from 120 to 180 kb in higher plants, and has a typical quadripartite structure including a large single copy (LSC) region, a small single copy (SSC) region, and a pair of inverted repeats (IRs) [17]. The chloroplast genome encodes about 110 to 130 genes, including up to 80 unique protein-coding genes, four ribosomal RNAs (rRNAs), and approximately 30 transfer RNAs (tRNAs) [18]. In recent years, many complete chloroplast genome sequences of higher plants have been reported, which were used to study population structure and phylogenetic relationships.

In this study, we report the *C. sclerophylla* chloroplast genome using Illumina sequencing technology. This is the first comprehensive analysis of *C. sclerophylla* chloroplast genomes, in conjunction these whole chloroplast genome sequences of three congeneric species which were previously published. In addition, we also used 22 complete chloroplast genome sequences from GenBank to analyzed phylogenic relationships and infer the phylogenetic position of *C. sclerophylla*. These results not only provide basic knowledge about characteristics of *C. sclerophylla* and also enhance our understanding of *Castanopsis* species evolution within the Fagaceae family. Our data will contribute to understand the genetic resources and evolutionary based on the diversity in the chloroplast genome, and facilitate to the exploration, utilization and conservation genetics of *C. sclerophylla*.

## Materials and Methods

### Plant material, DNA Extraction and Sequencing

Fresh young leaves of *C. sclerophylla* were collected from the Jiangxi agricultural university arboretum in Nanchang, China (28°45'N, 115°49'E). Total genomic DNA was extracted using the Plant Genomic DNA Kit (TIANGEN, Beijing, China). Agarose gel electrophoresis and Microplate Spectrophotometer (Molecular Device, Sunnyvale, CA, USA) were used to measured DNA quality and concentration, respectively. Shotgun libraries with an average insert size of 350 bp were constructed using pure DNA and sequenced with an Illumina Hiseq 2500 platform. Approximately 10.44 GB of clean data from *C. sclerophylla* were yielded with 150 bp paired-end read lengths.

### Chloroplast Genome Assemblage and Annotation

The high throughput raw reads were trimmed by Fastqc. Next, trimmed paired-end reads and references (*C. hainanensis*, *C. echinocarpa*, and *C. concinna*) were used to extract chloroplast like reads. And then, these reads were assembled by NOVOPlasty [19]. NOVOPlasty assembled part reads and stretched as far as possible until a circular genome formed. Finally, high quality complete chloroplast genome was obtained. The assembled genome was annotated using CpGAVAS [20]. BLAST and Dual Organellar Genome Annotator (DOGMA) were used to check the annotation results [21]. These tRNAs were identified by the tRNAscanSE [22]. The circular gene maps of *C. sclerophylla* was drawn by the OGDRAW v1.2 program [23]. To analyze the variation of synonymous codon usage, MEGA7 was used to compute relative synonymous codon usage values (RSCU), codon usage, and the GC content [24].

### Comparative Analysis and Phylogenetic Analysis

MUMmer [25] was conducted for pairing sequence alignment of the chloroplast genomes. The mVISTA [26] program was used to compare the complete chloroplast genome of *C. sclerophylla* to

other three published chloroplast genomes of the genus *Castanopsis*, i.e., *C. concinna* (NC\_033409), *C. echinocarpa* (NC\_023801), *C. hainanensis* (NC\_037389) with the shuffle-LAGAN mode, adopting the annotation of *C. concinna* as a reference.

Phylogenies were constructed using the 19 cp genome of the Fagaceae species sequences from the NCBI Organelle Genome and Nucleotide Resources database: *Castanea henryi* (NC\_033881), *Castanea mollissima* (NC\_014674), *Castanopsis concinna* (NC\_033409), *Castanopsis echinocarpa* (NC\_023801), *Castanopsis hainanensis* (NC\_037389), *Fagus engleriana* (NC\_036929), *Lithocarpus balansae* (NC\_026577), *Quercus aliena* (NC\_026790), *Quercus aquifolioides* (NC\_026913), *Quercus baronii* (NC\_029490), *Quercus dolicholepis* (NC\_031357), *Quercus glauca* (NC\_036930), *Quercus rubra* (NC\_020152), *Quercus sichouensis* (NC\_036941), *Quercus spinosa* (NC\_026907), *Quercus tarokoensis* (NC\_036370), *Quercus tungmaiensis* (NC\_036936), *Quercus variabilis* (NC\_031356), and *Trigonobalanus doichangensis* (NC\_023959). The sequences were initially aligned using MAFFT[27]. Then, the visualization and manual adjustment of multiple sequence alignment was conducted in BioEdit[28]. The maximum likelihood (ML) analysis was conducted by RAxML web servers[29]. The bootstrap test was conducted 1000 times adopting Kimura 2-parameter model. *Corylus fargesii* (NC\_031854), and *Eucalyptus umbra* (NC\_022387) were used as the outgroups.

## Results and discussion

### Characteristics of *C. sclerophylla* cpDNA

A total of 65 million pair-end reads were obtained, with 10.44 Gb clean data. The complete chloroplast genome sequence of *C. sclerophylla* is 160,497 bp in length, which deposited in GenBank with accession number MK387847. Its genome possesses a typical quadripartite structure including a pair of inverted repeats (IRa and IRb) regions of 25,675bp that was separated by a large single copy (LSC) region of 90,255bp, and a small single copy (SSC) region of 18,892bp (Table 1 and Fig. 1). The overall GC content of the chloroplast genome was 36.82%, and has a similar GC content to those of other Fagaceae species. Whereas, few differences in the GC contents were found among the chloroplast genome. The contents of the LSC, SSC, and IR regions are 34.65%, 30.94%, and 42.78% respectively (Table 2). The GC content is highest in the IR regions (42.78%), the phenomena was found in the chloroplast genome of *C. hainanensis* [30]. The overall GC content is an important indicator of species affinity[31].

