The complete chloroplast genome sequence of *Clerodendranthus spicatus*, a medicinal plant for preventing and treating kidney diseases from Lamiaceae family

Qing Du¹,²,³, Mei Jiang¹,³, Sihui Sun⁴, Liqiang Wang⁵, Shengyu Liu⁶, Chuanbei Jiang⁷, Haidong Gao⁷, Haimei Chen¹, Yong Li⁴, Bin Wang⁴, Chang Liu¹

Received: 28 May 2021 / Accepted: 10 January 2022 / Published online: 21 January 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

Background *Clerodendranthus spicatus* (Thunb.) C. Y. Wu ex H. W. Li is one of the most important medicines for the treatment of nephrology in the southeast regions of China. To understand the taxonomic classification of *Clerodendranthus* species and identify species discrimination markers, we sequenced and characterized its chloroplast genome in the current study.

Methods and results Total genomic DNA were isolated from dried leaves of *C. spicatus* and sequenced using an Illumina sequencing platform. The data were assembled and annotated by the NOVOPlasty software and CpGAVAS2 web service. The complete chloroplast genome of *C. spicatus* was 152,155 bp, including a large single-copy region of 83,098 bp, a small single-copy region of 17,665 bp, and a pair of inverted repeat regions of 25,696 bp. The Isoleucine codons are the most abundant, accounting for 4.17% of all codons. The codons of AUG, UUA, and AGA demonstrated a high degree of usage bias. Twenty-eight simple sequence repeats, thirty-six tandem repeats, and forty interspersed repeats were identified. The distribution of the specific *rps19*, *ycf1*, *rpl2*, *trnH*, *psbA* genes were analyzed. Analysis of the genetic distance of the intergenic spacer regions shows that *ndhG-ndhI*, *accD-psal*, *rps15-ycf1*, *rpl20-clpP*, *ccsA-ndhD* regions have high K2p values. Phylogenetic analysis showed that *C. spicatus* is closely related to two Lamiaceae species, *Tectona grandis*, and *Glechoma longituba*.

Conclusions In this study, we sequenced and characterized the chloroplast genome of *C. spicatus*. Phylogenomic analysis has identified species closely related to *C. spicatus*, which represent potential candidates for the development of drugs improving renal functions.

Keywords *Clerodendranthus spicatus* · Chloroplast genome · Hypervariable region · Phylogenetic analysis · Codon usage

Abbreviations

LSC Large single copy
SSC Small single copy
IR Inverted repeat
SSR Simple sequence repeat
ISSR Inter-simple sequence repeat
cpSSR Chloroplast simple sequence repeat

Qing Du, Mei Jiang, Sihui Sun contributed equally to this work and are co-first authors.

✉ Bin Wang
beinwang@126.com
✉ Chang Liu
clin6688@yahoo.com

Extended author information available on the last page of the article
GC    Guanine and cytosine
IGS   Inter genic spacer regions

Introduction

*Clerodendranthus spicatus* (Thunb.) C. Y. Wu ex H. W. Li, is commonly called Yanumiao, Maoxugong and Maoxucao [1]. It is a perennial herb and is the first reported species among the *Clerodendranthus* genus in the Lamiaceae family. It is mainly produced in Guangdong, Hainan, Guanxi, Yunnan, Fujian provinces of China, and Taiwan. And it is also distributed in India, Myanmar, Thailand, Indonesia, Philippines to Australia and adjacent islands [2]. In the wild, *C. spicatus* commonly grows in the wet place under the forest. They can also be cultivated in areas up to 1050 m above sea level [3]. *C. spicatus* has numerous chemical constituents, such as flavonoids [4], phenols [5, 6], terpenes [7], volatile oils [8], and lignans [9]. The entire plants or aerial parts are widely used to treat chronic nephritis, cystitis, lithangiuria and rheumatoid arthritis, and kidney diseases [10, 11].

The genetic study of *C. spicatus* is still in its infancy. The genetic clustering and genetic similarity analysis of 87 *C. spicatus* samples based on inter-simple sequence repeat (ISSR) markers have been reported previously [12]. The results showed that there were 120 polymorphic bands. And the samples were classified into two groups using the UPGMA method. The ISSR markers were feasible for the genetic diversity analysis of *C. spicatus* resources and can well reveal the genetic characteristics and genetic relationships in *C. spicatus* germplasm resources. It provides a theoretical basis for species identification, variety selection, product development, and application of *C. spicatus*.

