Complete chloroplast genome of *Campsis grandiflora* (Thunb.) Schum and systematical comparative analysis within the family *Bignoniaceae*.

Haimei Chen
institute of medicinal plant development
Zhuoer Chen
Chinese Academy of Medical Sciences
Qing Du
CAMS PUMC IMPLAD: Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development;qinghai nationalities university
Mei Jiang
Chinese Academy of Medical Sciences and Peking Union Medical College Hospital and Institute of Dermatology: Chinese Academy of Medical Sciences & Peking Union Medical College Hospital of Skin Diseases and Institute of Dermatology, Shandong Academy of Sciences
Chang Liu (cliu@implad.ac.cn)
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development https://orcid.org/0000-0003-3879-7302

Research Article

Keywords: Campsis grandiflora, chloroplast genome, gene rearrangement, SSR and repeat analysis, Phylogenetic analysis, Ka/Ks analysis

Posted Date: October 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-992863/v1

Abstract

Background The plants in the *Bignoniaceae* have a wide distribution in the tropics and large populations around the world. However, the research information of *Bignoniaceae* is still scarce or even blank. To excavating the research information of *Bignoniaceae* plants and provide data support for the study of plant plastid genomes.

Methods and results In this study, We have done a particular exploration of the chloroplast genome bioinformation of *Campsis grandiflora*. The chloroplast DNA of *C. grandiflora* was extracted, sequenced with HTS platform, assembled, and annotated with corresponding software. Results show that the complete chloroplast genome of *C. grandiflora* is 154,303 bp long, having a quadripartite structure with large single copy (LSC) of 85,064 bp and a small single copy (SSC) of 18,009 bp separated by inverted repeats (IRs) of 25,615 bp. A total of 110 genes in *C. grandiflora* is comprised of 79 protein-coding genes, 27 transfer RNA (tRNA) coding genes and 4 ribosomal RNA (rRNA) genes. The location and distribution of simple sequence repeats (SSRs) and long repeat sequences were determined. We finished the phylogenetic analysis based on homologous amino acid sequence among 45 species which derived from *Bignoniaceae*.

Conclusions The chloroplast genomes have been used for molecular markers, species identification and phylogenetic studies. The outcome strongly supported that *C. grandiflora* and genus *Incarvillea* formed a cluster within *Bignoniaceae*. This study identified the unique characteristics of the *C. grandiflora* cp genome, which will provide a theoretical basis for species identification and biological research.

Introduction

The chloroplast genome is an important role of plant plastid genetic system, and its highly conserved circular quadripartite double-stranded structure which consisted of a large single-copy region (LSC; 80–90 kb) and a small single-copy region (SSC; 16–27 kb), separated by two inverted repeat regions (IRs) of 20 to 28 kb. Leading its mutation rate at a low level in the process of plant evolution. Therefore, the character that stable gene content, simple structure, nonrecombinant, and mostly maternally inherited meaning that the chloroplast genomes contain a great deal of valuable biological information as an ideal material to support phylogeny and evolution studies[1]. With the rapid development of high-throughput sequencing technology in recent years, researchers have been able to efficiently extract and sequence chloroplast genomes from plants, thus greatly advancing the process of chloroplast genome sequencing. Chloroplast genome sequencing information has been widely used to build the basis of phylogenetic analysis, and the evolutionary history of many plant groups has been deeply explored and supported[2].

The abundance of species in *Bignoniaceae* includes a total of 650 species in 120 genera, including *Catalpa*, *Campsis*, *Adenomocalymma*, *Amphilophium*, and *Anemopaegma*, etc[3]. *Bignoniaceae* plants, which mainly for trees, shrubs, or woody vines, are widely distributed in the tropics and subtropics and constitute an important part of tropical plants. The vast majority of species of *Bignoniaceae* have dazzling large and beautiful flowers, as well as a variety of exotic fruit shapes, and are cultivated in botanical gardens around the world, as ornamental, scenic, and street trees, and as an ideal shade pergola plant for the tropics[4]. *Campsis grandiflora* is a climbing vine affiliated with the genus *Campsis*, family *Bignoniaceae*. Distinguished from *Campsis radicans*, other plants of the same genus that derived from North America, *C. grandiflora* is mainly distributed in China and Japan, cultivated in Vietnam, India, and Pakistan[5]. *Campsis grandiflora* can be used for ornamental and medicinal purposes. Pharmacological studies have shown that It has antibacterial, antithrombotic, and antitumor effects[6]. According to the Chinese Pharmacopoeia (2020 Edition)[7], *C. grandiflora* has the functions of promoting blood circulation, and its flower is a diuretic for meridional treatment and can also cure the disease of falling and injury [8].

Although there are numerous species in the family Bignoniaceae, only more than 40 chloroplast data have been recorded[9]. In particular, the chloroplast genome study of the entire genus *Campsis*, an important branch of *Bignoniaceae*, is still blank. In this study, we gained the chloroplast genomes of the *C. grandiflora* by using high-throughput sequencing technology and utilized more than 40 species uploaded chloroplast genomes of the *Bignoniaceae*, with the aim of 1. To explore the biodiversity and evolution process of the genus *Campsis*. 2. To characterize the gene contents and gene loss within the family *Bignoniaceae* by chloroplast genome assembly and annotation. 3. To get the phylogenetic information of *C. grandiflora* and making valid hypotheses about homology between different lineages of *Bignoniaceae*. 4.To explore the gene rearrangement structure that occurred in the family *Bignoniaceae*.

Materials And Methods

Plant material, DNA purification, and genome sequencing

The *Campsis grandiflora* sample was collected in Huazhong Medicinal Botanical Garden, China. located in 109.76 E,30.18 N, voucher sample IDs is implad201808016, IMPLAD, China. The whole-genome DNA of *Campsis grandiflora* was extracted using the plant genomic DNA kit (Tiangen Biotech, Beijing, China). The process of library construction and Genome sequence was completed by the Hiseq 2500 platform (Illumina, San Diego, CA, USA)[10].

Chloroplast genome assembly and annotation

The raw data of the sequence was assembled into a complete chloroplast genome with NOVOplasty(ver. 4.0.1) [11].

The genome annotation and repeat analysis works were finished by CPGAVAS2, DB 2[12].

Phylogenetic analysis

At present, the commonly accepted phylogenetic classification of *C. grandiflora* is that the *C. grandiflora* of genus *Campsis* of family *Bignoniaceae*. We used The maximum likelihood method[13] to construct an evolutionary tree with the cpREV model of IQ-Tree[14] for 56 common protein sequences of 45 species

which included genus *Adenocalymma*[15], *Neojobertia*[16], *Pleonotoma*[17], *Amphilophium*[18], *Anemopaegma, Tanaecium*[19], *Dolichandra*[20], *Oroxylum*[21], *Catalpa*[22, 23], *Incarvillea*[24-26], *Spathodea*[27] and 2 outer groups (*Paulownia tomentosa*[28] and *Arabidopsis thaliana*[29]) of species from the family *Bignoniaceae*. For phylogenic tree building, we used Phylosuite (version 1.2.2)[30] to extract the GenBank files of 47 species to get the common protein-coding genes information. Then we did the multiple sequence alignment by using MAFFT(v7.313) without the duplicated FASTA files. The protein-coding gene MAFFT outcome was concatenated by Gblocks(v0.91b), to select conserved blocks from multiple alignments for use in phylogenetic analysis.