In the *C. sclerophylla* chloroplast genome, a total of 131 genes were found, including 86 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Fig.1, Table 1). All of these 131 genes, 110 genes were unique and annotated, and were divided into three categories including 79 protein-coding genes, 27 tRNA genes, and four rRNA genes (Table 3). In addition, 21 functional genes (seven protein-coding genes, four rRNA, and 10 tRNA genes) are duplicated in the IR regions (Fig.1). The LSC region comprises 62 protein-coding genes and 22 tRNA genes, whereas the SSC region comprises 11 protein-coding genes and one tRNA gene (Table S1). There are 14 intron-containing genes, including eight protein-coding genes and six tRNA genes. 12 genes contain one intron, with *clpP* and *ycf3* genes having two introns. The *trnK-UUU* gene has the longest intron (2511bp), and the *trnL-UAA* gene has the smallest (485bp) (Table 4). The similar phenomenon is also present in *Quercus acutissima*[32]. The *ycf3* gene has been shown to stable accumulation of photosystem I complexes[33]. So, we will focus on the *ycf3* intron gain of *C. sclerophylla* may be helpful for further study the photosynthesis mechanism.

**Fig. 1 Chloroplast genome annotation map of *C. sclerophylla*. Genes inside the circle are transcribed in a clockwise, whereas genes outside are transcribed in a counterclockwise direction. Different colors represent**

different functional Genes. The darker gray and lighter gray in the inner circle shows the GC content and AT content of chloroplast genome, respectively.

According to the protein-coding sequences and tRNA genes, the codon usage frequency was computed for *C. sclerophylla* chloroplast genome (Table 5). All genes are encoded by 23131 codons. Among these, leucine (10.61%) is the maximum commonly encoded amino acid, with 2454 codons, and cysteine (1.13%) is the minimum commonly encoded amino acid with 262 codons. Similar ratios for amino acids are found in previously studied chloroplast genomes [34, 35]. The codons ending with A and U are common. Furthermore, the most preferred synonymous codons which relative synonymous codon usage values (RSCU) were bigger than 1 end with A or U[36].

## Comparative Analysis of Genomic Structure

Three complete chloroplast genomes within the *Castanopsis* genus (*C. hainanensis*, *C. echinocarpa*, and *C. concinna*) were selected for comparison with that of *C. sclerophylla*. *C. sclerophylla* has the smallest chloroplast genome (160,497bp), whereas *C. echinocarpa* has the largest chloroplast genome (160,647) with the smallest IR region (25,629 bp). Additionally, the LSC region lengths varied in these four species, from 90,255 bp of *C. sclerophylla* to 90,394 bp of *C. echinocarpa* (Table 1). The different lengths of LSC region are the main reason for four specie in sequence lengths, this result is coincident with the genus of *Oryza*[34]. To investigate the levels of the genome divergence, sequence identity was plotted for four species chloroplast genomes with *C. concinna* as a reference by the program mVISTA (Fig.2). The results of sequence analysis detected that the LSC and SSC regions were more divergent than the IR regions among the four genomes within the *Castanopsis*, furthermore, a higher divergence was found in non-coding regions than in coding regions. Significant variation in coding regions of the four chloroplast genomes were *ndhF*, *ndhG*, and *ycf1*, which were all located in the SSC regions. However, the most divergent regions were observed in the intergenic regions, including *trnK-rps16*, *trnS-trnT*, *atpA-atpF*, *trnC-petN*, *trnT-psbD*, *Ihba-trnG*, *ycf3-trnS*, *rps4-trnT*, *trnT-trnL*, *atpB-rbcL*, *petA-psbJ*, *psbE-petL*, *rpl16-rpl3*, *ndhF-rpl32*.

The expansion and contraction of IR regions at the borders are the major reason for chloroplast genome size variations and play vital roles in evolution [37-39]. A detailed comparison of four junctions (JLA, JSB, JSA, and JLA) between the two single-copy regions (LSC and SSC) and the two IRs (IRa and IRb) was performed among *C. sclerophylla*, *C. hainanensis*, *C. echinocarpa* and *C. concinna* by analyzing the exact IR border positions and adjacent genes (Fig.3). The IR regions are relatively conserved in the genus of *Castanopsis*, this result is agreed with the reports in the genus of *Quercus*[32]. The *rps19* gene was located between the junction of LSC and IRb regions in *C. concinna*. However, in *C. sclerophylla*, *C. hainanensis*, and *C. echinocarpa* chloroplast genomes, the *rps19* gene was completely located in LSC region and had an 11bp, 11bp, and 10bp distance to the border of LSC, respectively. Some studies indicate that *ycf1* is requisite for plant viability and encodes Tic214 which is an vital component of TIC complex in *Arabidopsis*[40, 41]. The *ycf1* gene crossed the SSC/IRb and SSC/IRa regions. The SSC/IRb junction is located in the *ycf1* region in all four *Castanopsis* species chloroplast genomes and extends into the SSC region by different lengths depending on the genome (*C. sclerophylla*, 21bp; *C. hainanensis*, 24 bp; *C. echinocarpa*, 59 bp; *C. concinna*, 59 bp); the IRb region includes 1131, 1157, 1107, and 1092bp of the *ycf1* gene. The SSC/IRa junction extends into the SSC region by different lengths depending on the genome (*C. sclerophylla*, 4581bp; *C. hainanensis*, 4568 bp; *C. echinocarpa*, 4608 bp; *C. concinna*, 4581bp); the IRa region includes 1092, 1114, 1107, and 1092bp of the *ycf1* gene.

