Comparative Chloroplast Genomes of *Camellia* Species

Jun-Bo Yang[1☯], Shi-Xiong Yang[2☯], Hong-Tao Li[1*], Jing Yang[1], De-Zhu Li[1,2*]

1 Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China, **2** Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China

Abstract

Background: *Camellia*, comprising more than 200 species, is a valuable economic commodity due to its enormously popular commercial products: tea leaves, flowers, and high-quality edible oils. It is the largest and most important genus in the family Theaceae. However, phylogenetic resolution of the species has proven to be difficult. Consequently, the interspecies relationships of the genus *Camellia* are still hotly debated. Phylogenomics is an attractive avenue that can be used to reconstruct the tree of life, especially at low taxonomic levels.

Methodology/Principal Findings: Seven complete chloroplast (cp) genomes were sequenced from six species representing different subdivisions of the genus *Camellia* using Illumina sequencing technology. Four junctions between the single-copy segments and the inverted repeats were confirmed and genome assemblies were validated by PCR-based product sequencing using 123 pairs of primers covering preliminary cp genome assemblies. The length of the *Camellia* cp genome was found to be about 157kb, which contained 123 unique genes and 23 were duplicated in the IR regions. We determined that the complete *Camellia* cp genome was relatively well conserved, but contained enough genetic differences to provide useful phylogenetic information. Phylogenetic relationships were analyzed using seven complete cp genomes of six *Camellia* species. We also identified rapidly evolving regions of the cp genome that have the potential to be used for further species identification and phylogenetic resolution.

Conclusions/Significance: In this study, we wanted to determine if analyzing completely sequenced cp genomes could help settle these controversies of interspecies relationships in *Camellia*. The results demonstrate that cp genome data are beneficial in resolving species definition because they indicate that organelle-based "barcodes", can be established for a species and then used to unmask interspecies phylogenetic relationships. It reveals that phylogenomics based on cp genomes is an effective approach for achieving phylogenetic resolution between *Camellia* species.

Citation: Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z (2013) Comparative Chloroplast Genomes of *Camellia* Species. PLoS ONE 8(8): e73053. doi: 10.1371/journal.pone.0073053

Editor: Turgay Unver, Cankiri Karatekin University, Turkey

Received April 18, 2013; **Accepted** July 16, 2013; **Published** August 23, 2013

Funding: This work was funded by the National Natural Science Foundation of China (30870169, 31161140350), the National 863 Project of China (2012AA021801) and the Chinese Academy of Sciences through a Large-Scale Scientific Facilities Research Project (2009-LSFGBOWS-01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: lihongtao@mail.kib.ac.cn (H-TL); dzl@mail.kib.ac.cn (D-ZL)

☯ These authors contributed equally to this work.

Introduction

Camellia, a genus containing shrubs and trees, is the largest and most economically, socially, and ecologically valuable genus in the family Theaceae [1–8]. It is native to eastern Asia and is found throughout East and Southeast Asia [3,6], and originated in South and Southwest China [2]. The genus *Camellia*, consisting of more than 200 species [9,10], is not only famous for its ornamental flowers, beverages, and plant oils, but also for its phylogenetic significance. *Camellia* plants provide excellent samples for studying the evolution of the species, interspecific hybridization, and other fundamental life science questions [11]. In addition, *Camellia* tea leaves harbor more than 700 chemical compounds that have been found to promote human health [7,12]. *Camellia* plants are therefore some of the most popular and desirable plants for agriculture, horticulture, and scientific research. Currently, more than 40 countries produce tea for commercial purpose. The annual value of the tea industry in China alone is more than $5 billion USD [12,13]. Many *Camellia* species are domesticated as ornamental plants, while the weeds of others produce high-quality edible oils. Because *Camellia* plants are grown for a variety of uses, they are now found all over the world [3,7,14,15].

Because of their enormous value in commercial, social, and scientific fields, *Camellia* plants have garnered much attention from scientists. The genus *Camellia* represents an excellent example of a taxonomic group under controversial

circumscription and having uncertain phylogenetic affinities that require detailed investigations. The traditional classifications of the genus *Camellia* were mainly based on morphology. The three most recently developed traditional classification methods applied to this genus were established by Sealy [16], Chang [9,10,17] and Ming [1,3,5,6], but these systems have given rise to many conflicting results. Sealy, Chang and Ming disagreed on the boundaries of subgenera, sections, and species, as well as the circumscription and relationships between species. Chang identified about 280 species, while Ming only recognized 119 species of *Camellia*. The genus *Camellia* was divided into 12, 20, and 14 sections by Sealy, Chang, and Ming, respectively. Furthermore, the Sealy system did not offer any subgeneric divisions, but Chang divided *Camellia* into four subgenera and Ming divided it into two. So far, it is uncertain which of these systems most accurately describes the phylogenetic relationships within the genus *Camellia*. As a result, it is necessary to seek other evidence that can be used to rebuild the classification system of *Camellia*.

Molecular methods based on DNA sequence analysis provide useful information for taxonomy, species identification, and phylogenetics. In the last few decades molecular phylogenetics has rapidly developed, and is gaining increasing importance in resolving phylogenetic relationships. Efforts to explore the taxonomy issues, relationships, and the evolution of subdivisions in *Camellia* have involved the use of molecular phylogenetic methods [18–27]. Xiao and Parks [22,23] attempted to resolve *Camellia* taxonomy using introns 11-16 and 23 of the RNA polymerase II (RPB2) gene. However, the poorly resolved results of this study presented completely different findings than the traditional classification methods. Another study based on molecular phylogenetics, the Vijayan et al. [7] study, inferred phylogenetic relationships within the genus *Camellia* using internal transcribed spacer (ITS) sequences of 112 species. These results resolved the 112 species into eight major clades, but the interrelationships between clades remained unresolved. Overall, the results from molecular phylogenetic studies have largely differed from the results of studies using traditional classification methods. In addition, recent studies on *Camellia* leaf the morphology have further complicated the classification of *Camellia* [28–32]. Molecular phylogenetic research on *Camellia* has been extensive applied, but there is no apparent structure associated with its molecular phylogeny, which would help to reveal the true phylogenetic relationships between its species. The major reason for the lack of phylogenic structure is because the genus *Camellia* contains a wide variety of species with complex evolutionary relationships. In addition, the lack of appropriate DNA sequences greatly limits the ability to perform adequate molecular phylogenetic research on *Camellia*. Most of the phylogenetic studies performed to date have suggested that the limited availability of suitable DNA sequences has resulted in finding relatively little genetic variation within the genus *Camellia*. Consequently, achieving phylogenetic resolution and performing species identification have been almost impossible. Currently, the interspecies relationships within the genus *Camellia* remain highly controversial.

Owning to the high cost of DNA sequencing and technological restrictions, molecular phylogenetic analyses have typically been limited. These roadblocks severely restricted the extent to which investigators could analyze DNA, only being able to sequence short segments of DNA contain a small number of informative loci. At present, DNA sequencing costs have fallen dramatically with the rapid development of next-generation DNA sequencing technologies [33–38]. Simultaneously, genomics research has also rapid developed. Phylogenomics [39], which combines genomics with phylogenetics, has become an attractive avenue to help reconstruct the tree of life [40]. The technology behind phylogenomics allows large quantities of entire organellar genomes and even nuclear genomes to be rapidly sequenced. Phylogenomics therefore brings the benefits of affordable genome-scale data collection to the area of phylogenetic resolution. As a result, phylogenetic resolution, especially at low taxonomic levels such as genus, has been substantially improved [41].