After we got the contree file, the visual work of the evolutionary tree was performed by iTOL (Interactive Tree of Life) [31].

SSR and repeat analysis

The SSR locus and distribution was identified with MISA (MIcroSAtellite identification tool)[32]. The long tandem repeats (matching parameter = 2, mismatching and indel parameter = 7, minimun identity score = 50, maximun repeat period = 500, minimum repeat size = 30bp, repeat unit similarity >= 90%) identified with TRF (Tandem Repeats Finder)[33]. The long interspersed repeats (repetition length >= 30bp, Hamming distance = 3) identified with VMATCH (The Vmatch large scale sequence analysis software)[34].

Synteny analysis

In this study, we compared each single 45 *Bignoniaceae* species in the phylogenic tree with *A. thaliana* to perform gene scale dot-plot analysis with Gepard (ver. 1.40 final.)[35].

The detailed synteny analyses of 12 species [These species are *A. oligoneuron* (NC_037232.1)[36], *A. gnaphalanthum* (NC_042903.1), *T. tetragonolobum* (NC_027955.1), *A. paniculatum* (NC_042918.1), *I. compacta* (NC_050666.1), *I. sinensis* (NC_051523.1), *N. candolleana* (NC_036503.1), *A. allamandiflorum* (NC_036494.1), *A. biternatum* (NC_036496.1), *A. divaricatum* (NC_037456.1), *A. marginatum* (NC_037457.1), *C. grandiflora* (MW430049).] with genomic structure rearrangement was revealed by Easyfig (ver. win2.1) [37]. We first got the B.O format files with blastn of particular 12 species, the contrasting species is *A. thaliana*. Then import the B.O files and GB or GBK files of related species into the Easyfig software. The synteny visualizes outcome will be generated.

Junction sites visualize analysis

We used the GenBank files of 12 representative species with genomic structural variations from 45 species of *Bignoniaceae* that were used for phylogenetic analysis to get the gene distribution on LSC, SSC, IRa, IRb border. The location of genes on the boundaries was visualized by IRSCOPE[38]

Ka/Ks analysis

We used the aBSREL model of Hyphy Vision software to contribute the selective pressure analysis[39] among 45 species in *Bignoniaceae*. We first acquired the corresponding chloroplast genome GB files and FASTA files according to the accession number in NCBI. Then got 63 clusters of orthologous genes among these species to used to calculate the Ka/Ks. The outcome was listed in aBSREL.json format. In this study, we selected genes with the p-value < 0.05. The detailed information was shown in the web version of aBSREL.

Results

Genome organization and compositions

The chloroplast genome sequence (GenBank accession no.: MW430049) of *Campsis grandiflora* was a typical circular DNA molecule with a total length of 154,303 bp. It has a conservative tetrad structure consisting of an LSC region, an SSC region, and a pair of IR regions, with lengths of 85,064 bp, 18,009 bp, and 25,615 bp, respectively (**Figure 1**). The G/C content of the chloroplast genome of *Campsis grandiflora* was 38.09 %. The G/C content in the IR region (43.17%) was higher than that in the SSC region (32.74%) and LSC region (36.16%).

Gene Content

The chloroplast genome of *Campsis grandiflora* encodes a total of 110 unique genes, including 79 protein-coding genes, 27 transfer RNA (tRNA) coding genes, and 4 ribosome RNA (rRNA) coding genes (**Table 1**). Among them, eight protein coding genes (*rps*12, *ndh*B, *rpl*2, *rpl*23, *rps*7, *ycf*1, *ycf*2, *ycf*15), 7 tRNA coding genes (*trn*A-UGC, *trn*E-UUC, *trn*L-CAA, *trn*M-CAU, *trn*N-GUU, *trn*R-ACG, *trn*V-GAC) and 4 rRNA coding genes (*rrn*16S, *rrn*23S, *rrn*5S, *rrn*4.5S) were located in IR region. 12 protein-coding genes (*rps*16, *atp*F, *rpo*C1, *pet*B, *pet*D, *rpl*16, *rpl*2 (+), *rpl*2(-), *ndh*B(+), *ndh*B(-), *ndh*A) contain one intron, and two protein-coding genes (*ycf*3, *acc*D) contain two introns. 8 tRNA coding genes (*trn*K-UUU, *trn*S-CGA, *trn*L-UAA, *trn*V-UAC, *trn*E-UUC(-), *trn*E-UUC(+), *trn*A-UGC(-), *trn*A-UGC(+)) contain one intron (**Table S1**). In the process of counting intron and exon of gene, we found the *clp*P gene has lost all of its introns and exons.

The coding sequence (CDS) in the Chloroplast genome of *C. grandiflora* was 79,170 bp, accounting for 51.31% of the total genome length. The length of the rRNA gene was 9,388 bp, accounting for 6.08% of the whole genome length. The length of the tRNA gene was 2,811 bp, accounting for 1.82% of the whole genome length. The non-coding region of the *Campsis grandiflora* chloroplast genome mainly includes introns and gene spacers, whose length accounts for 40.79% of the whole genome length.

SSR and repeat sequences analysis

The repeat sequences are particular nucleic characteristic sequences repeat units that have multiple copies in the genome. On the one hand, these repeats might play a significant role in the evolution of the chloroplast genome. On the other hand, they can also be used for species identification and molecular breeding as molecular markers. The repeat sequences are respectively classified into simple sequence repeat (SSR), long tandem repeats, and long interspersed repeated sequence 3 forms according to their length and correlation[1].

The SSR is also named microsatellite sequence. It is a piece of DNA that consists of multiple duplicate basic repeat units made of 1-6 nucleotides. The SSR is widespread all around the different places of the gene. Their length usually below 200bp. In the chloroplast genome of *C. grandiflora*, the SSR is mainly A/T type, with 54. The second and third types are C/G and AT/AT, with 3 and 2 respectively (**Table S2**). Besides, we analyzed and listed the quantity, type, size, and locus of SSRs in the chloroplast genome of *C. grandiflora*. In total, there were 51 SSRs identified in the *C. grandiflora* chloroplast genome. These SSRs are mainly composed of mononucleotide repeat units and dinucleotide repeat units (**Table S3**). Beyond that, we found no other forms like tri-, tetra-, penta-, hexa-nucleotide repeat units. Among 51 SSRs we found, most of them appear in intergenic spacers (35 SSRs), 9 SSRs located in the coding sequence, and 7 SSRs situated in the non-coding region of particular genes.

The long tandem repeats refer to the repeated repetition of a sequence on a chromosome. A total of 40 tandem repeats have been found, satisfying the two conditions that the total length is over 20bp and the similarity between repeating units is greater than or equal to 90% (**Table S4**). We also listed the related property in the table. Among the long tandem repeats, more than half (22) repeats were located in IGS, and 13 repeats were shown in the CDS, the remainder 5 repeats were located in the non-coding region.