**Fig. 2 Visualization alignment of the complete chloroplast genome of four species with *C. concinna* as a reference by the program mVISTA.** The grey arrows and thick black lines above the alignment indicate the orientation of genes. Blue bars represent exons, sky-blue ones represent untranslated region (UTR), and pink ones represent non-coding sequences (CNS). The vertical scale represents the percent identity within 50–100%.

**Fig. 3 Comparison of junctions the large single copy (LSC), small single copy (SSC), and inverted repeat (IR) among four chloroplast genomes of congeneric species.**

## Phylogenetic Analysis

Phylogenetic analysis was constructed based on the 22 alignment sequences chloroplast genomes by maximum likelihood (ML) analysis (Fig. 4). *Corylus fargesii* and *Eucalyptus umbra* were used as the outgroups. The maximum likelihood (ML) phylogenetic analysis showed that these four species of the genus *Castanopsis* formed a monophyletic clade and that *C. sclerophylla* is closely related to *Castanopsis hainanensis* with strong bootstrap values. The ML tree indicated that *Castanopsis* was closely related to *Castanea*. Surprisingly, the species of *Quercus* genus were not formed a clade, in addition, this genus *Quercus* were not divided to two clusters according to evergreen tree species deciduous tree species. The phylogenetic status of these genus were consistent with the previous report[32]. Until now, little has been known about the chloroplast genome of the *Castanopsis*, and only three chloroplast genome sequences of the *Castanopsis* species are found in GeneBank, which has greatly hampered for studying the phylogenetic relationships of *Castanopsis*. So, it is necessary for more research of the complete chloroplast genome within the *Castanopsis* genus in the future.

**Fig. 4** A maximum likelihood (ML) phylogenetic tree was constructed based on 22 species chloroplast genomes. *Corylus fargesii* and *Eucalyptus umbra* were used as the outgroups.

## Conclusions

*C. sclerophylla* is an important evergreen broad-leaved species of the *Castanopsis* genus in the Fagaceae family. In this study, the complete chloroplast genome sequences of *C. Castanopsis sclerophylla* was reported based on the Illumina Hiseq 2500 platform. *C. sclerophylla* chloroplast genomes exhibited a typical quadripartite and circular structure, which is similar to other three congeneric species. Compared to the chloroplast genomes of three *Castanopsis* species, *C. sclerophylla* has the smallest chloroplast genome (160,497bp). The phylogenetic ML tree showed that Phylogenetic relationships among 22 angiosperms strongly supported the known classification of *C. sclerophylla*. The ML analysis showed that these four species of the genus *Castanopsis* formed a monophyletic clade and that *C. sclerophylla* is closely related to *C. hainanensis* with strong bootstrap values, in addition, *Castanopsis* was closely related to *Castanea*. The genus *Castanopsis* contains about 120 known species, nearly half of these species are native to China. China has plenty of *Castanopsis* germplasm resources, the availability of the chloroplast genomes provides a powerful genetic resource for the phylogenetic analysis and biological study. So, it is necessary for research of the complete chloroplast genome of the genus *Castanopsis* in the future. The data will contribute to understand the genetic resources and identification of evolutionary relationships, and facilitate to the exploration, utilization and conservation genetics of the genus.

## Acknowledgments

We are grateful to Yuelong Liang for his assistance in sample collection.

## References

1. Yi-Su S, Jian Z, Kai J, Min-Yan C, Yuan-Yuan L. Development and characterization of polymorphic microsatellite markers in *Castanopsis sclerophylla* (Fagaceae). American Journal of Botany. 2011; 98(2): 19-21.
2. Conden PJ. Propagation of *Castanopsis sclerophylla* and *Lindera umbellata* by Stem Cuttings and Nitrogen Nutrition of Containerized *Ternstroemia gymnanthera*. 2002.
3. Wang R, Compton SG, Shi YS, Chen XY. Fragmentation reduces regional-scale spatial genetic structure in a wind-pollinated tree because genetic barriers are removed. Ecology and evolution. 2012; 2(9): 2250-2261. <https://doi.org/10.1002/ece3.344> PMID: 23139883.