Plastids are essential organelles in plant cells. Molecular differences that arise in the chloroplast genomes between plant species and individuals offer promising tools to achieve phylogenetic resolution. The chloroplast (cp) genomes in vascular plants have a conserved quadripartite structure composed of two copies of a large inverted repeat (IR) and two sections of unique DNA, which are referred to as the large single-copy (LSC) regions and small single-copy (SSC) regions, respectively [42,43]. There are many advantages to using the chloroplast genome to achieve phylogenetic resolution rather than the nuclear genome, afforded by its haploid nature, maternal inheritance, single structure, gene content, and high conserved genome structure [44,45]. Complete cp genome sequences have been widely used for phylogenetic resolution in plants. Moore et al. [46] resolved the relationships between basal angiosperms using plastid genome-scale data. Similarly, Jansen et al. [47] used 64 plastid genomes to infer relationships between angiosperms. Moore et al. [48] used 83 chloroplast genomes to further resolve the early diversification of eudicots. Parks et al. [41] increased the phylogenetic resolution at low taxonomic levels using chloroplast genomes. Because plastids offer a complete yet relatively small genome, plastid genome sequencing has become a universal method to obtain evolutionary information that can be used for taxonomical and phylogenetic analyses on plants.

Here, we present the complete nucleotide sequences of cp genomes from seven *Camellia* individuals of six species using Illumina sequencing technology applied to total cp DNA. We aimed to evaluate the suitability of using the analyzed cp genome sequences for taxonomy and phylogenetic resolution between *Camellia* species. A phylogenetic tree formed by seven complete cp genomes belonging to six species was reconstructed. Our analyses of seven *Camellia* individuals provided detailed genetic data that was able to differentiate individuals and species. This study supports the method of applying information from complete chloroplast genome sequencing to taxonomy and phylogenetic resolution of *Camellia*.

Materials and Methods

Plant Materials

Seven plants from six different species, representing different subdivisions of the genus *Camellia*, were sampled. Healthy, clean, fresh green leaves were collected from the seven adult plants. The voucher herbarium specimens for the seven sampled tea plants were deposited at the Herbarium of Kunming Institute of Botany of the Chinese Academy of Sciences (KUN) (Table S1).

Chloroplast DNA Extraction, Sequencing, Genome Assembly, and PCR-based Validation

Total DNA enrichment for chloroplast DNA (cp DNA) extraction was performed as described in Zhang et al. [49] from 100 g of fresh leaves. A 5 mg sample of purified DNA was fragmented and used to construct short-insert libraries according to the manufacturer's manual (Illumina). The DNA from different individuals was indexed using tags and pooled together in one lane of the Illumina's Genome Analyzer for sequencing at the Beijing Genomics Institute (BGI) in Shenzhen, China. The deep-sequencing datasets of seven plants of *Camellia* were deposited into the NIH Short Read Archive (Table S1).

Because the raw sequence reads included non-cp DNA from the nucleus and mitochondria mixed in with the cp DNA, we isolated the cp sequence reads from the raw sequence reads based on all known angiosperm cp genome sequences. The filtered cp sequence reads were used to assemble cp genomes. First, the filtered short reads were assembled into non-redundant contigs using SOAPdenovo [50], a de novo sequence assembly software, with k=31 bp and scaffolding contigs having a minimum size of 100 bp. Then, all contigs were aligned with reference cp genomes, including the cp genomes of plants in the Solanaceae [51,52] and Araliaceae [53,54] families, using the Basic Local Alignment Search Tool (BLAST) database (http://blast.ncbi.nlm.nih.gov/), provided by the National Center for Biotechnology Information (NCBI), using the default search parameters. Next, the order of the aligned contigs was determined according to the reference genomes, and the gaps between the de novo contigs were replaced with consensus sequences of raw reads mapped to the reference genomes. Finally, we acquired preliminary assembly genomes.

The four junctions between the single-copy segments and the inverted repeats were confirmed using PCR-based product sequencing of the preliminary assembled genomes. To avoid assembly errors and to obtain high-quality complete cp genome sequences, we validated genome assembly using intensive PCR-based sequencing. We designed 123 pairs of primers to cover the seven preliminary cp genome assemblies. PCR products were sequenced using the BigDyeV3.1 Terminator Kit for ABI 3730xl (Life Technologies). Sequences obtained using Sanger method were aligned with the assembled genomes using Geneious [55] assembly software to determine if there were any differences. The final complete cp genome sequences of six species of *Camellia* were deposited into the GenBank (Table S1).

Genome Annotation and Repeat Analysis

We annotated the sequenced genomes using the Dual Organellar GenoMe Annotator (DOGMA) database [56], and then manually corrected for start and stop codons and for intron/exon boundaries in order to match the gene predictions of sequenced cp genomes within GenBank and the Chloroplast Genome Database. The sequences of identified tRNA genes were achieved using DOGMA and tRNAscan-SE (version 1.23) [57]. The functional classification of cp genes was determined by referring referred to the CpBase (http://chloroplast.ocean.washington.edu/). The annotated GenBank files of the *Camellia* cp genomes were used to obtain gene maps using the OrganellarGenomeDRAW tool (OGDRAW) [58].

Both direct and inverted repeats were assessed using REPuter [59]. Four types of repeats—dispersed, tandem, palindromic, and gene similarity repeats—were determined within the *Camellia* cp genomes. The maximal length of the gap size between palindromic repeats was restricted to 3 kb. Overlapping repeats were incorporated into one repeat motif whenever possible. Furthermore, a given region in the genome was defined as having only one type of repeat, when one repeat motif could be described as both tandem and dispersed, the region was described as a tandem repeat rather than a dispersed repeat.

Molecular Markers Identification

To examine divergence regions within the seven *Camellia* cp genomes for phylogenetic applications, we extracted all regions, including coding regions, introns and intergenic spacers. Every homologous region was aligned using the Multiple Sequence Alignment Tool (MUSCLE) [60], followed by making additional manual adjustments where necessary. Afterward, the percentage of variable characters within each region was calculated.

For regions that were hotspots of divergence, the maximum parsimony method was used to construct phylogenetic trees using PAUP4.0b10 [61,62], which also allowed us to check the congruence of the phylogenetic tree with the evolution and life history of each species. Heuristic tree searches were conducted using 10,000 random taxon addition replicates (holding 20 trees at each step) and tree bisection-reconnection (TBR) branch swapping with MulTrees in effect. A non-parametric bootstrap analysis was conducted using 1,000 replicates with TBR branch swapping.

Phylogenomic Analyses

We aligned the seven *Camellia* cp genome sequences using the Muitiple Sequence Alignment Program (MAFFT version 5) [63] and made manual adjustments where necessary. Unambiguously aligned DNA sequences were used for phylogenetic analyses, but ambiguously aligned regions were excluded. To check the utility of different genomic regions for phylogenetic resolution, simultaneous analyses were carried out on the following data: (1) the complete cp DNA sequences; (2) the protein-coding exons; (3) the large single-copy region; (4) the small single-copy region; (5) the inverted repeat region; and (6) the introns and spacers.

Maximum likelihood (ML) and maximum parsimony (MP) analyses were conducted using PAUP 4.0b10. Characters were treated as unordered and unweighted. For ML analyses, the best model and parameter settings were chosen using the Akaike information criterion (AIC) as suggested by Modeltest V 3.7 [64,65]. Heuristic searches were conducted with tree bisection-reconnection (TBR) branch swapping, MulTrees in effect, and 10,000 random taxon addition replicates holding 20 trees at each step. Bootstrap support (BS) values for individual clades were calculated by running 1,000 bootstrap replicates of the data, with starting trees acquired by a single replicate of random stepwise addition of taxa, under TBR branch swapping with MulTrees in effect. The consistency index (CI), retention index (RI) and rescaled consistency index (RC) were obtained through PAUP 4.0b10 as the actual number of site differences excluding insertions and deletions (indels).