The chloroplasts are important organelles for plant photosynthesis [13]. The chloroplast genome (cpgenome) study is significant in revealing the mechanism and metabolic regulation of plant photosynthesis [14]. A large number of proteins that function in the chloroplast are coded and transcribed in the nuclear genome. For this reason, the chloroplast is only a semi-autonomous organelle [15]. Studies of the chloroplast genomes help to understand the interaction between nuclear and chloroplast genomes [16]. Moreover, chloroplast genomes have an important value in revealing the plants' origin, molecular evolution process, and relationship among different species [17].

Here, the cpgenome of *C. spicatus* was sequenced, annotated, and characterized for the first time. After we had characterized its structure features, we chose other 14 species known to have pharmacological effects on the kidney system in clinical research for phylogenetic analysis. *Epimedium brevicornu* and *Dioscorea polystachya* were used as outgroups. In addition, we compared their gene contents to identify possibly related to the chemical biosynthesis and disease treatment [18]. We systematically identified simple sequence repeat markers that can be used in the future for genetic diversity analysis as described above [19]. The results from this study laid the foundation for taxonomic classification and species discrimination of *C. spicatus* related species. This information will be indispensable for the exploration of novel medicinal materials used for treating kidney diseases.

Materials and methods

Plant materials

Young leaves of *C. spicatus* were collected from the Guangxi Medical Botanical Garden, Nanning, Guanxi, China (Geospatial coordinates: N22.859968, E108.383475) and treated with Silica gel immediately for total genomic DNA isolation. The voucher specimen was deposited at the Institute of Medicinal Plant Development under the voucher number: implad201910126, (Contact person: HM Chen; Email: hmchen@implad.ac.cn).

DNA extraction and determination of DNA quality

Total genomic DNA was extracted from the dried leaves using the plant genomic DNA kit (Tiangen Biotech, Beijing, China) [20, 21]. The DNA purity was evaluated with 1.0% agarose gel. Moreover, we measured the DNA concentration using the Nanodrop spectrophotometer 2000.

Chloroplast genome sequencing and assembly

DNA extracts were fragmented for 300 bp short-insert library construction. The library was sequenced in pair-end mode with the read length of 150 bp on an Illumina Hiseq 2500 platform following the manufacturer’s recommendations [22, 23]. The raw reads were filtered using Trimomatic 0.35 to remove adapters and low-quality bases [24]. We used BLASTN software to compare the clean reads and cpgenome sequences downloaded from GenBank. The reads that were similar to the cpgenome sequences satisfying the BLASTn e value cutoff of 1e-5 were filtered. Then, the filtered reads were assembled using NOVOPlasty (v 4.2) software [25] with the following parameters (Type=chloroplast, K-mer=39, Read Length = 150, Insert size = 300, Single/ Paired = PE, Insert Range = 1.9) and the *rbcL* sequences as the seed.
**Genome annotation and manual curation**

We annotated the genome using the CpGAVAS2 web service (http://www.herbalgenomics.org/cpgavas2/) [26], including the prediction of protein-coding genes (PCGs), tRNA gene, rRNA genes, and the analysis of the repeat sequences. The annotation problems were manually corrected using the Apollo software [27]. At last, the assembly and the annotation results of *C. spicatus* cpgenome was deposited in GenBank under the accession number MZ063774.

**Characterization of cis- and trans-splicing genes**

Schematic representation of cpgenome structure and the detailed structures of the cis-splicing genes and trans-splicing PCGs were depicted using CPGview-RSG software (http://www.herbalgenomics.org/cpgview/).

**Analysis of relative synonymous codon usage (RSCU) and codon adaptation index (CAI)**

The Relative Synonymous Codon Usage (RSCU) was calculated as the ratio of the observed codon frequency to the expected frequency of the same codon within a synonymous codon group in the entire coding sequence of the gene concerned. The RSCU patterns and quantity were estimated by using CodonW (v1.4.4) (http://codonw.sourceforge.net/) [28]. The Codon Adaptation Index (CAI) value was calculated by using the CAIcal server (http://genomes.urv.es/CAIcal/) [29].

**Repeat analysis**

The interspersed repeats, tandem repeats and simple sequence repeats were analyzed using the REPuter (parameter: Hamming distance = 3, Minimum repetition unit = 8 bp, https://bibiserv.cebitec.uni-bielefeld.de/reputer/) [30], Tandem Repeat Finder (TRF4.09) (parameter: matching weight = 2, mismatching penalty = 7, Delta = 7, match probability = 80, PI = 10, Minscore = 50, MaxPeriod = 500, http://tandem.bu.edu/trf/trf.html) [31] and MISA (parameter: unit_size, min_repeats = 1–10, 2–6, 3–5, 4–5, 5–5, 6–5; max_difference_between_2_SSRs = 100, http://pgrc.ipk-gatersleben.de/misa/) [32].