4. Tu ZC LG, Dou YX, Liu CM, Lin DR. Study on extraction and purification total flavones from the fruits of *Castanopsis sclerophylla*. Food Res Dev. 2008; 29(1): 33-36.
5. Xiao QH, Zong CT, Xue CZ, Hui W, Hui X, Qiu TZ, et al. Isolation and Physicochemical Properties of Hardleaf Oatchestnut (*Castanopsis sclerophylla*) Starch. Applied Mechanics & Materials. 2012; 140(2): 360-368.
6. Xin Z, Xu G, Shen D, Gu Y, Hui G, Luo X, et al. Maintenance and natural regeneration of *Castanopsis sclerophylla* populations on islands of Qiandao Lake Region, China. Acta Ecologica Sinica. 2007; 27(2): 424-430.
7. Tang X, Fehrmann L, Guan F, Forrester DI, Guisasola R, Pérez-Cruzado C, et al. A generalized algebraic difference approach allows an improved estimation of aboveground biomass dynamics of *Cunninghamia lanceolata* and *Castanopsis sclerophylla* forests. Annals of Forest Science. 2017; 74(1): 12.
8. Yue HF, Shao WH, Jing ZH, Lu P, Huang L, Jiang JM. Geographic variation of seed characters of *Castanopsis sclerophylla*. Forest Research. 2010.
9. Huang YL, Tanaka T, Matsuo Y, Kouno I, Li DP, Nonaka GI. Two new phenolic glucosides and an ellagitannin from the leaves of *Castanopsis sclerophylla*. Phytochemistry Letters. 2012; 5(1): 158-161.
10. Lin-Jiang YE, Wang J, Sun P, Dong SP, Zhang ZY. The Transferability of Nuclear Microsatellite Markers in Four *Castanopsis* Species to *Castanopsis tibetana* (Fagaceae). Plant Diversity & Resources. 2014.
11. Rong W, Compton SG, Yi-Su S, Xiao-Yong C. Fragmentation reduces regional-scale spatial genetic structure in a wind-pollinated tree because genetic barriers are removed. Ecology & Evolution. 2012; 2(9): 2250-2261.
12. Moore MJ, Soltis PS, Bell CD, J Gordon B, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(10): 4623-4628.
13. Cécile A, J Gordon B, McMahon MM, Sanderson MJ. Covarion structure in plastid genome evolution: a new statistical test. Molecular Biology & Evolution. 2005; 22(4): 914-924.
14. Raman G, Choi KS, Park S. Phylogenetic Relationships of the Fern *Cyrtomium falcatum* (Dryopteridaceae) from Dokdo Island Based on Chloroplast Genome Sequencing. Genes. 2016; 7(12): 115.
15. Daniell H, Lin CS, Ming Y, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology. 2016; 17(1): 134.
16. And HEN, Emes MJ. NONPHOTOSYNTHETIC METABOLISM IN PLASTIDS. Annu Rev Plant Physiol Plant Mol Biol. 2000; 51(51): 111-140.
17. Wicke S, Schneeweiss GM, Depamphilis CW, Kai FM, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Molecular Biology. 2011; 76(3-5): 273-297.
18. Park I, Kim WJ, Yeo SM, Choi G, Kang YM, Piao R, et al. The Complete Chloroplast Genome Sequences of *Fritillaria ussuriensis* Maxim. and *Fritillaria cirrhosa* D. Don, and Comparative Analysis with Other *Fritillaria* Species. Molecules. 2017; 22(6): 982.

19. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle  
genomes from whole genome data. *Nucleic Acids Research*. 2017; doi:  
10.1093/nar/gkw955(4).
20. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, et al. CpGAVAS, an integrated web server  
for the annotation, visualization, analysis, and GenBank submission of completely sequenced  
chloroplast genome sequences. *Bmc Genomics*. 2012; 13(1): 715-715.
21. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with  
DOGMA. *Bioinformatics*. 2004; 20(17): 3252-3255.
22. Peter S, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers  
for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*. 2005; 33(Web Server  
issue): 686-689.
23. 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(5-6): 267-274.
24. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis  
version 7.0 for bigger datasets. *Molecular Biology & Evolution*. 2016; 33(7): 1870.
25. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile  
and open software for comparing large genomes. *Genome Biology*. 2004; 5(2): R12.
26. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, et al. VISTA  
: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*. 2000;  
16(11): 1046-1047.
27. Kazutaka K, Kei-Ichi K, Hiroyuki T, Takashi M. MAFFT version 5: improvement in  
accuracy of multiple sequence alignment. *Nucleic Acids Research*. 2005; 33(2): 511-518.
28. Hall T A. BioEdit: A user-friendly biological sequence alignment editor and analysis  
program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 1999, 41: 95–98.
29. Alexandros S, Paul H, Jacques R. A rapid bootstrap algorithm for the RAXML Web  
servers. *Systematic Biology*. 2017; 57(5): 758.
30. Chen X. Yang J. Yang Y. Zhang X. Zhao G. Characterization of the complete plastid  
genome of *castanopsis hainanensis merrill*. *Conservation Genetics Resources* 2017;10 (4):1-  
4. <https://doi.org/10.1007/s12686-017-0940-9>.
31. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, et al. Complete Chloroplast Genome  
Sequence and Phylogenetic Analysis of the Medicinal Plant *Artemisia annua*. *Molecules*.  
2017; 22(8): 1330.
32. Xuan L, Li Y, Zang M, Li M, Fang Y. Complete Chloroplast Genome Sequence and  
Phylogenetic Analysis of *Quercus acutissima*. *International Journal of Molecular Sciences*.  
2018; 19(8): 2443.
33. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD. The chloroplast *ycf3* and  
*ycf4* open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation  
of the photosystem I complex. *Embo Journal*. 2014; 16(20): 6095-6104.
34. Asaf S, Waqas M, Khan AL, Khan MA, Kang SM, Imran QM, et al. The Complete  
Chloroplast Genome of Wild Rice (*Oryza minuta*) and Its Comparison to Related Species.  
*Frontiers in plant science*. 2017; 8: 304.