Bayesian analyses (BA) were conducted using MrBayes 3.2 software [66,67]. The best model and parameters settings were chosen using the Akaike information criterion (AIC) as suggested by ModelTest v 3.7. The results were based on the best-fit models of the AIC test. Four independent Markov Chain Monte Carlo algorithms were performed simultaneously and sampled every 100 generations for 1,000,000 generations. To establish the burn-in phase, i.e., the phase before the log probability values reached stationarity, we plotted generations against log likelihood scores using Excel 2003 (Microsoft, Redmond, WA, USA); the trees identified in the burn-in period were discarded from the analysis.

Results

Genome Assembly and PCR-based Validation

Seven individuals were sequenced to produce 6,539,876 to 7,233,285 paired-end reads (90 bp average read length) using the Illumina Hiseq 2000 system. After screening these paired-end reads by aligning them with reference cp genomes, 108,851 to 112,589 reads were mapped to the reference genomes, reaching, on average, over 100× coverage of the cp genome. After de novo and reference-guided assembly, three complete cp genomes were obtained. The other four cp genomes had four to six gaps, but were complete using PCR-based sequencing.

The four junction regions in each resulting cp genome were validated using PCR-based sequencing. We simultaneously corrected potential errors using PCR-based validation in order to eliminate assembly errors caused by heterogeneous indels from homopolymeric repeats, resulting in complete, high-quality cp genome sequences [38,68]. We designed 123 pairs of primers for the preliminary cp genome assemblies to validate these sequences in each cp genome (Table S2). The validated sequences from each individual reached 172,100 bp. We assembled the high-quality sequences into complete cp genomes using Phred, Phrap, Consed software [69,70]. We then compared these sequences directly to the assembled genomes, and we observed no nucleotide mismatches or indels. These results validated the accuracy of our genome sequencing and assembly methods. We obtained complete cp genome sequences ranging from 156,577 bp to 156,976 bp in length.

Genome Features and Sequence Divergence

As seen in other angiosperms, *Camellia* cp genomes showed a typical quadripartite structure consisting of a pair of IRs (26,025–26,057 bp) separated by the LSC (86,204–86,673 bp) and SSC (18,232–18,318 bp) regions (Figure 1). The cp genomes were found to encode an identical set of 146 predicted functional genes, of which 123 were unique and 23 were duplicated in the IR regions. The 123 unique genes comprised 81, 38 and 4 protein-coding, transfer RNA and ribosomal RNA genes, respectively. Eighteen distinct genes, namely *atp*F, *ndh*A, *ndh*B, *pet*B, *pet*D, *rpl*16, *rpl*2, *rpo*C1, *rps*12, *rps*16, *trn*A-UGC, *trn*G-GCC, *trn*I-GAU, *trn*K-UUU, *trn*L-UAA, and *trn*V-UAC, contained one intron, while two genes (*clp*P and *ycf*3) contained two introns. These introns of all protein-coding genes shared the same splicing mechanism as group II introns [71]. In addition, we identified some unusual start codons, such as ATC for *ndh*D, GTG for *rps*19. Similar noncanonical start codons have been detected in other angiosperms [68,72] and tree fern plants [73].

We found no genes with lost or reduced functioning in *Camellia* cp genomes. The *ycf*1_like gene in the junction region of IRb and SSC was the only pseudogene found, and arose because of incomplete duplication of the normal copy of *ycf*1 in the IRa and SSC junction region (Figure 1). Similar mutations have been identified in the cp genomes of other angiosperm species [68].

A total of 60.52%-60.71% of the *Camellia* cp genomes were made up of coding regions. Overall, 52.82%-53%, 1.91%-1.92%, and 5.76%-5.78% of the genome sequence encoded proteins, tRNAs, and rRNAs, respectively. The remaining 39.29%-39.48% of the genome was made up of non-coding regions filled with introns, intergenic spacers, and pseudogenes. Similar to other angiosperm cp genomes [72,73], *Camellia* cp genomes was also found to be AT-rich, with overall AT and GC content is 62.7% and 37.3%, respectively. In general, the genome features of the seven *Camellia* cp genomes analyzed in this study were found to be quite similar in terms of gene content, gene order, introns, intergenic spacers, and AT content, and the sequences identity to 98.5%.

Sequences were plotted to check their identity using the mVISTA tool [74] by aligning the seven *Camellia* cp genomes with *Panax ginseng* [53] as a reference. The sequences identity percentage is 93% of *Camellia* species and reference. Moderate genetic divergence in *Camellia* species was detected. Taken together, the aligned sequences showed moderate divergence with more than 20 regions having sequence similarities below 60%. These results suggested that *Camellia* cp genomes contain moderate genetic differentiation especially in the noncoding and single-copy regions. More than 10 divergent hotspot regions were identified (Figure 2).

SNPs analyses were conducted using SAMtools [75] and Venn diagram showing overlap of SNPs identified were made (Figure S1). Simultaneously, the variant positions (substitution, indels) and variation types (transition, transversion) were