Interspersed repeats are another kind of repeated sequence which are different from tandem repeats. It include palindromic repeats and direct repeats. With the e-value less than 1E-4 as the threshold, the scattered repeats of plumbic chloroplast genomes included 49 direct repeats. It's worth mentioning that all of the interspersed repeats of *C. grandiflora* chloroplast genome are D type (direct repeat sequence), and these interspersed repeats are all in the range of 62,500-63,700 of *acc*D gene and almost all of them are located in the non-coding region except one sequence that its repeat unit I in the CDS of *acc*D (**Table 2**).

Phylogenetic analysis

To get the phylogenetic information of *C. grandiflora* and making valid hypotheses about homology between different lineages of *Bignoniaceae*. We used 45 *Bignoniaceae* species and 2 outgroup species chloroplast genomes to build the phylogenetic tree of *Bignoniaceae*. (**Figure 2**)

The tree shows that that two primary branches initially diverged from the tree root There are 15 species from the genus *Adenocalymma*, *Neojobertia* and *Pleonotoma* gathered into a branch on top of the tree. 11 species of the genus *Amphilophium* converged into a branch. 8 species of the genus *Anemopaegma* got into a branch. Then the genus *Amphilophium*, *Anemopaegma*, *Tanaecium*, and *Dolichandra* gathered into a big branch with *Adenocalymma*, *Neojobertia*, and *Pleonotoma*. Furthermore, the grand branch congregated a branch with genus *Oroxylum*, and then the genus *Spathodea*. At the bottom of the tree, 2 species of genus *Catalpa* gathered to a branch. From this view, the eight genera mentioned above have contributed to the upper grand branch of the evolutionary tree of the family *Bignoniaceae*. In the remaining part of the tree, 3 species of the genus *Incarvillea* gathered into a branch, then the sole branch *Tecomaria* have aggregated a branch with the genus *Incarvillea*. At last, the genus *Campsis*, *Incarvillea*, *Tecomaria* have converged into another grand branch of the tree.

In the phylogenetic tree of the family *Bignoniaceae*, the bootstrap scores of all branches of the evolutionary tree were high ($\geq$ 47%), indicating that the evolutionary tree has high reliability. The results of the phylogenetic analysis are consistent.

Synteny analysis.

The initial study of *C. grandiflora* chloroplast genome rearrangement structure started from synteny analysis[40]. We first performed a genome comparing dot plot by using Gepard (ver. 1.40 final). The visualization result shows that the rearrangement was occurred at about 48 772-73 286bp in the *C. grandiflora* chloroplast genome (**Figure S1**).

Comparative analysis of gene loss in family Bignoniaceae.

This study explores whether there is a correlation between gene loss and the rearrangement of genome structure. We made detailed statistics of the protein-coding gene loss in the particular plants of *Bignoniaceae* (**Table 3**). All the plants involved in the statistics are derived from phylogenetic trees (**Figure 2** with supplement **Figure 4**). The genes listed in the table are all existed quantity varies. Oppositely, genes that did not differ in number were not listed. The result of the statistic shows: The number of genes in 8 species from the genus *Anemopaegma* was highly conserved and consistent. In terms of gene loss, the *acc*D gene has been lost in the genus *Incarvillea*. The *clp*P gene was found lost in *I. arguta* , *T. tetragonolobum* but also found incomplete structure in *C. grandiflora*. The *ndh*D gene was found lost in *T. tetragonolobum* and *I. arguta*. The *pet*B, *rpl*16, *rpo*A gene has lost in *T. tetragonolobum*, *I. arguta*. The *I. sinensis* was also found to have lost the *pet*B and *rpl*16 gene. The *rpl*32 gene was found only missing in *C. grandiflora*. The *rps*16 was found lost in *T. tetragonolobum*, *I. compacta*, *I. arguta*. The *rps*19 gene was found only lost in *I. sinensis*. The *rps*4 gene has lost in *T. tetragonolobum*. The *ycf*4 gene has been lost in *A. peregrinum* and *A. biternatum*. And the *ycf*15 gene was only found in *T. tetragonolobum*, *I. arguta*, and *C. grandiflora*. The *ycf*1 gene was found only lost in *A. gnaphalanthum*. In general, the majority of gene loss occurs in the genus *Incarvillea and Tanaecium*.

Ka/Ks Selective pressure analysis

In genetics, Ka/Ks or dN/dS represents the ratio between non-synonymous replacement (Ka) and synonymous replacement (Ks). This ratio can be used to determine whether there is selective pressure acting on the protein-coding gene[41].

Nucleotide variations that do not lead to amino acid changes are called synonymous mutations, whereas non-synonymous mutations occur. It is generally believed that synonymous mutations are not subject to natural selection, while non-synonymous mutations are. In evolutionary analysis, it makes sense to understand the rate at which synonymous and non-synonymous mutations occur[41].

In this study, we used the phylogenetic tree (**Figure 2**) as species reference, and we utilized the aBSREL (adaptive branch-site random effects likelihood) model of software Hyphy to carry the selection pressure analysis of protein-coding genes (**Table S5**). A total of 6 *Bignoniaceae* genes were positively selected: *ndh*G, *rbc*L, *rpl*22, *rpl*23, *rps*12, *rps*15. In species *A. bracteatum*, the *ndh*G gene is positively selected. In species *A. glaucum*, *A. divaricatum*, the *rbc*L gene was positively selected. The *rpl*22 gene was positively selected in species *A. steyermarkii*, *D. cynanchoides*. In species *A. allamandiflorum*, *A. chamberlaynii*, *rpl*23 gene was positively selected. In *C. ovata*, *rps*12 and *rps*15 are positively selected. In species *C. grandiflora*, *rps*15 is positively selected.

Junction sites visualize analysis

To unravel the gene distribution of junction sites border, and compare the distinction between *C. grandiflora* and other species which have genome rearrangement structure in the family *Bignoniaceae*. We have visualized the gene distribution with IRSCOPE (**Figure S5**).

In the result of the visualizing analysis, we can see the complete genome was divide into 5 parts with 4 vertical bars. The 5 parts are respective LSC, IRb, SSC, IRa, LSC. In the 4 species from the genus *Adenocalymma*, the *rps*15 gene has crossed the JSA between SSC and IRa, and the gene distribution in genome boundary is highly conservative. The bp number of each gene from the boundary or across the boundary is highly consistent or similar. In the *I. sinensis*, the *ndh*F was found to cross the JSB between IRb and SSC region. In the *I. compacta*, the gene located in the boundary between IRb/IRa and SSC is *trn*N. It's worth mentioning that, there are significant differences in the length of SSC and IR regions between these two species from the genus *Incarvillea*. The SSC region in the *I. sinensis* was only 8,666 bp, and the IR regions were 35,394 bp respectively. But in the *I. compacta*, the SSC region has reached 21,925 bp. This difference in the length of genomic regions has also occurred in *A. paniculatum* and *A. oligoneuron*. Their IR regions have reached 37,372 bp and 39,614 bp respectively, much longer than the normal length of the IR region. Accordingly, in the *A. paniculatum*, the gene that crossed the IRb and LSC are *pet*D, and in the *A. oligoneuron*, the counterpart gene is *pet*B.