35. Chen J, Hao Z, Xu H, Yang L, Liu G, 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.
36. Shetty SM, Md Shah MU, Makale K, Mohd-Yusuf Y, Khalid N, Othman RY. Complete Chloroplast Genome Sequence of Corroborates Structural Heterogeneity of Inverted Repeats in Wild Progenitors of Cultivated Bananas and Plantains. *Plant Genome*. 2016; 9(2).
37. 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(1): 36.
38. Yao X, Tang P, Li Z, Li D, Liu Y, Huang H. The First Complete Chloroplast Genome Sequences in Actinidiaceae: Genome Structure and Comparative Analysis. *PloS one*. 2015; 10(6): e0129347.
39. Yang Y, Dang Y, Li Q, Lu J, Li X, Wang Y. Complete chloroplast genome sequence of poisonous and medicinal plant *Datura stramonium*: organizations and implications for genetic engineering. *PloS one*. 2014; 9(11): e110656. <https://doi.org/10.1371/journal.pone.0110656> PMID: 25365514.
40. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. ycf1, the most promising plastid DNA barcode of land plants. *Scientific reports*. 2015; 5: 8348.
41. Shingo K, Jocelyn B, Minako H, Yoshino H, Maya O, Midori I, et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science*. 2013; 339(6119): 571-574.

## Figure Legends

**Fig. 1 Chloroplast genome annotation map of *C. sclerophylla*.** Genes inside the circle are transcribed in a clockwise, whereas genes outside are transcribed in a counterclockwise direction. Different colors represent different functional Genes. The darker gray and lighter gray in the inner circle shows the GC content and AT content of chloroplast genome, respectively.

**Fig. 2 Visualization alignment of the complete chloroplast genome of four species with *C. concinna* as a reference by the program mVISTA.** The grey arrows and thick black lines above the alignment indicate the orientation of genes. Blue bars represent exons, sky-blue ones represent untranslated region (UTR), and pink ones represent non-coding sequences (CNS). The vertical scale represents the percent identity within 50–100%.

**Fig. 3 Comparison of junctions the large single copy (LSC), small single copy (SSC), and inverted repeat (IR) among four chloroplast genomes of congeneric species.**

**Fig. 4 A maximum likelihood (ML) phylogenetic tree was constructed based on 22 species chloroplast genomes.** *Corylus fargesii* and *Eucalyptus umbra* were used as the outgroups.

## Supporting information

### S1 Table

**S1 Table.** The number of genes *C. sclerophylla* chloroplast genome chloroplast genome

| Region     | Number of CDS | Number of tRNA | Number of rRNA | Total |
|------------|---------------|----------------|----------------|-------|
| LSC region | 62            | 22             | 0              | 84    |
| SSC region | 11            | 1              | 0              | 12    |
| IRA region | 7             | 7              | 4              | 18    |
| IRB region | 6             | 7              | 4              | 17    |

## Tables

**Table 1. Summary of four *Castanopsis* chloroplast genome Characteristics.**

| Genome                         | <i>C. sclerophylla</i> | <i>C. hainanensis</i> | <i>C. echinocarpa</i> | <i>C. concinna</i> |
|--------------------------------|------------------------|-----------------------|-----------------------|--------------------|
| Genome size (bp)               | 160,497                | 160,631               | 160,647               | 160,606            |
| LSC length (bp)                | 90,255                 | 90,328                | 90,394                | 90,368             |
| SSC length (bp)                | 18,892                 | 18,929                | 18,995                | 18,884             |
| IR length (bp)                 | 25,675                 | 25,687                | 25,629                | 25,677             |
| Number of genes                | 131                    | 132                   | 132                   | 136                |
| Number of protein-coding genes | 86                     | 84                    | 84                    | 82                 |
| Number of tRNA genes           | 37                     | 40                    | 40                    | 46                 |
| Number of rRNA genes           | 8                      | 8                     | 8                     | 8                  |

**Table 2. Base content of the *C. sclerophylla* chloroplast genome.**

| Region | A (%) | T (%) | C (%) | G (%) | A+T (%) | G+C (%) |
|--------|-------|-------|-------|-------|---------|---------|
| LSC    | 31.94 | 33.4  | 17.74 | 16.91 | 65.34   | 34.65   |
| SSC    | 34.4  | 34.66 | 16.29 | 14.65 | 69.06   | 30.94   |
| IR     | 28.61 | 28.61 | 21.39 | 21.39 | 57.22   | 42.78   |
| Total  | 31.65 | 32.23 | 18.47 | 17.65 | 63.18   | 36.82   |