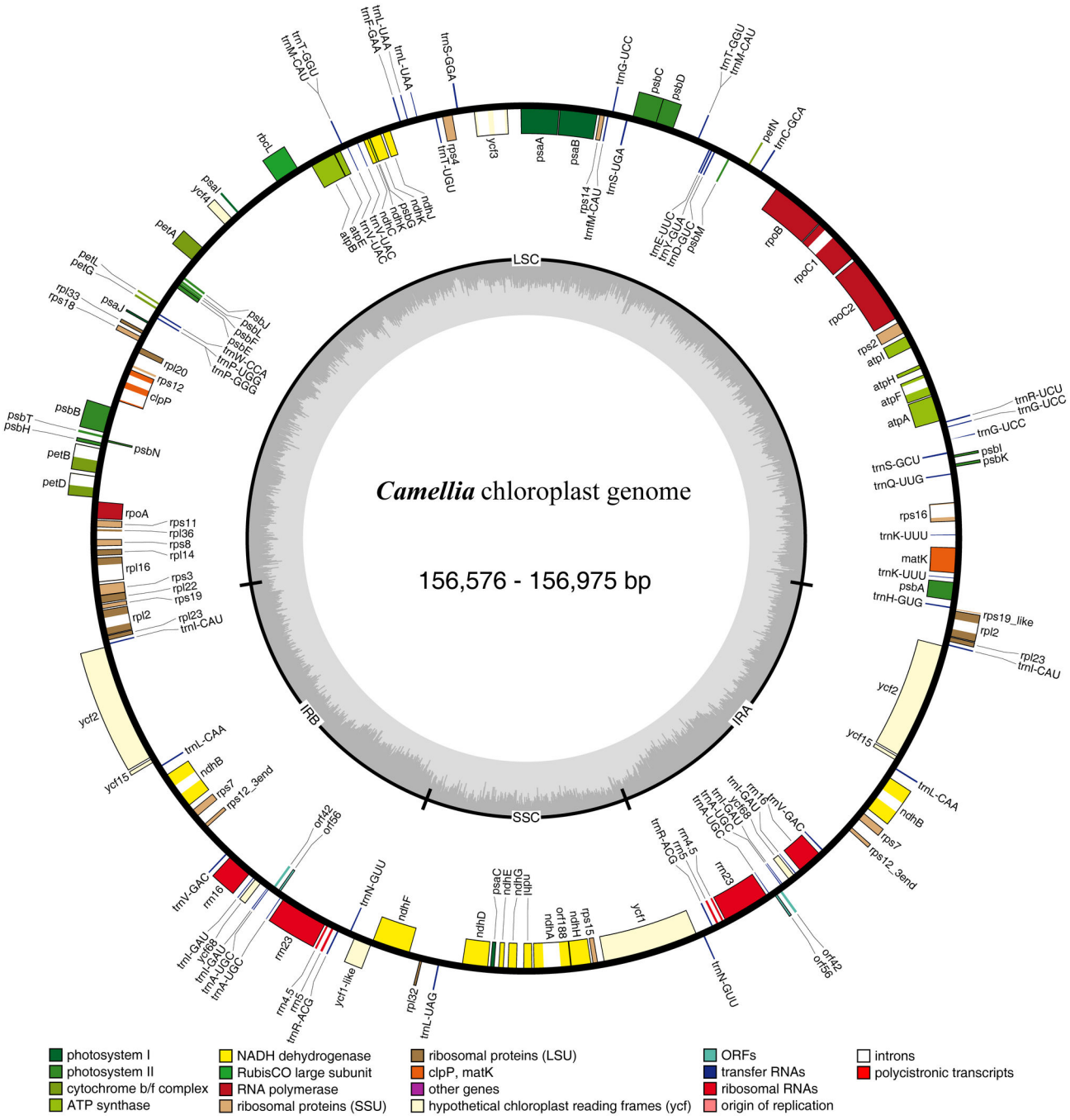

Figure 1. Gene map of the *Camellia* chloroplast genomes. Genes shown outside the outer circle are transcribed clockwise and those inside are transcribed counterclockwise. Genes belonging to different functional groups are color-coded. Dashed area in the inner circle indicates the GC content of the chloroplast genome.

doi: 10.1371/journal.pone.0073053.g001

aggregated and summarized according to the coding, intron-spacers, IR, LSC and SSC regions (Figure S2).

P-distances were used to estimate the average genetic divergences of the seven *Camellia* individuals. The results showed that the p-distances in all individuals, between species, and within individuals were 0.000829, 0.00118 and 0.00003, respectively. These results suggest that moderate interspecies genetic divergence existed within the genus *Camellia*. In addition, we found that interspecies sequence divergence was

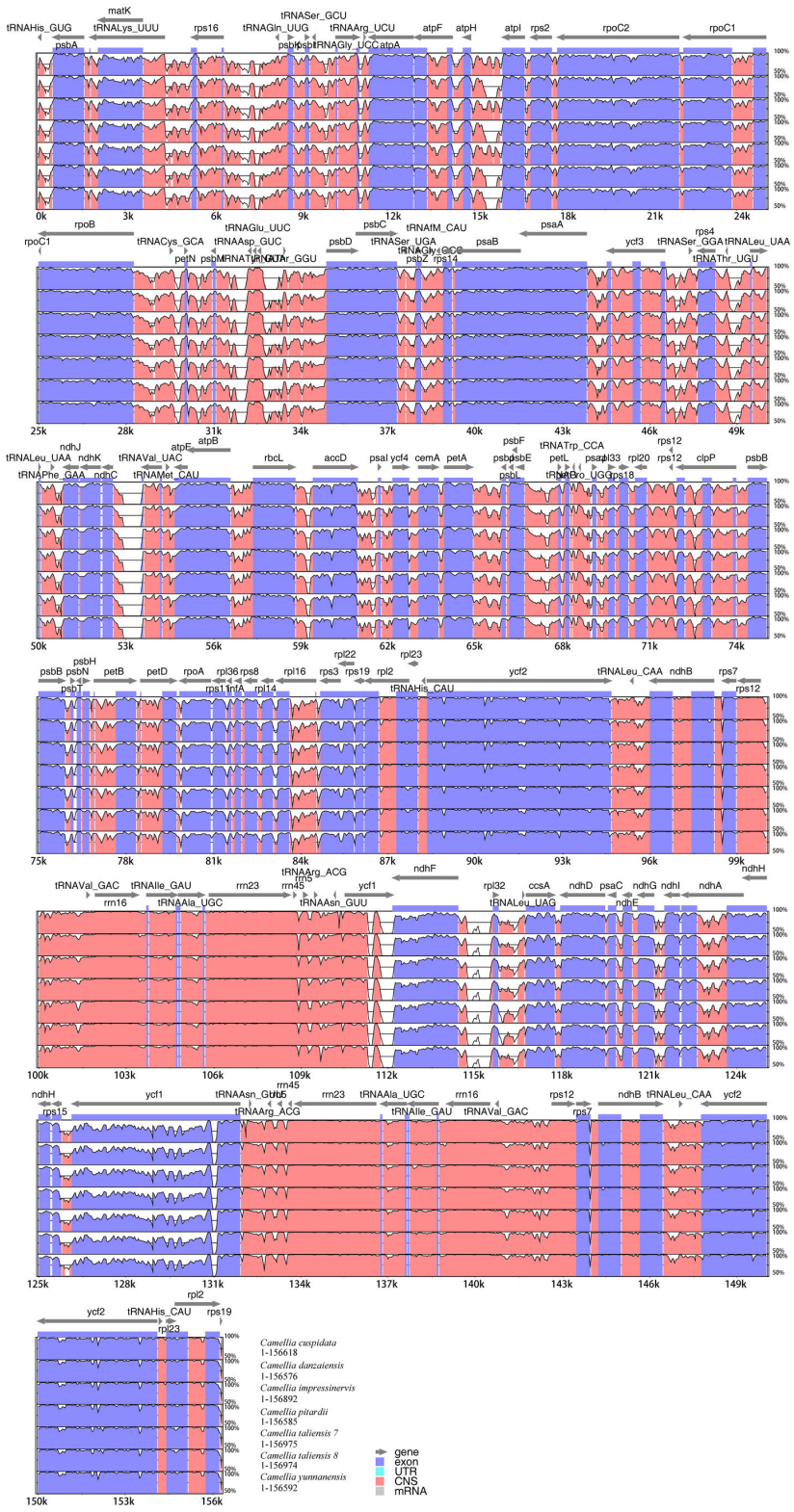

Figure 2. Visualization of alignment of the seven *Camellia* chloroplast genome sequences. VISTA-based identity plots showing sequence identity between seven sequenced chloroplast genomes and the one published chloroplast genomes of Araliaceae, with *Panax ginseng* as a reference. Genome regions are color coded as protein coding, rRNA coding, tRNA coding or conserved noncoding sequences.

doi: 10.1371/journal.pone.0073053.g002

much more pronounced than intraspecies sequence divergence.

Repetitive Sequences

Four categories of repeats—dispersed, tandem, palindromic and gene similarity repeats [49,76] — were identified using REPuter [59] and manual verification of sequence having a copy size of 30 bp or longer and a sequence similarity greater than 90%. Repeat analysis identified more than 300 repeats in the seven *Camellia* cp genomes. The longest repeat, other than the IRs, was 65 bp in length. Most of the repeated sequences were located in the intergenic regions, while some were found in protein-coding regions.

Analysis of IRs

Our study showed that the IRs of *Camellia* were representative of the typical dicot cp genome structure, in which the IRs expanded to the *rps*19 and *ycf*1 genes. In IR-LSC, the 5'-end of *rps*19 partially fell within the IRs, and the IRs expanded to the 5'-end of *ycf*1 in IR–SSC.

Genome Divergent Hotspot Regions

Hotspot regions of sequence divergence were identified using a genome-wide comparative analysis of seven *Camellia* whole cp genomes. The results suggested that 11 hotspot regions (accD-*psa*I, *atp*F-*atp*H, *ccs*A-*ndh*D, *clp*P-*psb*B, *ndh*C-*trn*V, *ndh*F-*rpl*32, *pet*D-*rpo*A, *psb*H-*pet*B, *rpl*32-*trn*L, *trn*G_intron, *trn*S-*trn*G) could be applied to the phylogenetic analysis of *Camellia*. All hotspot regions contained more than 1% variable characters.