**Analysis of the gene contents at the borders of the LSC, SSC, and IR regions**

The boundaries of the LSC, SSC, and IR regions boundary of cpgenomes from *C. spicatus* and six species were visualized using the IRscope software (https://irscope.shinyapps.io/irapp/). These six species include five closely related Lamiales species (*Cistanche deserticola*, *Glechoma longituba*, *Salvia miltiorrhiza*, *Salvia miltiorrhiza f. alba*, and *Tectona grandis*) and *Epimedium brevicornu* from the Berberidaceae family [33].

**Analysis of the hypervariable regions from related species**

The genetic distances of intergenic spaces (IGS) from the cpgenomes of *C. spicatus*, *G. longituba*, *S. miltiorrhiza*, *S. miltiorrhiza f. alba*, and *T. grandis* were calculated by using the distmat program from EMBOSS (v6.3.1) [34] with the Kimura 2-parameters (K2p) [35] evolutionary model.

**Phylogenetic analysis**

We carried out phylogenetic analysis using the concatenated coding sequences (CDS) of the cpgenomes of *C. spicatus* and the other 14 species. These include six Lamiales species (*C. spicatus*, *G. longituba*, *S. miltiorrhiza*, *S. miltiorrhiza f. alba*, *T. grandis*, and *C. deserticola*), one Eucommiaceae species (*E. ulmoides*), one Caprifoliaceae species (*Dipsacus asper*), one Fabaceae species (*Cullen corylifolium*), four Polygonaceae species (*Rheum officinale*, *Rheum palmatum*, *Rheum tanguticum*, and *Rheum nobile*), while two species *E. brevicornu* from the Berberidaceae family and *D. polystachya* from the Dioscoreaceae family was used as the outgroup. The cpgenome sequences were downloaded from GenBank (Table S1). The shared CDS were extracted, concatenated, and aligned by using PhyloSuite (v1.2.2) [36] and MAFFT (v7.313) [37]. Phylogenetic analysis was conducted based on the maximum likelihood (ML) method implemented in IQ-TREE (v1.6.8) [38] under the TVM + F + I + G4 nucleotide substitution model. The reliability of the phylogenetic tree was assessed by bootstrap analysis with 1000 replications. Finally, the phylogenetic tree was visualized using MEGA 5.0 [39].

**Results**

**Isolation, quality control, and sequencing of total genomic DNA**

We analyzed the genomic DNA using electrophoresis on a 1% agarose gel. A clear band corresponding to the genome DNA was observed. The detection value of OD260/OD280 by the microspectrophotometer is 1.8–1.9, and that of OD260/OD230 is over 2.0, indicating that the purity of the DNA
sample is of high quality for PCR experiment and library construction. 11.5 GB of paired-end reads were obtained after filtering, resulting in 10.0 GB clean data. We mapped the reads to the assembled *C. spicatus* cpgenome. For a total of 15,926,314 reads, 354,188 pairs of reads (1.11%) were mapped to the genome exactly one time, corresponding to a depth of 354,188*151/152155 = 351.5.

**General features of *C. spicatus* cpgenome**

The assembled cpgenome of *C. spicatus* was 152,155 bp, including a large-single copy (LSC) region of 83,098 bp, a small-single copy (SSC) region of 17,665 bp, and a pair of inverted repeat regions (IRa and IRb) of 25,696 bp each (Fig. 1). The GC content of the whole cpgenome was 37.86%, and those of LSC, SSC, and IR areas were 35.90%, 31.75%, and 43.12%, respectively (Table 1). The