359

**Table 3. List of genes annotated in the sequenced *C. sclerophylla* chloroplast genome chloroplast genome.**

| Category         | Function                                | Genes                                                                                                                                                                                                                                                                                                                                     |
|------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Photosynthesis   | Photosystem I                           | <i>psaA, psaB, psaC, psaI, psaJ</i>                                                                                                                                                                                                                                                                                                       |
|                  | Photosystem II                          | <i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT</i>                                                                                                                                                                                                                                                 |
|                  | Cytochrome b/f complex                  | <i>petA, petB, petD, petG, petL, petN</i>                                                                                                                                                                                                                                                                                                 |
|                  | ATP synthase                            | <i>atpA, atpB, atpE, atpF*, atpH, atpI</i>                                                                                                                                                                                                                                                                                                |
|                  | NADH dehydrogenase                      | <i>ndhA*, ndhB*(X2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>                                                                                                                                                                                                                                                             |
|                  | Rubisco large subunit                   | <i>rbcL</i>                                                                                                                                                                                                                                                                                                                               |
| Self-replication | RNA polymerase                          | <i>rpoA, rpoB, rpoC1*, rpoC2</i>                                                                                                                                                                                                                                                                                                          |
|                  | Ribosomal proteins (LSU)                | <i>rpl14, rpl16, rpl2*(X2), rpl20, rpl22, rpl23(X2), rpl32, rpl33, rpl36</i>                                                                                                                                                                                                                                                              |
|                  | Ribosomal proteins (SSU)                | <i>rps11, rps12(X2), rps14, rps15, rps16*, rps18, rps19, rps2, rps3, rps4, rps7(X2), rps8</i>                                                                                                                                                                                                                                             |
|                  | Transfer RNAs                           | <i>trnA-UGC*(X2), trnC-ACA*, trnC-GCA, trnD-GUC, trnE-UUC*(X3), trnF-GAA, trnG-GCC, trnH-GUG, trnK-UUU*, trnL-CAA(X2), trnL-CAA, trnL-UAA*, trnL-UAG, trnM-CAU(X4), trnN-GUU(X2), trnP-UGG, trnQ-UUG, trnR-ACG(X2), trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-CGU*, trnT-GGU, trnT-UGU, trnV-GAC(X2), trnV-GAC, trnW-CCA, trnY-GUA</i> |
|                  | Ribosomal RNAs                          | <i>rrn16, rrn23, rrn4.5, rrn5</i>                                                                                                                                                                                                                                                                                                         |
|                  | Hypothetical chloroplast reading frames | <i>ycf1(X2), ycf2(X2), ycf3**, ycf4</i>                                                                                                                                                                                                                                                                                                   |
| others           | Other genes                             | <i>ccsA, cemA, clpP**, infA, lhba, matK</i>                                                                                                                                                                                                                                                                                               |

(×n) n gene copies in the IR.\* Genes containing one intron; \*\* genes containing two introns.

360

**Table 4. The lengths of exons and introns in genes with introns in the *C. sclerophylla* chloroplast genome.**

| Gene            | Location | Exon I (bp) | Intron I (bp) | Exon II (bp) | Intron II (bp) | Exon III (bp) |
|-----------------|----------|-------------|---------------|--------------|----------------|---------------|
| <i>clpP</i>     | LSC      | 70          | 851           | 291          | 654            | 227           |
| <i>trnK-UUU</i> | LSC      | 36          | 2511          | 34           |                |               |
| <i>rpoC1</i>    | LSC      | 429         | 838           | 1618         |                |               |
| <i>trnC-ACA</i> | LSC      | 37          | 610           | 55           |                |               |
| <i>ndhB</i>     | IRA      | 776         | 681           | 755          |                |               |
| <i>ndhA</i>     | SSC      | 550         | 1049          | 540          |                |               |
| <i>rpl2</i>     | IRB      | 390         | 685           | 433          |                |               |
| <i>trnA-UGC</i> | IRB      | 36          | 801           | 35           |                |               |
| <i>trnL-UAA</i> | LSC      | 34          | 485           | 49           |                |               |
| <i>trnE-UUC</i> | LSC      | 31          | 956           | 39           |                |               |
| <i>trnT-CGU</i> | LSC      | 34          | 720           | 42           |                |               |
| <i>rps16</i>    | LSC      | 41          | 903           | 227          |                |               |
| <i>ycf3</i>     | LSC      | 125         | 727           | 225          | 768            | 154           |
| <i>atpF</i>     | LSC      | 144         | 789           | 409          |                |               |

388 **Table 5. The codon-anticodon recognition pattern and codon usage for *C. sclerophylla* chloroplast genome.**