Phylogenomic Analyses

Six data partitions (complete cp DNA sequences, protein-coding exons, the large single-copy region, the small single-copy region, the inverted repeat region and introns and spacers) from the seven *Camellia* cp genomes and four outgroups (NC_006290, NC_016430, NC_004561, NC_007062 from GenBank) [51–54] were used for phylogenetic analyses. Excluding outgroups, the sequence characteristics of the ingroups associated with the six datasets are shown in Table S3. The small single-copy region harbored the highest percentage of variable characters, at 0.67%, followed by the introns and spacers with 0.61%. The large single-copy region and the protein-coding exons also possessed moderate genetic variation, reporting 0.48% and 0.34% variable characters, respectively. The inverted repeat region was highly conserved, having the fewest, less than 0.2%, variable characters.

Phylogenetic trees with bootstrap values (BS) and posterior probabilities (PP) were built based on the previously discussed six datasets (Figure 3). The method of data analysis (ML, MP, or BA) had no effect on the resulting phylogenetic trees, and their topologies were also found to be highly similar. Phylogenetic trees produced according to each of the six datasets were largely congruent with each other. These findings suggest that there were no conflicts between partitions of the cp genome. The results also revealed that the phylogenetic resolution and the support values of nodes increased significantly with the increasing of the sequences (Figure 3).

All analyses (ML and MP) generated a single phylogenetic tree in each dataset. The topology of these phylogenetic trees consisted of dichotomous branches for resolving phylogenetic relationships using the complete cp DNA sequences and the introns and spacers. By contrast, phylogenetic analyses using the other four datasets did not provide much information to help in the phylogenetic resolution of *Camellia*.

Discussion

Genome Organization

Structural rearrangements and gene loss-and-gain events often occur in some angiosperms, and are especially common in monocot cp genomes. A representative example is the cp genome of the Poaceae, in which three inversions within the LSC regions and gene translocation of the *rpl*23 gene from the IR to the LSC regions constitutes a disruption to the canonical order [77]. Indels and gene loss (deletions or becoming a pseudogene) are also frequently found in Poaceae cp genomes, as evidenced by intron loss within *rpo*C1, insertion within *rpo*C2, and gene loss in accD, *ycf*1, and *ycf*2 [78,79]. Other monocot families also display rearrangements and gene-loss events in their cp genomes. *Phalaenopsis* and *Oncidium* have lost most of their *ndh* genes [80,81], while *Lemna*, *Dioscorea* and two Acoraceae members each lost a single gene: *inf*A, *rps*16, and accD, respectively [78,82,83]. Rearrangements have also occurred in *Dioscorea*, such as the inversion of SSC [83]. Similarly, rearrangements and gene loss-and-gain events have also occurred in dicots. Geraniaceae cp genomes have experienced remarkable genomic changes [84], such as the loss of *ndh* genes in *Erodium* [85]. Some legumes do not have the IR and have lost the *rps*16 gene [36,86–88]. Usually, IR expansion is quite common, such as the expansion of single-copy *rps*19 and *rpl*22 genes from the LSC into the IRs as a result of gene duplications [89,90]. The *Pelargonium* cp genomes contain massive IR expansions, also due to gene duplications [91]. IR contractions are also common, such as those observed in the subfamily Apioideae [92]. However, we found that the genome organization of *Camellia* wais relatively well conserved. The gene order within the *Camellia* genomes was identical to the gene order in the published Solanaceae and Araliaceae genomes. The cp genomes of the six *Camellia* species were very similar to the cp genome of standard angiosperms, and were distinctly different from the cp genomes of monocots in structure and content. We detected no structural rearrangements, IR expansions, or gene loss-and-gain events in *Camellia* cp genomes. And, as the previous study [93], the *ycf*15 gene, employing an ATG start codon, is likely a functional gene.

Repetitive Sequences

The presence of repeats in cp genomes, especially in intergenic spacer regions, has been reported in all published angiosperm lineages. Compared with other angiosperm species, the number of repeats found in *Camellia* is rather high.

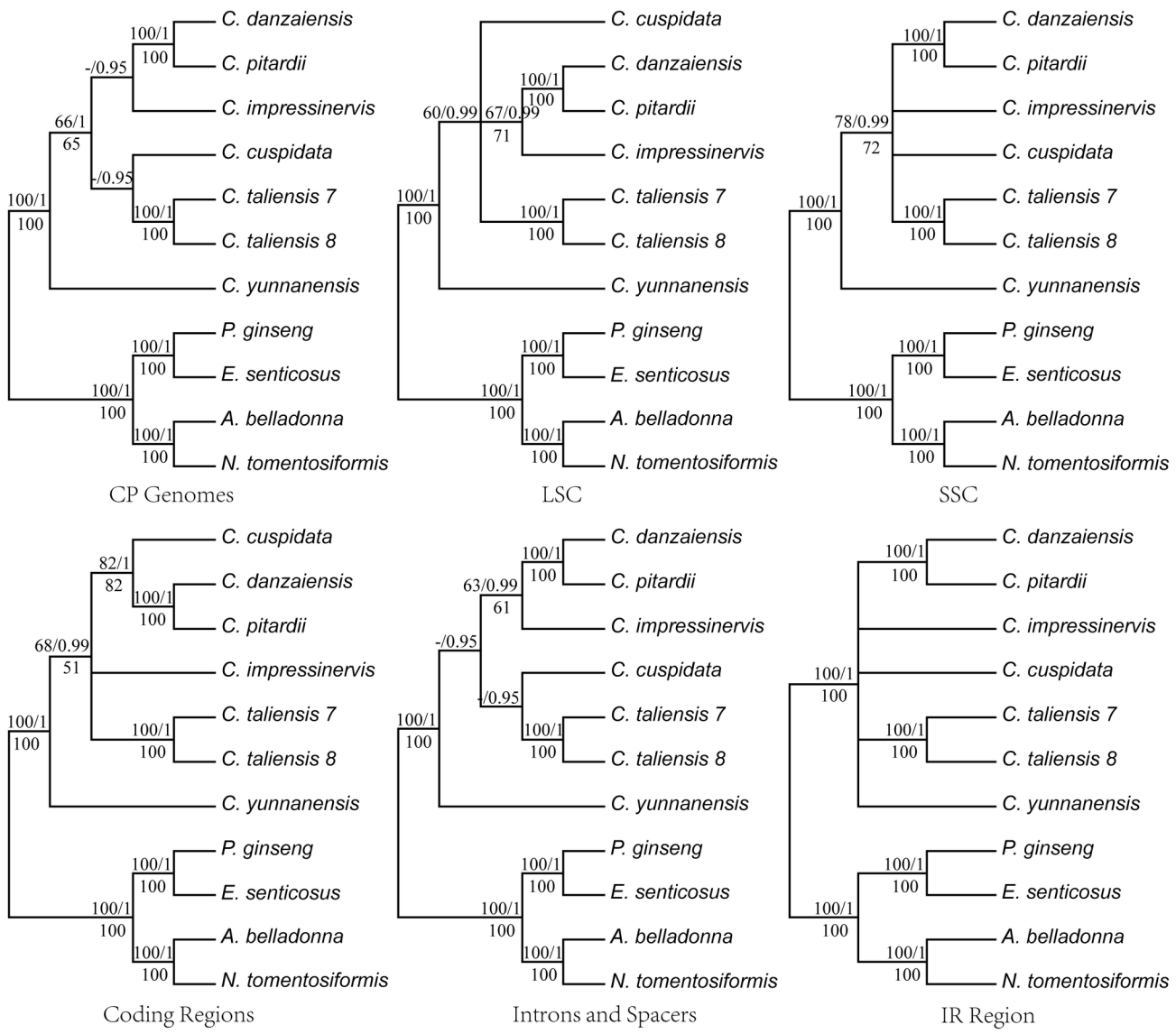

Figure 3. Maximum parsimony trees of all the six chloroplast datasets for seven *Camellia* individuals. Numbers above the lines on the left indicate the maximum parsimony bootstrap of each clade >50%, numbers above the lines on the right indicate the Bayesian posterior probabilities, numbers below each branch are the maximum likelihood bootstrap of each clade >50%.

doi: 10.1371/journal.pone.0073053.g003

In all, more than 300 repeats were detected in the seven *Camellia* cp genomes. The numbers and distributions of the four repeat types were found to be remarkably similar and conserved among the seven cp genomes. Among these repeats, tandem repeats were the most common, accounting for 42% of the total number of repeats, while gene similarity repeats only made up 4%. Except for a few repeats, which were found in the genes *inf*A, *rpo*C2, *rps*18 and *rps*3, the majority of repeats were located in noncoding regions. The lengths of repeats found in *Camellia* range from 30 to 61, representing much shorter repeats than those in the Poaceae, some of which have measured 91-bp and 132-bp [49,94].