Discussion

In the current study, we extracted and sequenced the chloroplast genome of Campsis grandiflora. The raw data were assembled and annotated with relevant tools, and the complete information of the transiting chloroplast genome was obtained. Furthermore, the phylogenetic analysis of Campsis grandiflora was performed. Otherwise, we found the gene rearrangement structure in the genome of Campsis grandiflora after we used the tool (Gepard, ver. 1.40 final.) to compare the synteny of chloroplast genome sequences between Campsis grandiflora and Arabidopsis thaliana (Figure S1). It could provide us a new direction of chloroplast genome research of Campsis grandiflora. Special distribution of interspersed repeated sequences in accD gene In the process of statistical analysis of repeated sequences, we found the particularity of interspersed sequences. Compared with other species in this family, the interspersed sequences in C. grandiflora chloroplast genome showed obvious centralization and uniformity. The results showed that, except for one sequence located in the coding region of accD gene, all the remaining sequences were distributed in the non-coding region of accD gene. And the distribution range is concentrated in 62000bp-64000bp. In addition, the types of repeated sequences are only direct sequences, and palindrome sequences are not found (Table 2). The accD gene, full name acetyl-CoA carboxylase gene, is present in plastids such as chloroplasts in most flowering plants, including non-photosynthetic parasites. Its function is to encode the β -carboxylase subunit of acetyl-CoA carboxylase, thereby participating in plant life activities and material metabolism. Previous studies on tobacco have shown that if the accD gene is knocked out or destroyed and cannot be successfully expressed in plastids, the leaf development of the plant will be severely affected. For example, the loss of tissue cells leads to the stagnation of leaf division and differentiation, which leads to the failure of photosynthesis and the death of plants. This indicates that accD gene is an indispensable and important gene in plants. In this study,the special distribution of interspersed sequences raised the possibility of molecular markers for the unique sequence in the gene coding region, and at the same time, through the statistics and analysis of the location of different repeat sequence families in different genes, new interspecies relationships or evolutionary processes can be found. These new directions are expected to be realized in future research. Phylogenetic tree From the distribution of species displayed in the phylogenetic tree, the genus Adenocalymma has a distant genetic relationship with the genus Campsis. In contrast, the genus Incarvillea, Tecomaria, and Catalpa have a more close genetic relationship with the genus Campsis. And because of the Campsis grandiflora located in the base of the whole tree, we reckon that the divergence event occered in an earlier period of the evolution process in Bignoniaceae. Junction sites visualize The results showed that the location and species of boundary genes were different with the length of genome sequence (Figure S4). From this phenomenon, we can deduce that variation in the length of genomic regions leads to differences in the genes located at the boundaries. In the C. grandiflora and T. tetragonolobum, the location of the ycf1 gene was all at the JSB and JSA. Whereas in the C. grandiflora, the rps19 was located at the LSC region but crossed the JLB in the T. tetragonolobum. The chloroplast genome rearrangement occurred in Campsis grandiflora The initial study of C. grandiflora chloroplast genome rearrangement structure started from synteny analysis[40]. After using different software, we focused on the analysis of the inverted chloroplast genome structure of C. grandiflora. We amplified and detailed the results of EasyFig synteny analysis of the rearranged region, and it was clearly and directly to see that in the 48772 bp-73286 bp, both gene type and location were reversed, indicating a typical local genome sequence reversal. (Figure 3 with supplement Figure S1). The systematic analyze of genome rearrangement occurred in Bignoniaceae To verify whether other species in the Bignoniaceae have occurred the genome rearrangement. We then analyzed other 44 species from the phylogenetic tree with Gepard (ver. 1.40 final). Finally, we identified 13 genomic structures in chloroplast genomes from 45 species

of Bignoniaceae, including 12 different types of chloroplast genome rearrangement structures and 1 conventional structure without variation. We used EasyFig to visualize these 12 rearrangement structures (Figure 4 with supplement Figure S2 and Figure S3). In summary, 4 species of genus Adenocalymma, 2 species of Amphilophium, 2 species of genus Incarvillea, 8 species of genus Anemopaegma (In fact, eight species of the genus Anemopaegma share the same genomic rearrangement structure[42], and for ease of comparison, we used only one of the eight species randomly to participate in the comparative analysis), and Neojobetia candolleana (NC_036503.1), Tanaecium tetragonolobum (NC_027955.1), Campsis grandiflora (MW430049) have occurred different genome structure rearrangement. Combined with the above-mentioned statistical results of gene loss and the results of synteny analysis (Figure 4 and Table 3), it is not difficult to see that: In the genus Anemopaegma, 8 species have the same genome structure and keep highly conservative gene number. We can think of it as an intergeneric characteristic of the genus Anemopaegma. However, among species from the genus Incarvillea with a large variety in cp genomic length and genomic region length within the genus, the number of gene loss was also significantly different. Compared with I. arguta, the same genus without rearrangement of genome structure, we found that the genes in the LSC region of I. compacta and I. sinensis such as clpP, ndhD, petB, rpl16, and rpoA, were not found in I. arguta cp genome. In short, genome rearranged species have more genes than non-rearranged species. Thus, variation in gene structure indeed affects the category and number of genes. But whether the relationship holds for other species remains unknown.

Conclusions

In this study, we extracted, assembled, sequenced, and annotated the complete chloroplast genome of *C. grandiflora*, filling in the gaps in chloroplast genome information of genus *Campsis*. The phylogenetic analysis not only reveals the phylogenetic information of *Bignoniaceae* but also shows the overall evolutionary history of 45 species of the family. And the repeat sequence analysis provided us the genetic characteristic information. The Ka/Ks analysis indicated the direction of evolution of *Bignoniaceae*. We conducted a detailed and in-depth analysis of the chloroplast genome of *C. grandiflora*, and found that the chloroplast genome has an inverted rearrangement structure through synteny analysis, We also found and sorted out the rearrangement structures of 12 chloroplast genomes of *Bignoniaceae* from the available data by synteny analysis. which was of great significance to the phylogenetic information of *C. grandiflora*. Then the gene loss analysis inspired us to the relationship between rearrangement structure and the gene quantity variation.

There are a lot of species in the *Bignoniaceae*, but only a drop in the bucket of information that is currently available. The results of this study are based on all the released chloroplast genome sequences available so far. Please correct any irregularities or gaps. With the acceleration of sequencing progress, the database of *Bignoniaceae* will be enriched day by day in the future, and there will be more discoveries, new features, and breakthroughs.

Declarations

Author Contribution

CL conceived the study; MJ collected samples of *C. grandiflora*, extracted DNA for next-generation sequencing, assembled and validated the genome; ZEC performed data analysis and drafted the manuscript; HMC and QD reviewed the manuscript critically. All authors have read and agreed the contents of the manuscript.

Funding

This work was supported by the National Science &Technology Fundamental Resources Investigation Program of China [2018FY100705], National Science Foundation Funds [81872966],Chinese Academy of Medical Sciences, Innovation Funds for Medical Sciences (CIFMS) [2017-I2M-1-013], Qinghai Provincial Key Laboratory of Phytochemistry of Qinghai Tibet Plateau [2020-ZJ-Y20]. The funders were not involved in the study design, data collection, analysis, decision to publish, or manuscript preparation.