---

**Clerodendranthus spicatus**

![Fig. 1 Graphic representation of features identified in *C. spicatus* cp genome by using CPGview-RSG (http://www.herbalgenomics.org/cpgview). The map contains seven circles. The first circle shows the distributed repeats connected with red (the forward direction) and green (the reverse direction) arcs from the center going outward. The next circle shows the tandem repeats marked with short bars. The third circle shows the microsatellite sequences as short bars. The fourth circle shows the size of the LSC and SSC. The fifth circle shows the IRA and IRB. The sixth circle shows the GC contents along the plastome. The seventh circle shows the genes having different colors based on their functional groups.](image)
PCGs (rps12, rps7 (× 2), rpl2 (× 2), rpl23 (× 2), ndhB (× 2), ycf1, ycf15 (× 2) and ycf2 (× 2)), 14 tRNA genes (trnM-CAU (× 2), trnL-CAA (× 2), trnV-GAC (× 2), trnE-UUC (× 2), trnA-UGC (× 2), trnR-ACG (× 2), trnN-GUU (× 2)), and 8 rRNA genes (rrn16S (× 2), rrn23S (× 2), rrn5S (× 2), rrn4.5S (× 2)) were located in the IR regions (Fig. 1). Among these genes, 19 cis-splicing genes contained one or two introns, e.g., 10 genes (rps16, atpF, rpoC1, petD, rpl16, rpl2 (× 2), ndhB (× 2)) and ndhA and 7 tRNA genes (trnK-UUU, trnT-CGU, trnL-UAA, trnE-UUC, trnA-UGC (× 2), trnE-UUC) contained one intron. Two PCGs: ycf3 and clpP contained two introns and three exons (Table S2, Figs. S1, and S2). The structures of trans-splicing gene rps12 in the cp genome of C. spicatus are shown in Figure S3. The rps12 contained three exons (Figure S3). The white arrow represents one copy of exon two located in the IRa. The black arrow represents another copy of the exon 2 located in the IRb. The grey arrow represents exon one located in the LSC region. The exon1 and the two copies of exons are trans-spliced together to form two transcripts. The arrow indicated the sense direction of the genes (Figure S3).

Characteristics of tRNA genes

We found that the tRNA genes corresponded to 18 types of amino acids. The amino acid and their corresponding anticodon were Arg (TCT and ACG), Asn (GTT), Asp (GTC), Cys (GCA), Gly (GCC), Gln (TTG), Glu (TTC), His (GTG), Ile (GAT), Leu (CAA and TAG), Met (CAT), Phe (GAA), Pro (TGG), Ser (TGA, GCT, and GGA) Trp (CCA), Thr (ACU, AGC) and Val (GTG, ACT, and ATT). The tRNA genes in the IR regions were transcribed in antisense orientation, whereas the tRNA genes in the LSC and SSC regions were transcribed in the same orientation as the DNA.

Table 1 Base composition in the cp genome of C. spicatus

| Region  | Total | A (bp) | T (bp) | C (bp) | G (bp) | A+T (bp) | C+G (bp) | GC content (%) |
|---------|-------|--------|--------|--------|--------|----------|----------|---------------|
| LSC     | 83,098| 26,002 | 27,263 | 15,257 | 14,575 | 53,265   | 29,832   | 35.90         |
| SSC     | 17,665| 6034   | 6021   | 2677   | 2932   | 12,055   | 5609     | 31.75         |
| IRb     | 25,696| 7324   | 7292   | 5337   | 5744   | 14,616   | 11,081   | 43.12         |
| IRA     | 25,696| 7292   | 7324   | 5337   | 5744   | 14,616   | 11,081   | 43.12         |
| Total   | 152,155| 46,652 | 47,900 | 29,015 | 28,588 | 94,552   | 57,603   | 37.86         |

GC content of the IR region was higher than those of the SSC and LSC regions.

Gene content

We predicted 131 genes in the cp genome, including 87 PCGs, 36 tRNA genes, and 8 rRNA genes (Table 2). 14 PCGs (rps12, rps7 (× 2), rpl2 (× 2), rpl23 (× 2), ndhB (× 2), ycf1, ycf15 (× 2) and ycf2 (× 2)), 14 tRNA genes (trnM-CAU (× 2), trnL-CAA (× 2), trnV-GAC (× 2), trnE-UUC (× 2), trnA-UGC (× 2), trnR-ACG (× 2), trnN-GUU (× 2)), and 8 rRNA genes (rrn16S (× 2), rrn23S (× 2), rrn4.5S (× 2), rrn5S (× 2)) were located in the IR regions (Fig. 1). Among these genes, 19 cis-splicing genes contained one or two introns, e.g., 10 genes (rps16, atpF, rpoC1, petD, rpl16, rpl2 (× 2), ndhB (× 2) and ndhA) and 7 tRNA genes (trnK-UUU, trnT-CGU, trnL-UAA, trnE-UUC, trnA-UGC (× 2), trnE-UUC) contained one intron. Two PCGs: ycf3 and clpP contained two introns and three exons (Table S2, Figs. S1, and S2). The structures of trans-splicing gene rps12 in the cp genome of C. spicatus are shown in Figure S3. The rps12 contained three exons (Figure S3). The white arrow represents one copy of exon two located in the IRa. The black arrow represents another copy of the exon 2 located in the IRb. The grey arrow represents exon one located in the LSC region. The exon1 and the two copies of exons are trans-spliced together to form two transcripts. The arrow indicated the sense direction of the genes (Figure S3).