Previous research has suggested that repeat sequences may play roles in rearranging sequences and producing variation which cp genomes through illegitimate recombination and slipped-strand mispairing [76,95,96]. Our research also showed that divergent regions of the cp genome were associated with repeat sequences; for example, the *rpo*C2 gene harbored various repeats. It is possible that repeat sequences also correlate with genome rearrangement in *Camellia* cp genomes.

Genome Divergent Hotspot Regions

Aligning entire chloroplast genomes revealed that *Camellia* chloroplast genomes are relatively well conserved. Furthermore, similar to other angiosperms, the noncoding regions show greater sequence divergence than the coding regions, among the six *Camellia* species studied. Although the gene order and content between *Camellia* cp genomes were found to be highly conserved, the differences that do exist may indicate of species variation and differentiation. The phylogenetic analyses on the complete cp genomes of six *Camellia* species provided enough evidence for unique variations between the different lineages. The observed rates of interspecies nucleotide polymorphism were moderate at 0.12%.

In this study, 11 hotspot regions of divergence were checked, and were reported to have more than 1% variable characters. Of these regions, 11 intergenic regions harboring high phylogenetic information were newly identified in our study. Previous studies have also shown that noncoding regions of chloroplast genomes could be successfully used for phylogenetic studies in angiosperms [80,97,98]. The new divergence hotspot regions found in our study could potentially be used as molecular targets for future phylogenetic studies. Furthermore, developing universal primers for these hotspot regions could aid in revealing the molecular phylogeny of other *Camellia* species.

Phylogenetic Implications

Phylogenomic analyses have revealed that different species within a genus are associated with moderate genetic differentiation. Furthermore, individuals of the same species but from different distributions also have moderate genetic differentiation and can therefore be distinctively classified. For example, two individuals of *C. taliensis*, both share a common monophyletic node, yet they harbor 16 variable sites. Regardless of the level moderate genetic differentiations may provide enough phylogenetic information to distinguish between species or even individuals. The sites of sequence variation occur primarily in intergenic regions, such as *ndh*C-*trn*V, *pet*D-*rpo*A, *trn*S-*trn*G, etc. The results of our study show that analyses of entire cp genomes significantly contribute to species identification and phylogenetic studies.

Our phylogenetic analysis of *Camellia* did not agree with any of the traditional classification methods used recently in *Camellia* taxonomy. Such as *C. danzaiensis*, *C. impressinervis* and *C. taliensis*, belonging to the Subgen. *Thea* according Chang, did not form a monophyletic clade. Similarly, *C. pitardii* and *C. yunnanensis*, belonging to Subgen. *Camellia* according to Ming, dispersed into the clade of Subgen. *Thea* comprising *C. taliensis*, *C. cuspidata*, *C. impressinervis* and *C. danzaiensis* (Figure 3). Taxonomic studies on *Camellia* are very controversial. Traditional classification systems conflict with each other, especially in terms of species definition; the number of *Camellia* species has been reported any where from 119 to 280, depending on the classification systems used. However, defining species of *Camellia* using analyses of entire cp genomes provides a feasible way to resolve the controversial taxonomy of *Camellia*.

Previous molecular phylogenetic research failed to resolve the phylogenetic relationships of *Camellia* for a variety of reasons. Overall, previous phylogenetic studies did not contain enough informative characters, used samples that may have undergone hybridization, resulted in incomplete lineage sorting, involved stochastic properties, or used non-concerted evolution ITS markers. A comparative analysis using the entire cp genome revealed many informative characters; compared with prior analyses of short sequences in *Camellia*, our analyses on the entire cp genomes contained more than 100 times the number of parsimony-informative characters, and resulted in phylogenetic trees with better-resolved nodes and higher support values. While analyses using entire cp genomes may still be insufficient to fully resolve all phylogenetic relationships [41,99,100], our results suggest that this type of whole-genome phylogenomic analyses will provide solutions to many disputes and guide the way for phylogeny in *Camellia*.

Furthermore, with the rapid development of next-generation DNA sequencing technologies, the sequencing costs have dramatically fallen and the sequencing accuracy has significantly improved. As a result, genome sequencing of organelles and phylogenomic analyses are becoming a reasonable way to improve resolution in phylogenetic studies, especially at low taxonomic levels. In the near future, sequencing the genomes of thousands of organelles will greatly benefit to break the current limitations that arise from using short sequences to carry out phylogenetic studies [41,101,102]. The “barcodes” associated with entirely sequenced cp genomes [101,103] will significantly improve our ability to distinguish between and identify different species. Especially for groups mired in controversy over species definition, organelle-based genome barcodes will help promote taxonomic studies and contribute to the establishment of natural classification systems.

In this study, we sequenced seven individuals, representing six species of *Camellia* using Illumina sequencing-by-synthesis technology. The sequenced cp genomes provided large amounts of genetic information to aid in the species identification and phylogenetics of these economically important plants. The analyzed cp genomes showed moderate genetic variations, which may provide enough genetic information to further species identification and species definition efforts. At the same time, this information may also provide enough adequate phylogenetic information to resolve the evolutionary relationships between species of *Camellia*. Our results show that whole-genome analyses using *Camellia* chloroplast genomes provide an effective and feasible approach to resolve species identification issues and support phylogenetic applications in the study of *Camellia*.

Supporting Information

Figure S1. Venn diagram showing overlap of seven *Camellia* individuals of SNPs identified. (TIF)

Figure S2. Bar graph summarizing the variant positions and variation types in the different regions. (TIF)

Table S1. Sampled species and voucher specimens of *Camellia* used in this study. (DOC)

Table S2. Primers used for gap closure, assembly and junction verification. (DOC)

Table S3. DNA site variation and tree statistics for the six datasets used in the phylogenomic analyses presented in this study. (DOC)

Acknowledgements

We are very grateful to Mr. Zhi-Rong Zhang, Mr. Wu-Xiang Fu, Ms. Juan-Hong Zhang and Ms. Na Yang of Kunming Institute of Botany for their help with the experiments.

Author Contributions

Conceived and designed the experiments: J-BY S-XY H-TL D-ZL. Performed the experiments: J-BY H-TL. Analyzed the data: H-TL. Contributed reagents/materials/analysis tools: J-BY H-TL JY. Wrote the manuscript: J-BY H-TL D-ZL.