Compliance with ethical standards

Conflict of interest All the authors declare no conflicts of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Data availability statement

The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at (https://www.ncbi.nlm.nih.gov/) under the accession no.MW430049. The associated BioProject, SRA, and Bio-Sample numbers are PRJNA704532, and SAMN18043523 respectively.

ORCID

Chang Liu: 0000-0003-3879-7302

Qing Du: 0000-0002-0732-3377

Mei Jiang: 0000-0002-8266-7233

Zhuoer Chen: 0000-0001-7782-6992

References

1. Liu C, Huang L-F (2020) **Chloroplast genomic mapping of Chinese medicinal plants [M]**. Science Press: Beijing, China. **Vol.1, pp 3-4**.

2. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) **Comparison of next-generation sequencing systems**. *J Biomed Biotechnol*, **2012**:251364. https://doi.org/10.1155/2012/251364

3. Lohmann, Lúcia G (2006) **Untangling the phylogeny of neotropical lianas (Bignonieae, Bignoniaceae)**. *Am J Bot*, **93**(2):304-318. https://doi.org/10.3732/ajb.93.2.304

4. Flora of China (2021) **Bignoniaceae**. IBCAS Publishing iPlant.**http://www.iplant.cn/info/Bignoniaceae?t=z**.Accessed 07 Jan 2021

5. Flora of China (2021) **Campsis grandiflora**. IBCAS Publishing iPlant.**http://www.iplant.cn/info/Campsis%20grandiflora?t=foc**. Accessed 07 Jan 2021

6. Cui XY, Kim JH, Zhao X, Chen BQ, Lee BC, Pyo HB, Yun YP, Zhang YH (2006) **Antioxidative and acute anti-inflammatory effects of Campsis grandiflora flower**. *J Ethnopharmacol*, **103**(2):223-228. 10. https://doi.org/1016/j.jep.2005.08.007

7. National Pharmacopoeia Commission (2015) **The Pharmacopoeia of the People's Republic of China (2015 edition) [S]**. Medical Technology Publishing House: Beijing, China, vol.1, pp 299

8. Xiao P-g, Zhao Z-z (2018) **Encyclopedia of Medicinal Plants[M]**. *World Book Inc*: Beijing, China, pp **11-14**.

9. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B*et al* (2020) **NCBI Taxonomy: a comprehensive update on curation, resources and tools**. *Database : the journal of biological databases and curation*. **2020**. https://doi.org/10.1093/database/baaa062

10. Steemers FJ, Gunderson KL (2005) **Illumina, Inc**. *Pharmacogenomics*, **6**(7):777-782. https://doi.org/10.2217/14622416.6.7.777

11. Dierckxsens N, Mardulyn P, Smits G (2017) **NOVOPlasty: de novo assembly of organelle genomes from whole genome data**. *Nucleic Acids Res*, **45**(4):e18. https://doi.org/10.1093/nar/gkw955

12. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C (2019) **CPGAVAS2, an integrated plastome sequence annotator and analyzer**. *Nucleic Acids Res*, **47**(W1):W65-W73. https://doi.org/10.1093/nar/gkz345

13. Hasegawa M, Kishino H, Saitou N (1991) **On the maximum likelihood method in molecular phylogenetics**. *J Mol Evol*, **32**(5):443-445. https://doi.org/10.1007/BF02101285

14. Nguyen, L. T, Schmidt H. A, Von Haeseler A, Minh, B. Q (2015) **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol*, **32**(1):268-274. https://doi.org/10.1093/molbev/msu300

15. Fonseca LHM, Lohmann LG (2017) **Plastome Rearrangements in the "Adenocalymma-Neojobertia" Clade (Bignonieae, Bignoniaceae) and Its Phylogenetic Implications**.*Front Plant Sci*. **8**:1875. https://doi.org/10.3389/fpls.2017.01875

16. Fonseca LHM, Lohmann LG (2018) **Combining high-throughput sequencing and targeted loci data to infer the phylogeny of the "Adenocalymma-Neojobertia" clade (Bignonieae, Bignoniaceae)**. *Molecular phylogenetics and evolution*. **123**:1-15. https://doi.org/10.1016/j.ympev.2018.01.023

17. Fonseca LHM, Lohmann LG (2019) **An updated synopsis of Adenocalymma (Bignonieae, Bignoniaceae): new combinations, synonyms, and lectotypifications**. *Systematic Botany*. **44**(4):893-912. https://doi.org/10.1600/036364419X15710776741341

18. Thode VA, Lohmann LG (2019) **Comparative Chloroplast Genomics at Low Taxonomic Levels: A Case Study Using Amphilophium (Bignonieae, Bignoniaceae)**. *Frontiers in plant science*. **10**:796. https://doi.org/10.3389/fpls.2019.00796

19. Nazareno AG, Carlsen M, Lohmann LG (2015) **Complete Chloroplast Genome of Tanaecium tetragonolobum: The First Bignoniaceae Plastome**. *PloS one*. **10**(6):e0129930. https://doi.org/10.1371/journal.pone.0129930

20. Fonseca LH, Cabral SM, Agra Mde F, Lohmann LG (2015) **Taxonomic updates in Dolichandra Cham. (Bignonieae, Bignoniaceae)**. *PhytoKeys*.(46):35-43. https://doi.org/10.3897/phytokeys.46.8421

21. Jiang Y, Wang J, Qian J, Xu L, Duan BZ (2020) **The complete chloroplast genome sequence of Oroxylum indicum (L.) Kurz (Bignoniaceae) and its phylogenetic analysis**. *Mitochondrial DNA Part B*. **5**(2):1429-1430. https://doi.org/10.1080/23802359.2020.1736961

22. Ma Q-g, Zhang J-g, Zhang J-p (2020) **The complete chloroplast genome of Catalpa ovata G. Don.(Bignoniaceae)**. *Mitochondrial DNA Part B*. **5**(2):1800-1801. https://doi.org/10.1080/23802359.2020.1750979

23. Yang J, Wang S, Huang Z, Guo P (2020): **The complete chloroplast genome sequence of Catalpa bungei (Bignoniaceae): a high-quality timber species from China**. *Mitochondrial DNA Part B*. **5**(4):3854-3855. https://doi.org/10.1080/23802359.2020.1841581

24. Wu X, Li H, Chen St (2021) **Characterization of the chloroplast genome and its inference on the phylogenetic position of Incarvillea sinensis Lam. (Bignoniaceae)**. *Mitochondrial DNA Part B*.**6**(1):263-264. https://doi.org/10.1080/23802359.2020.1860722

25. Ma G-T, Yang J-G, Zhang Y-F, Guan T-X (2019) **Characterization of the complete chloroplast genome of Incarvillea arguta (Bignoniaceae)**. *Mitochondrial DNA Part B*.**4**(1):1603-1604. https://doi.org/10.1080/23802359.2019.1601529

26. Wu X, Peng C, Li Z, Chen S (2019) **The complete plastome genome of Incarvillea compacta (Bignoniaceae), an alpine herb endemic to China**. *Mitochondrial DNA Part B*.**4**(2):3786-3787. https://doi.org/10.1080/23802359.2019.1681916