| Amino acid | Codon | No.  | RSCU | tRNA            | Amino acid | Codon | No. | RSCU | tRNA            |
|------------|-------|------|------|-----------------|------------|-------|-----|------|-----------------|
| Ala        | GCC   | 196  | 0.61 |                 | Asn        | AAU   | 858 | 1.55 |                 |
| Ala        | GCG   | 158  | 0.49 |                 | Asn        | AAC   | 250 | 0.45 |                 |
| Ala        | GCU   | 581  | 1.81 | <i>trnA-UGC</i> | Pro        | CCG   | 136 | 0.56 |                 |
| Ala        | GCA   | 348  | 1.08 |                 | Pro        | CCA   | 270 | 1.12 |                 |
| Cys        | UGU   | 189  | 1.44 | <i>trnC-ACA</i> | Pro        | CCU   | 362 | 1.5  | <i>trnP-UGG</i> |
| Cys        | UGC   | 73   | 0.56 | <i>trnC-GCA</i> | Pro        | CCC   | 195 | 0.81 |                 |
| Asp        | GAC   | 176  | 0.38 |                 | Gln        | CAG   | 179 | 0.44 |                 |
| Asp        | GAU   | 740  | 1.62 |                 | Gln        | CAA   | 638 | 1.56 |                 |
| Gln        | GAG   | 291  | 0.47 |                 | Arg        | CGA   | 312 | 1.38 |                 |
| Gln        | GAA   | 950  | 1.53 |                 | Arg        | CGC   | 94  | 0.41 |                 |
| Phe        | UUU   | 876  | 1.33 | <i>trnL-UAA</i> | Arg        | CGG   | 99  | 0.44 |                 |
| Phe        | UUC   | 438  | 0.67 | <i>trnN-GUU</i> | Arg        | CGU   | 293 | 1.29 | <i>trnR-ACG</i> |
| Gly        | GGG   | 282  | 0.7  |                 | Arg        | AGA   | 418 | 1.84 |                 |
| Gly        | GGC   | 191  | 0.47 | <i>trnG-GCC</i> | Arg        | AGG   | 145 | 0.64 |                 |
| Gly        | GGU   | 526  | 1.3  |                 | Ser        | UCA   | 339 | 1.2  |                 |
| Gly        | GGA   | 615  | 1.52 |                 | Ser        | UCC   | 285 | 1    | <i>trnS-GGA</i> |
| His        | CAU   | 419  | 1.54 |                 | Ser        | AGC   | 104 | 0.37 |                 |
| His        | CAC   | 126  | 0.46 |                 | Ser        | UCU   | 474 | 1.67 | <i>trnT-UGU</i> |
| Ile        | AUC   | 384  | 0.56 |                 | Ser        | AGU   | 345 | 1.22 |                 |
| Ile        | AUA   | 664  | 0.97 |                 | Ser        | UCG   | 155 | 0.55 | <i>trnT-CGU</i> |
| Ile        | AUU   | 1000 | 1.46 |                 | Thr        | ACU   | 474 | 1.61 |                 |
| Lys        | AAA   | 920  | 1.5  |                 | Thr        | ACC   | 217 | 0.74 |                 |
| Lys        | AAG   | 306  | 0.5  |                 | Thr        | ACA   | 355 | 1.2  |                 |
| Leu        | CUG   | 170  | 0.42 |                 | Thr        | ACG   | 133 | 0.45 |                 |
| Leu        | UUA   | 810  | 1.98 |                 | Val        | GUG   | 180 | 0.56 |                 |
| Leu        | CUA   | 316  | 0.77 |                 | Val        | GUU   | 461 | 1.43 | <i>trnE-UUC</i> |
| Leu        | CUU   | 496  | 1.21 | <i>trnL-UAG</i> | Val        | GUC   | 154 | 0.48 | <i>trnV-GAC</i> |
| Leu        | CUC   | 170  | 0.42 | <i>trnH-GUG</i> | Val        | GUA   | 499 | 1.54 |                 |
| Leu        | UUG   | 492  | 1.2  | <i>trnM-CAU</i> | Trp        | UGG   | 402 | 1    | <i>trnW-CCA</i> |
| Met        | AUG   | 533  | 1    |                 | Tyr        | UAC   | 179 | 0.41 |                 |
|            |       |      |      |                 | Tyr        | UAU   | 690 | 1.59 |                 |

Note: RSCU indicated Relative Synonymous Codon Usage.