References

1. Ming TL, Zhang WJ (1996) The evolution and distribution of genus *Camellia*. Acta Bot Yunnan 18: 1-13.
2. Kole C (2011) Wild crop relatives: Genomic and breeding resources cereals. New York: Springer Verlag. p. xxiii, 497
3. Ming TL (2000) Monograph of the genus Camellia. Kunming, China: Yunnan Science and Technology Press.
4. Lu HF, Pi EX, Peng QF, Wang LL, Zhang CJ (2009) A particle swarm optimization-aided fuzzy cloud classifier applied for plant numerical taxonomy based on attribute similarity. Expert Syst Appl 36: 9388-9397. doi:10.1016/j.eswa.2008.12.065.
5. Ming TL (1999) A systematic synopsis of the genus *Camellia*. Acta Bot Yunnan 21: 149-159.
6. Ming TL, Bartholomew B (2007) Theaceae. In: ZY WuPH RavenDY Hong. Flora of China. Beijing & St. Louis: Science Press & Missouri Botanical Garden Press. pp. 367.
7. Vijayan K, Zhang WJ, Tsou CH (2009) Molecular taxonomy of *Camellia* (Theaceae) inferred from nrITS sequences. Am J Bot 96: 1348-1360. doi:10.3732/ajb.0800205. PubMed: 21628283.
8. Vijayan K, Chung MC, Tsou CH (2012) Dispersion of rDNA loci and its implications on intragenomic variability and phylogenetic studies in *Camellia*. Sci Hort 137: 59-68. doi:10.1016/j.scienta.2012.01.021.
9. Chang HT (1998) Flora of Reipublicae Popularis Sinicae. Delectis Florae Republicae Popularis Sinicae, Agendae Academiae Sinicae Edita, Tomus. Beijing, China: Science Press. pp. 101-113.
10. Chang HT (1981) A taxonomy of the genus *Camellia*. J Sun Yatsen Univ 1: 1-180.
11. Ye CX (1997) Classification in the genus *Camellia* L. American Camellia Society: American Camellia Yeabook. 9-23.
12. Khan N, Mukhtar H (2007) Tea polyphenols for health promotion. Life Sci 81: 519-533. doi:10.1016/j.lfs.2007.06.011. PubMed: 17655876.
13. Chen L, Zhou ZX, Yang YJ (2007) Genetic improvement and breeding of tea plant (*Camellia sinensis*) in China: from individual selection to hybridization and molecular breeding. Euphytica 154: 239-248. doi: 10.1007/s10681-006-9292-3.
14. Gao JY, Parks CR, Du YQ (2005) Collected species of the genus *Camellia*: an illustrated outline. Zhejiang, China: Zhejiang Science and Technology Press.
15. Zhang DL, Yu JF, Chen YZ, Zhang R (2007) Ornamental teaoil from *Camellia* cultivars and their hypocotyl graft propagation. SNA Research Conference 52: 257-260.
16. Sealy JR (1958) A revision of the genus Camellia. Royal Horticultural Society.
17. Chang HT, Bartholomew B (1984) Camellias. Portland, Oregon, U S A: Timber Press.
18. Thakor BH, Parks CR (1997) Phylogenetic relationships within the genus *Camellia* (Theaceae). Am J Bot 84: 237.
19. Thakor BH (1998) Phylogenetic relationships among the species of genus *Camellia* (Theaceae). University of North Carolina at Chapel Hill.
20. Tang S, Zhong Y (2002) A phylogenetic analysis of nrDNA ITS sequences from ser. Chrysantha (Sect. Chrysantha, Camellia, Theaceae). J Genet Mol Biol 13: 105-107.
21. Yang JB, Li HT, Yang SX, Li DZ, Yang YY (2006) The application of four DNA sequences to studying molecular phylogeny of *Camellia* (Theaceae). Acta Bot Yunnan 28: 108-114.
22. Xiao T (2001) Phylogenetic relationships of the genus Camellia (Theaceae) based on the RNA polymerase II (RPB2) sequences. University of North Carolina at Chapel Hill.
23. Xiao TJ, Parks CR (2002) Molecular analysis of genus *Camellia*. American Camellia Society: American Camellia Yeabook: 52-58.
24. Fang W, Yang JB, Yang SX, Li DZ (2010) Phylogeny of Camellia sects. Longipedicellata, Chrysantha and Longissima (Theaceae) based on sequence data of four chloroplast DNA loci. Acta Bot Yunnan 32: 1-13.
25. Su MH, Hsieh CF, Tsou CH (2009) The confirmation of *Camellia formosensis* (Theaceae) as an independent species based on DNA sequence analyses. Bot Stud 50: 477-485.
26. Prince LM, Parks CR (2001) Phylogenetic relationships of Theaceae inferred from chloroplast DNA sequence data. Am J Bot 88: 2309-2320. doi:10.2307/3558391. PubMed: 21669662.
27. Liu Y, Yang SX, Ji PZ, Gao LZ (2012) Phylogeography of *Camellia taliensis* (Theaceae) inferred from chloroplast and nuclear DNA: insights into evolutionary history and conservation. BMC Evol Biol 12: 92. doi:10.1186/1471-2148-12-92. PubMed: 22716114.
28. Pi EX, Peng QF, Lu HF, Shen JB, Du YQ et al. (2009) Leaf morphology and anatomy of *Camellia* section *Camellia* (Theaceae). Bot J Linn Soc 159: 456-476. doi:10.1111/j.1095-8339.2009.00952.x.
29. Lu HF, Shen JB, Lin XY, Fu JL (2008) Relevance of Fourier transform infrared spectroscopy and leaf anatomy for species classification in *Camellia* (Theaceae). Taxon 57: 1274-1288.
30. Luna I, Ochoterena H (2004) Phylogenetic relationships of the genera of Theaceae based on morphology. Cladistics 20: 223-270. doi: 10.1111/j.1096-0031.2004.00024.x.
31. Lu HF, Jiang W, Ghiassi M, Lee S, Nitin M (2012) Classification of *Camellia* (Theaceae) species using leaf architecture variations and pattern recognition techniques. PLOS ONE 7(1): e29704. doi:10.1371/journal.pone.0029704. PubMed: 22235330.
32. Jiang B, Peng QF, Shen ZG, Moller M, Pi EX et al. (2010) Taxonomic treatments of *Camellia* (Theaceae) species with secretory structures based on integrated leaf characters. Plant Syst Evol 290: 1-20. doi: 10.1007/s00606-010-0342-x.
33. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8: 3-17. doi:10.1111/j.1471-8286.2007.02019.x. PubMed: 21585713.
34. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133-141. doi:10.1016/j.tig.2007.12.007. PubMed: 18262675.
35. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145. doi:10.1038/nbt1486. PubMed: 18846087.
36. Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T et al. (2010) The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17: 11-22. doi:10.1093/dnares/dsp025. PubMed: 20007682.
37. Cronn R, Liston A, Parks M, Gernandt DS, Shen R et al. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36: e122-. PubMed: 18753151.
38. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG et al. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes.