27. Wang Y, Yuan X, Li Y, Zhang J (2019) **The complete chloroplast genome sequence of Spathodea campanulata**. *Mitochondrial DNA Part B*.**4**(2):3469-3470. https://doi.org/10.1080/23802359.2019.1674710

28. Yi D-K, Kim K-J (2016) **Two complete chloroplast genome sequences of genus Paulownia (Paulowniaceae): Paulownia coreana and P. tomentosa**. *Mitochondrial DNA Part B*.**1**(1):627-629. https://doi.org/10.1080/23802359.2016.1214546

29. Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata Satoshi (1999) **Complete structure of the chloroplast genome of Arabidopsis thaliana**. *DNA Research* **6**(5):283-290. https://doi.org/10.1093/dnares/6.5.283

30. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX, Wang GT (2020) **PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies**. *Mol Ecol Resour.* **20**(1):348-355. https://doi.org/10.1111/1755-0998.13096

31. Letunic I, Bork P (2021) **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. Nucleic acids research. **49(W1):**W293-W296. https://doi.org/10.1093/nar/gkab301

32. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) **MISA-web: a web server for microsatellite prediction**. *Bioinformatics*. **33**(16):2583-2585. https://doi.org/10.1093/bioinformatics/btx198

33. Benson G (1999) **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic acids research.* **27**(2):573-580. https://doi.org/10.1093/nar/27.2.573

34. Gojman B, DeHon A (2009) **VMATCH: Using logical variation to counteract physical variation in bottom-up, nanoscale systems**. In: *2009 International Conference on Field-Programmable Technology*. IEEE. 2009: 78-87. https://doi.org/10.1109/FPT.2009.5377684

35. Krumsiek J, Arnold R, Rattei T (2007) **Gepard: a rapid and sensitive tool for creating dotplots on genome scale**. *Bioinformatics*. **23**(8):1026-1028. https://doi.org/10.1093/bioinformatics/btm039

36. Firetti F, Zuntini AR, Gaiarsa JW, Oliveira RS, Lohmann LG, Van Sluys MA (2017) **Complete chloroplast genome sequences contribute to plant species delimitation: A case study of the Anemopaegma species complex**. *American Journal of Botany*. **104**(10):1493-1509. https://doi.org/10.3732/ajb.1700302

37. Sullivan MJ, Petty NK, Beatson SA (2011) **Easyfig: a genome comparison visualizer**. *Bioinformatics*. **27**(7):1009-1010. https://doi.org/10.1093/bioinformatics/btr039

38. Amiryousefi A, Hyvönen J, Poczai P (2018) **IRscope: an online program to visualize the junction sites of chloroplast genomes**. *Bioinformatics*. **34**(17):3030-3031. https://doi.org/10.1093/bioinformatics/bty220

39. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL (2015) **Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection**. *Molecular biology and evolution*. **32**(5):1342-1353. https://doi.org/10.1093/molbev/msv022

40. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) **Synteny and collinearity in plant genomes**. *SCIENCE*.**320**(5875):486-488. https://doi.org/10.1126/science.1153917

41. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J (2006) **KaKs_Calculator: calculating Ka and Ks through model selection and model averaging**. *Genomics, Proteomics & Bioinformatics*.**4**(4):259-263. https://doi.org/10.1016/S1672-0229(07)60007-2

42. Firetti F, Zuntini AR, Gaiarsa JW, Oliveira RS, Lohmann LG, Van Sluys MA: **Complete chloroplast genome sequences contribute to plant species delimitation: A case study of the Anemopaegma species complex**. *Am J Bot*2017, **104**(10): 1493-1509. https://doi.org/10.3732/ajb.1700302

Tables

Table 1. Gene contents in the chloroplast of *Campsis grandiflora*.

Category of genes	Group of genes	Name of genes
rRNA	rRNA genes	*rrn*16S(×2), *rrn*23S(×2), *rrn*5S(×2), *rrn*4.5S(×2)
tRNA	tRNA genes	27 trn genes
Self-replication	Small subunit of ribosome	*rps*11, *rps*12(×2), *rps*14, *rps*15, *rps*16, *rps*18, *rps*19, *rps*2, *rps*3, *rps*4, *rps*7(×2), *rps*8
	Large subunit of ribosome	*rpl*14, *rpl*16, *rpl*2(×2), *rpl*20, *rpl*22, *rpl*23(×2), *rpl*33, *rpl*36
	DNA dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1, *rpo*C2
Genes for photo-synthesis	Subunits of NADH-dehydrogenase	*ndh*A, *ndh*B(×2), *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of photosystem Ⅰ	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunits of photosystem Ⅱ	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*I, *psb*J, *psb*K *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z, *ycf*3
	Subunits of cytochrome b/f complex	*pet*A, *pet*B, *pet*D, *pet*G, *pet*L, *pet*N
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I
	Large subunit of rubisco	*rbc*L
Other genes	Maturase	*mat*K
	Translational initiation factor	*inf*A
	Envelope membrane protein	*cem*A
	Protease	Ψ*clp*P
	Subunit of Acetyl-CoA-carboxylase	*acc*D
	c-type cytochrom synthesis gene	*ccs*A
Unkown	**Conserved open reading frames**	*ycf*1(×2), *ycf*2(×2), *ycf*4, *ycf*15(×2)

Table 3. Gene loss of protein-coding genes in the chloroplast genomes of the structurally variant species in family Bignoniaceae. [a] genome rearrangement structure type. [b] normal structure without rearrangement. a: gene quantity is 1. b: gene located in IR region. c: the gene is not found in this species.

type[a]	Name of species	*acc*D	*clp*P	*inf*A	*ndh*D	*pet*B	*pet*D	*rpl*14	*rpl*16	*rpl*20	*rpl*22	*rpl*32	*rpl*36	*rpo*A	*rps*11	*rps*12	*rps*15	*rps*16	*rps*19	*rps*3	*rps*4
type1	*Anemopaegma. oligoneuron*	1[a]	1	2[b]	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. acutifolium	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. album	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. arvense	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. chamberlaynii	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. foetidum	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. glaucum	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
	Anemopaegma. prostratum	1	1	2	1	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
type2	*Amphilophium. gnaphalanthum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1
type3	*Amphilophium. paniculatum*	1	1	2	1	1	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
N[b]	*Amphilophium. dusenianum*	1	1	2	1	1	2	2	2	1	2	1	2	2	2	2	2	1	2	2	1
type4	*Tanaecium. tetragonolobum*	0	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0
type5	*Incarvillea. compacta*	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
type6	*Incarvillea. sinensis*	0	1	1	1	0	1	1	0	1	1	1	1	1	1	2	2	1	0	1	1
N	*Incarvillea. arguta*	0	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	0	2	1	0
type7	*Neojobertia. candolleana*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
type8	*Adenocalymma. allamandiflorum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
type9	*Adenocalymma. biternatum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
type10	*Adenocalymma. divaricatum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1
type11	*Adenocalymma. marginatum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
N	*Adenocalymma. peregrinum*	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
type12	*Campsis. grandiflora*	1	Ψ1	1	1	1	1	1	1	1	1	0	1	1	1	2	1	1	1	1	1
N	*Arabidopsis. thaliana*	1	1	0	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1

Due to technical limitations, Table 2 is only available as a download in the Supplemental Files section.