389

390

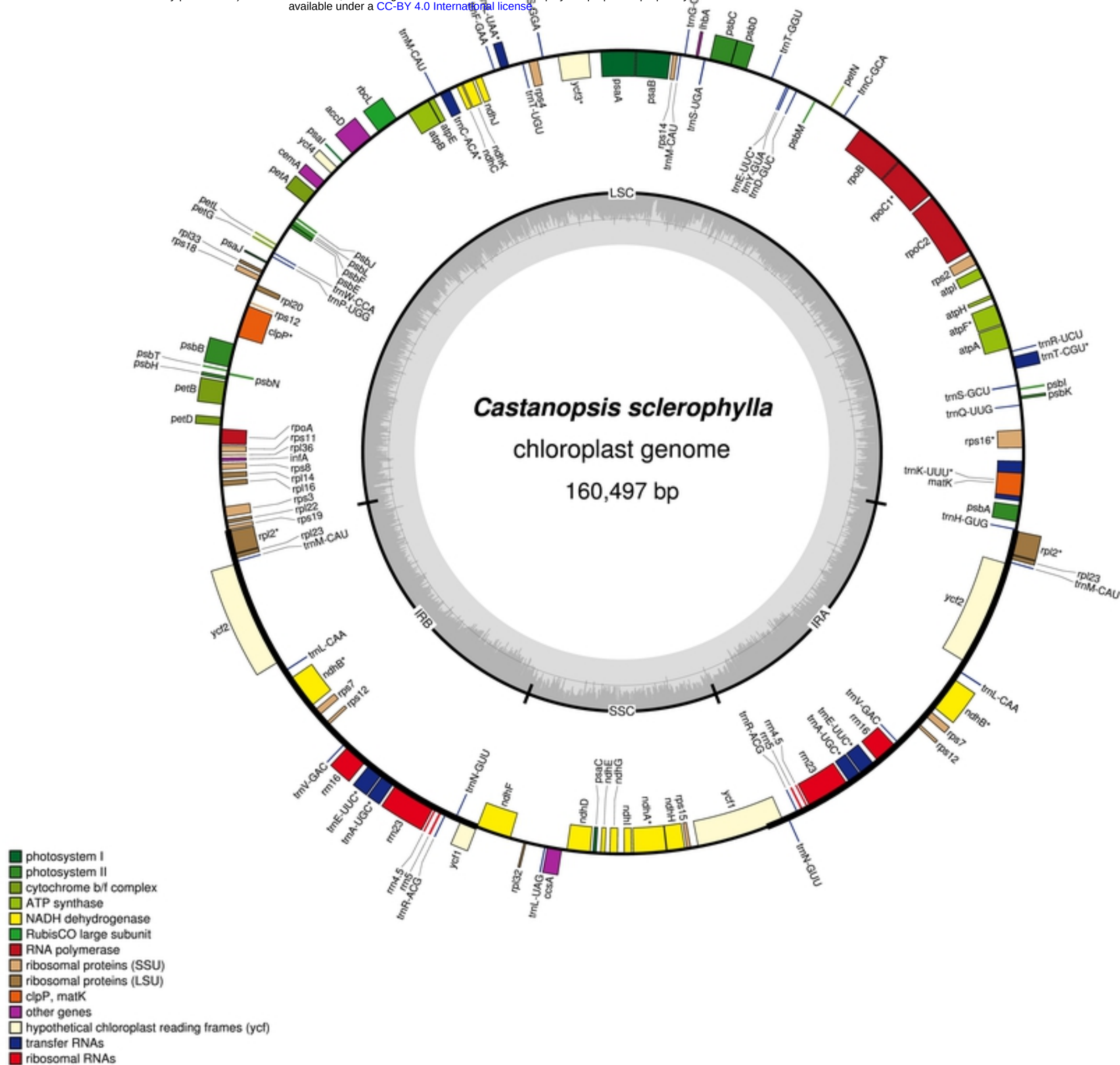


Figure1



# Inverted Repeats



Figure3

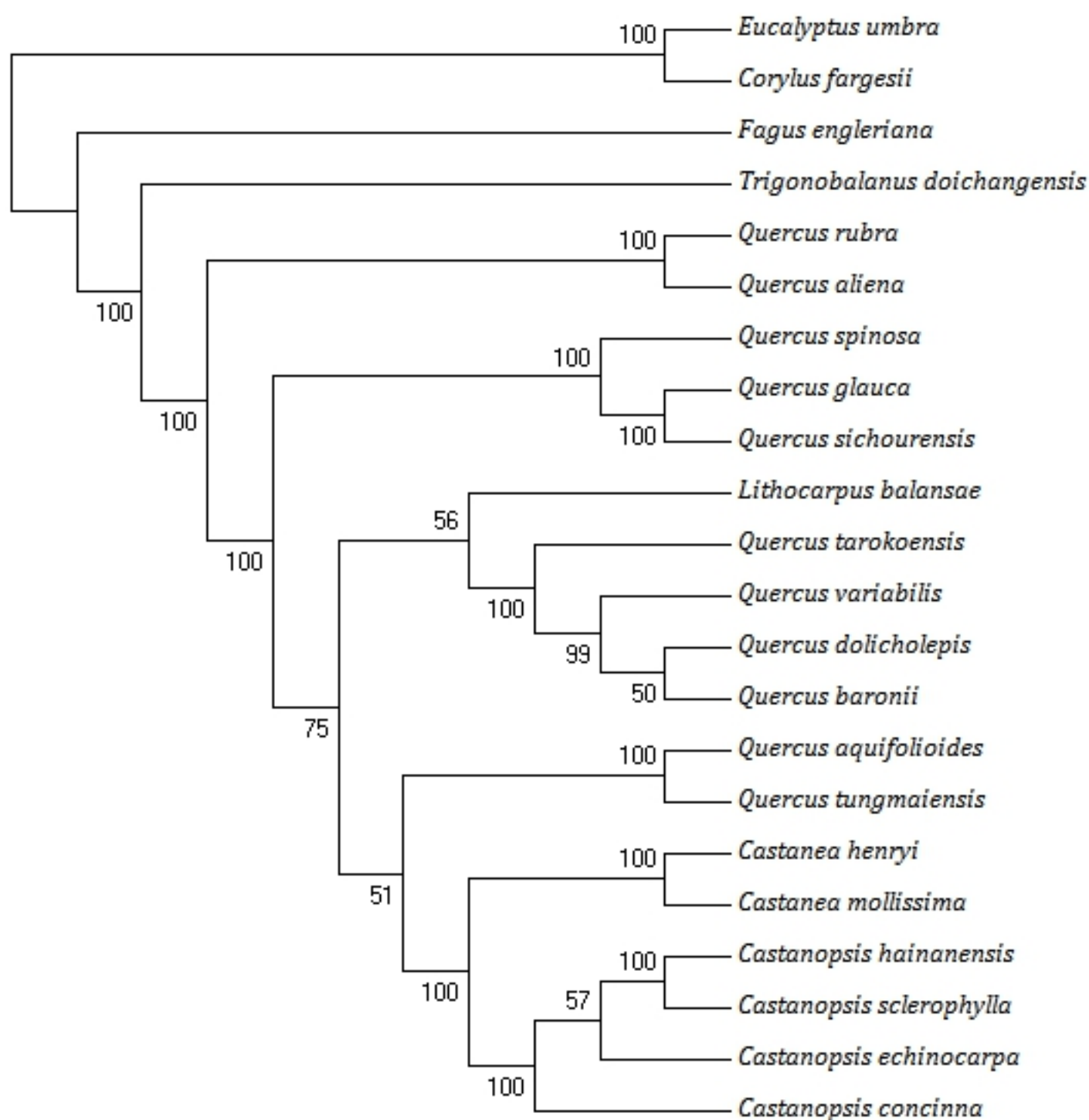


Figure4