BMC Plant Biol 6: 17. doi:10.1186/1471-2229-6-17. PubMed: 16934154.
39. Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics. Science 300: 1706-1707. doi:10.1126/science.1086292. PubMed: 12805538.
40. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361-375. doi: 10.1038/nrg1603. PubMed: 15861208.
41. Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol 7: 84. doi:10.1186/1741-7007-7-84. PubMed: 19954512.
42. Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW et al. (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol 395: 348-384. doi:10.1016/S0076-6879(05)95020-9. PubMed: 15865976.
43. Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10: 823-833. doi:10.1007/BF00418529.
44. Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Mo Bot Gard 75: 1180-1206. doi:10.2307/2399279.
45. Tian X, Li DZ (2002) Application of DNA sequences in plant phylogenetic study. Acta Bot Yunnan 24: 170-184.
46. Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A 104: 19363-19368. doi: 10.1073/pnas.0708072104. PubMed: 18048334.
47. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104: 19369-19374. doi:10.1073/pnas.0709121104. PubMed: 18048330.
48. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci U S A 107: 4623-4628. doi:10.1073/pnas.0907801107. PubMed: 20176954.
49. Zhang YJ, Ma PF, Li DZ (2011) High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLOS ONE 6: e20596. doi:10.1371/journal.pone.0020596. PubMed: 21655229.
50. Li R, Zhu H, Ruan J, Qian W, Fang X et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265-272. doi:10.1101/gr.097261.109. PubMed: 20019144.
51. Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG et al. (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: The role of RNA editing in generating divergence in the process of plant speciation. Mol Biol Evol 19: 1602-1612. doi:10.1093/oxfordjournals.molbev.a004222. PubMed: 12200487.
52. Yukawa M, Tsudzuki T, Sugiura M (2006) The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. Mol Genet Genomics 275: 367-373. doi:10.1007/s00438-005-0092-6. PubMed: 16435119.
53. Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res 11: 247-261. doi:10.1093/dnares/11.4.247. PubMed: 15500250.
54. Yi DK, Lee HL, Sun BY, Chung MY, Kim KJ (2012) The complete chloroplast DNA sequence of *Eleutherococcus senticosus* (Araliaceae); comparative evolutionary analyses with other three asterids. Mol Cells 33: 497-508. doi:10.1007/s10059-012-2281-6. PubMed: 22555800.
55. Meintjes P, Duran C, Kearse M, Moir R, Wilson A et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12): 1647-1649. doi:10.1093/bioinformatics/bts199. PubMed: 22543367.
56. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255. doi: 10.1093/bioinformatics/bth352. PubMed: 15180927.
57. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964. doi:10.1093/nar/25.5.955. PubMed: 9023104.
58. Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52: 267-274. doi:10.1007/s00294-007-0161-y. PubMed: 17957369.
59. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J et al. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633-4642. doi:10.1093/nar/29.22.4633. PubMed: 11713313.
60. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797. doi: 10.1093/nar/gkh340. PubMed: 15034147.
61. Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.
62. Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6: Unit 6.4. John Wiley & Sons, Inc., Malden, MA, United States. PubMed: 18428704
63. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518. doi:10.1093/nar/gki198. PubMed: 15661851.
64. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817-818. doi:10.1093/bioinformatics/14.9.817. PubMed: 9918953.
65. Posada D (2003) Using MODELTEST and PAUP* to select a model of nucleotide substitution. Curr Protoc Bioinformatics Chapter 6: Unit 6.5. John Wiley & Sons, Inc., Malden, MA, United States. PubMed: 18428705
66. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574. doi: 10.1093/bioinformatics/btg180. PubMed: 12912839.
67. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A et al. (2012) MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539-542. doi: 10.1093/sysbio/sys029. PubMed: 22357727.
68. Yang M, Zhang X, Liu G, Yin Y, Chen K et al. (2010) The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLOS ONE 5: e12762. doi:10.1371/journal.pone.0012762. PubMed: 20856810.
69. Gordon D, Abajian C, Green P (1998) *Consed*: a graphical tool for sequence finishing. Genome Res 8: 195-202. doi:10.1101/gr.8.3.195. PubMed: 9521923.
70. Ewing B, Green P (1998) Base-calling of automated sequencer traces using *phred*. II. Error Probabilities Genome Res 8: 186-194.
71. Sugita M, Sugiura M (1996) Regulation of gene expression in chloroplasts of higher plants. Plant Mol Biol 32: 315-326. doi:10.1007/BF00039388. PubMed: 8980485.
72. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM et al. (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 8: 174. doi:10.1186/1471-2164-8-174. PubMed: 17573971.
73. Gao L, Yi X, Yang YX, Su YJ, Wang T (2009) Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. BMC Evol Biol 9: 130. doi:10.1186/1471-2148-9-130. PubMed: 19519899.
74. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM et al. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-1047. doi:10.1093/bioinformatics/16.11.1046. PubMed: 11159318.
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079. doi:10.1093/bioinformatics/btp352. PubMed: 19505943.
76. Timme RE, Kuehl JV, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. Am J Bot 94: 302-312. doi:10.3732/ajb.94.3.302. PubMed: 21636403.
77. Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proc Natl Acad Sci U S A 89: 7722-7726. doi:10.1073/pnas.89.16.7722. PubMed: 1502190.
78. Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol 22: 1813-1822. doi:10.1093/molbev/msi173. PubMed: 15930156.
79. Maier RM, Neckermann K, Igloi GL, Kössel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251: 614-628. doi:10.1006/jmbi.1995.0460. PubMed: 7666415.
80. Wu FH, Chan MT, Liao DC, Hsu CT, Lee YW et al. (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol 10: 68. doi:10.1186/1471-2229-10-68. PubMed: 20398375.

81. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH et al. (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol 23: 279-291. PubMed: 16207935.
82. Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS et al. (2008) Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. J Mol Evol 66: 555-564. doi: 10.1007/s00239-008-9091-7. PubMed: 18463914.
83. Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV et al. (2007) Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). Mol Phylogenet Evol 45: 547-563. doi:10.1016/j.ympev.2007.06.004. PubMed: 17644003.
84. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28: 583-600. doi:10.1093/molbev/msq229. PubMed: 20805190.
85. Blazier J, Guisinger MM, Jansen RK (2011) Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol Biol 76: 263-272. doi:10.1007/s11103-011-9753-5. PubMed: 21327834.
86. Palmer JD, Osorio B, Aldrich J, Thompson WF (1987) Chloroplast DNA evolution among Legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr Genet 11: 275-286. doi: 10.1007/BF00355401.
87. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J et al. (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol Biol 59: 309-322. doi:10.1007/s11103-005-8882-0. PubMed: 16247559.
88. Stacey G (2008) Genetics and genomics of soybean. New York: Springer. xv, pp. 407.
89. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM et al. (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol 8: 36. doi:10.1186/1471-2148-8-36. PubMed: 18237435.
90. Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252: 195-206. doi:10.1007/BF02173220. PubMed: 8804393.
91. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ et al. (2006) The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23: 2175-2190. doi:10.1093/molbev/msl089. PubMed: 16916942.
92. Plunkett GM, Downie SR (2000) Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. Syst Bot 25: 648-667. doi:10.2307/2666726.
93. Shi C, Liu Y, Huang H, Xia EH, Zhang HB et al. (2013) Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf*15 function and evolution in angiosperms. PLOS ONE 8(3): e59620. doi:10.1371/journal.pone.0059620. PubMed: 23527231.
94. Saski C, Lee SB, Fjellheim S, Guda C, Jansen RK et al. (2007) Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. Theor Appl Genet 115: 571-590. doi: 10.1007/s00122-007-0567-4. PubMed: 17534593.
95. Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K (2004) Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: A comparative analysis of four monocot chloroplast genomes. DNA Res 11: 93-99. doi:10.1093/dnares/11.2.93. PubMed: 15449542.
96. Cavalier-Smith T (2002) Chloroplast evolution: secondary symbiogenesis and multiple losses. Curr Biol 12: R62-R64. doi: 10.1016/S0960-9822(01)00675-3. PubMed: 11818081.
97. Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear adh sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85: 1301-1315. doi:10.2307/2446640. PubMed: 21685016.
98. Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L et al. (2012) Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). PLOS ONE 7(5): e36869. doi:10.1371/journal.pone.0036869. PubMed: 22606302.
99. Wortley AH, Rudall PJ, Harris DJ, Scotland RW (2005) How much data are needed to resolve a difficult phylogeny? case study in Lamiales. Syst Biol 54: 697-709. doi:10.1080/10635150500221028. PubMed: 16195214.
100. Petersen G, Aagesen L, Seberg O, Larsen IH (2011) When is enough, enough in phylogenetics? A case in point from *Hordeum* (Poaceae). Cladistics 27: 428-446. doi:10.1111/j.1096-0031.2011.00347.x.
101. Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N et al. (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9: 328-333. doi:10.1111/j.1467-7652.2010.00558.x. PubMed: 20796245.
102. Gao L, Su YJ, Wang T (2010) Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J Syst Evol 48: 77-93. doi:10.1111/j.1759-6831.2010.00071.x.
103. Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ et al. (2011) Complete chloroplast genome sequence of Magnolia *kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54: 663-673. doi:10.1139/g11-026. PubMed: 21793699.