Supplementary

FigureS5 is not available with this version.

Figures

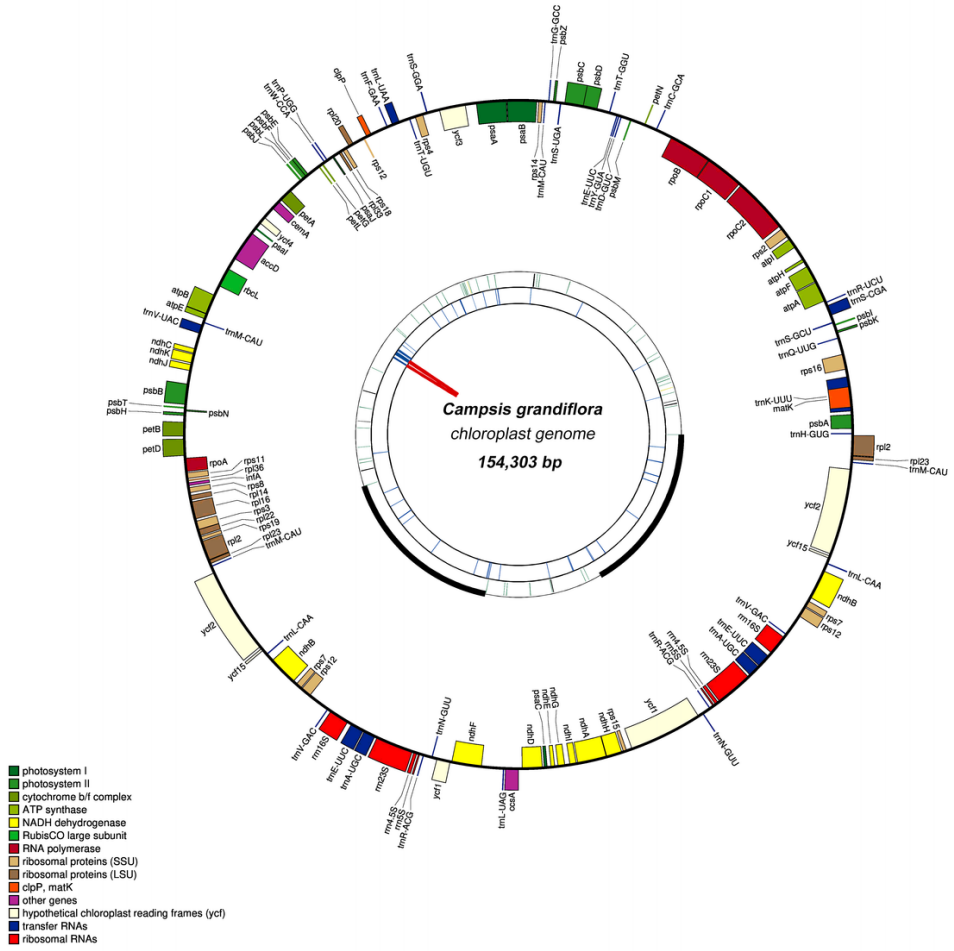

Figure 1

Map of the chloroplast genome of Campsis grandiflora. There are four rings in the figure: from the center outwards, the red and green arcs in the first circle represent the forward and reverse repeating sequence respectively; The short bars in the second circle represent tandem repeats; The short bar in the third circle represents the microsatellite repetition sequence; The fourth circle is the genetic structure and location map of the chloroplast genome. Genes with different functions are shown in different colors

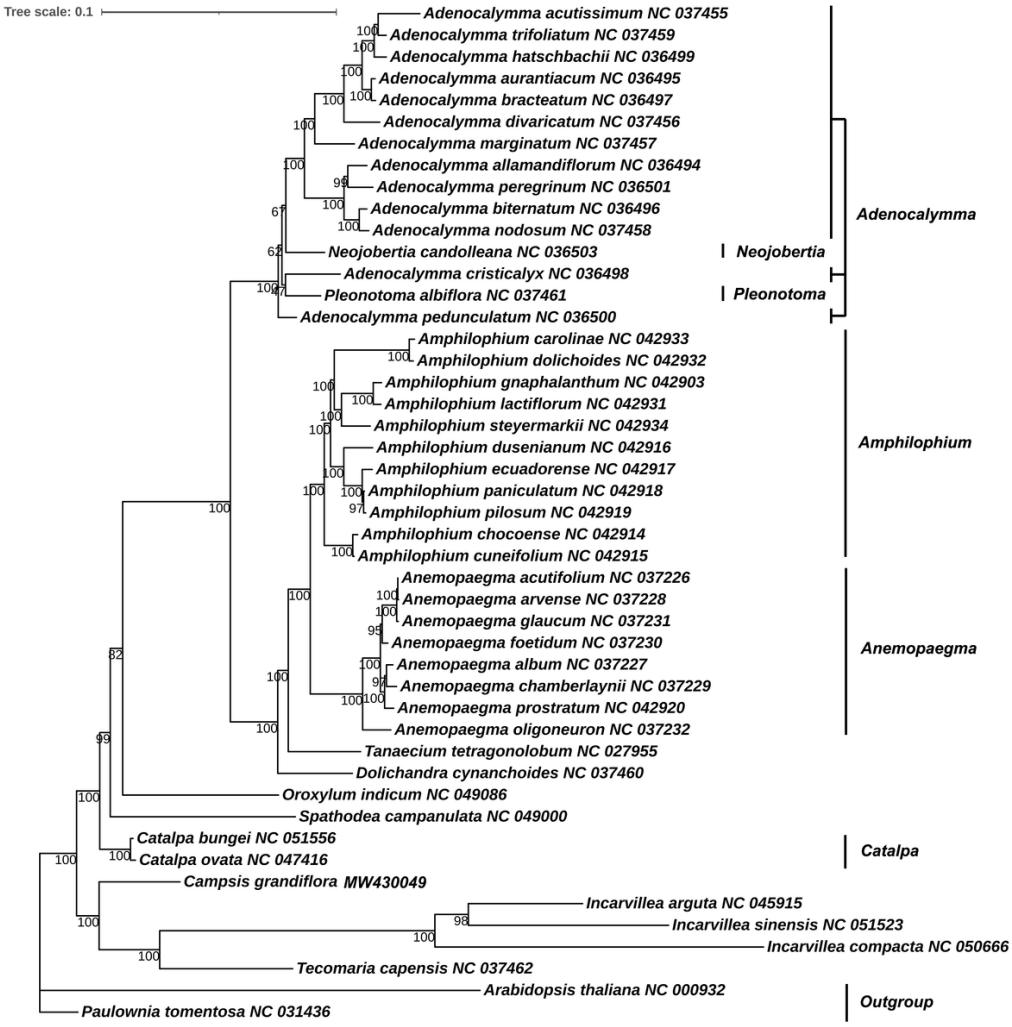

Figure 2

The evolutionary tree of family Bignoniaceae. The phylogenetic results included 45 species within families and 2 outer species. The N. candolleana and P. albiflora interspersed in 13 Adenocalymma species, converged into a large clade together with 11 Amphilophium species, 8 Anemopaegma species, T. tetragonolobum and D. cynanchoides. This large clade subsequently converged with two species in the genus Catalpa and eventually gathered at the base of the evolutionary tree with C. grandiflora, three species in the genus Incarvillea, and T. capensis. According to the evolutionary tree, the event of C. grandiflora differentiation occurred in a relatively early period and have closer genetic relationship with Incarvillea.

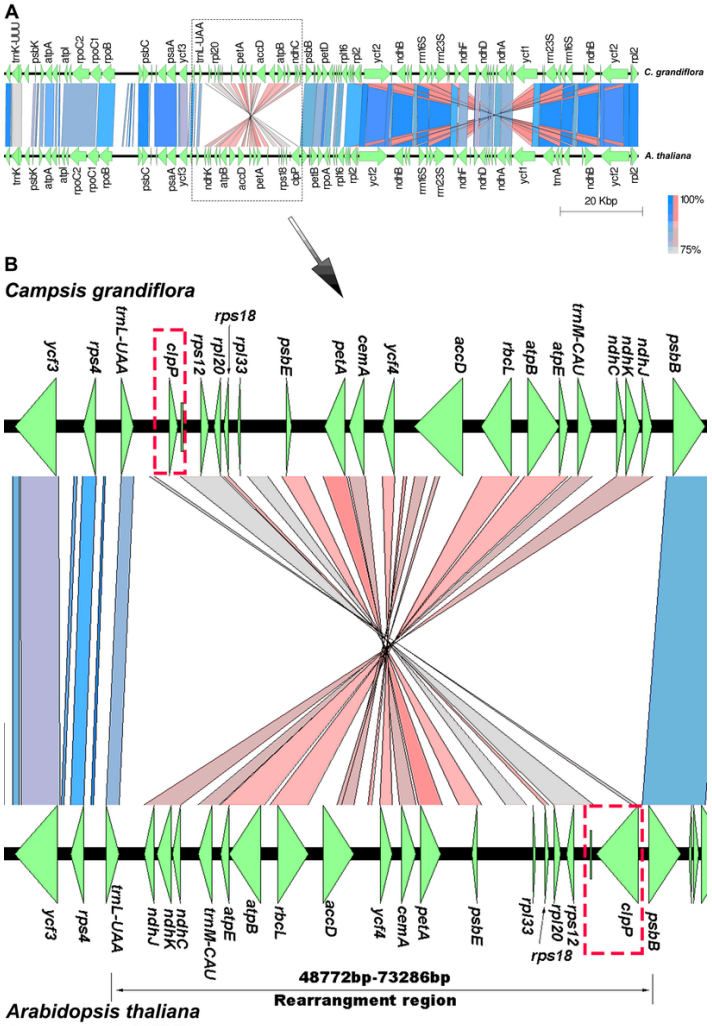

Figure 3

The Easyfig analysis outcome image of C. grandiflora and A.thaliana. A panel shows that each horizontal black line represents a genome. The species names are shown to the right of the corresponding line. The green arrows represent genes, and the direction of the arrows is where the genes start and end on the genome. In the alignment of the two sequences, the conserved regions are bridged by lines, the matching genes in the same direction are connected by blue lines; The reverse and matching genes are connected by red lines. The darker the color, the better the match, or the less the match. B panel shows the details of the genome rearrangement area of C. grandiflora.

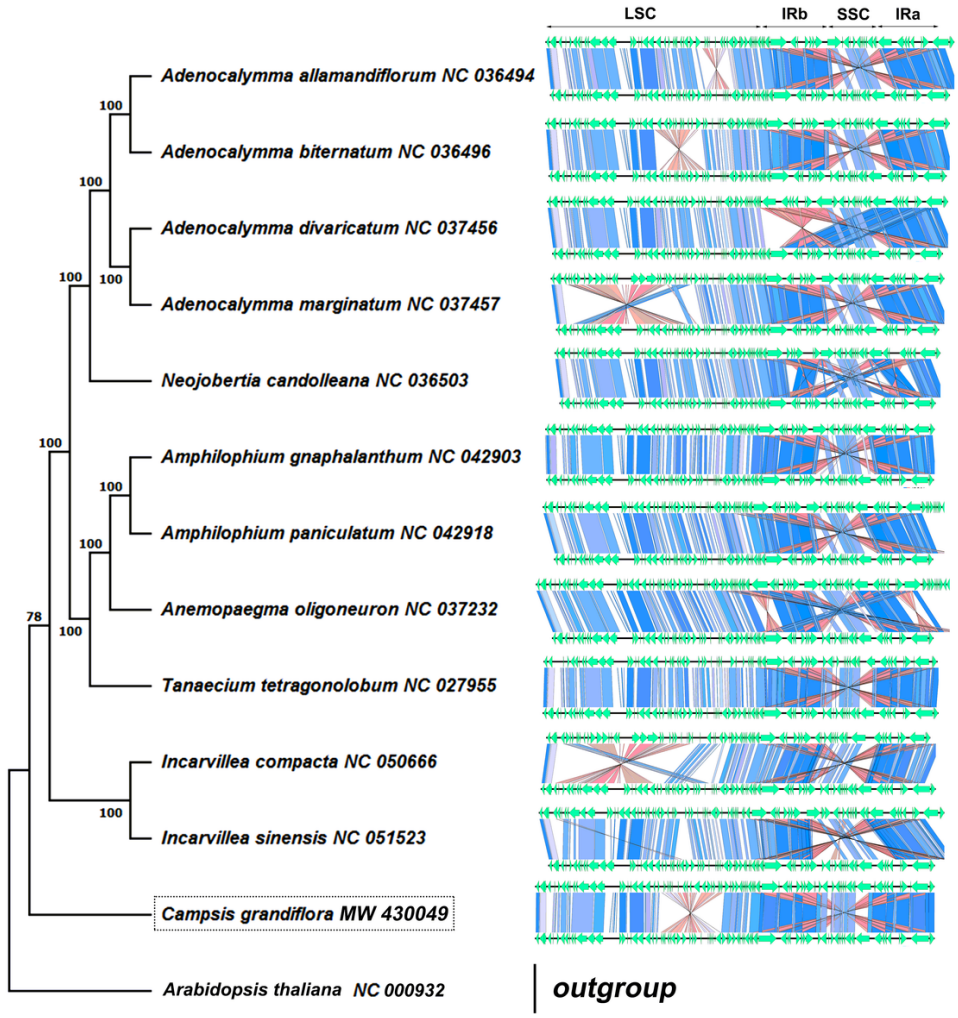

Figure 4

Comparative analysis results of the representative species which have gene rearrangement structure in family Bignoniaceae. The figure shows the phylogenetic analysis of those 12 species and the corresponding genome structure.

Supplementary Files

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

- Table2.docx
- TableS1.docx
- TableS2.docx
- TableS3.docx
- TableS4.docx
- TableS5.docx
- FigureS1.docx
- FigureS2.docx
- FigureS3.docx
- FigureS4.docx