Table 2 Genes contents of the cp genome of C. spicatus

| Category for genes | Group of genes | Name of genes |
|--------------------|----------------|--------------|
| rRNA               | rRNA genes     | rrs16S (× 2), rrs23S (× 2), rrs5S (× 2), rrs4.5S (× 2) |
| tRNA               | tRNA genes     | 36 unique trna genes (6 contains 1 intron) |
| Self-replication   |                |              |
| Photosynthesis     |                |              |
| Other genes        |                |              |
| Genes of unknown functions | | ycf1, ycf15 (× 2), ycf2 (× 2), ycf4 |

The “(×2)” indicates that the gene located in the IRs and thus had two copies.
(GGT, and TGT), Tyr (GTA) and Val (GAC), respectively (Table S3).

Analysis of RSCU and CAI regarding the codon usage

Codon usage bias provides critical clues for the plant evolution process and has an essential role in the expression of proteins [40]. We analyzed the codon usage in the cpgenome of *C. spicatus*. The PCGs contained a total of 26,421 codons. Among them, 1102 (4.17%) codons were for Isoleucine, the most abundant amino acid. Lysine was the second most abundant amino acid. Its codons accounted for 4.06% of all codons. The least abundant codons were for cysteine, representing only 0.24% of all codons (Table S4).

The RSCU bias affects the gene expression because synonymous codon usage can usually influence the protein secondary structural units. The RSCU values varied from 0.336 to 2.9904 in the cpgenome of *C. spicatus*. Codons having an RSCU value > 1 was preferred. In this study, the codons of AUG, UUA, and AGA had the higher RSCU values, indicating the presence of higher codon usage bias among the 64 codons. To understand the synonymous codon usage bias for the cpgenome of *C. spicatus*. We calculated the CAI value, which was 0.645, indicating that the presence of codon preference. It should be noted that only two amino acids (Trp and Met) were encoded by only one codon, respectively.

Repeat sequences analysis

Repeat sequences are important genetic markers and are closely related to the origin and evolution of species. Repeat sequences can be divided into scattered (interspersed) repetition and tandem repetition sequences. Interspersed repetition sequences are scattered and distributed in the genome. Multiple repeats of a sequence on a chromosome are called tandem repeats. A special form of tandem repeats is simple tandem repeats, also known as simple repeats (SSR). SSRs often have natural polymorphism and are the focus of chloroplast genome repeat analysis.

We analyzed the three types of repetitive sequences, including SSRs, tandem repeats, and interspersed repeats in the cpgenome of *C. spicatus*. For the SSR, we found 12 of 28 contained “A” as the repeat unit, 15 of 28 contained “T” as the repeat unit, and 1 SSR contained “TA” as the repeat unit (Table S5). The sizes of SSRs were between 10 and 15 bp (Table S6). 36 repeats were identified for the tandem repeats satisfying the conditions: the length of the repeat unit ≥ 20 bp, and the similarity among the repeat unit sequences ≥70% (Table S6). For the interspersed repeats, 40 repeats were identified, including 22 palindromic repeats and 18 direct repeats (Table S7), with the length of repeat units 1, 2 being between 30 and 60 bp. The e values of interspersed repeats varied from 7.80E–23 to 6.19E–04 (Table S7).

Structures of the IR boundaries from seven selected species

IR contraction and expansion have been a common phenomenon in the genome evolution of cpgenome. We analyzed the IR boundaries’ structure and calculated the length of various regions for seven species. The results showed that these cpgenomes had similar structures, but different lengths of corresponding regions (Fig. 2). Two genes, *rps*19 and *ndhF*, were found most frequently at the border regions. The *rps*19 genes were located at the border area of LSC and IRb in *C. spicatus, S. miltiorrhiza, S. miltiorrhiza f. alba*, and *T. grandis*. In contrast, the *rps*19 genes were located in the LSC of *E. brevicornu* cpgenome. In *C. deserticola* cpgenome, it was located in IRb. More interestingly, a small fragment of the *rps*19 gene was found at the border area of IRa and LSC in Salvia species and *C. deserticola*. No *rps*19 and *ycf1* gene fragments were found in the *G. longituba* cpgenome. According to the IR regions boundary and sizes analysis of *C. spicatus*, the *rps*19 genes have various distributions across the borders of the single-copy and the inverted regions [41].

In contrast, *ndhF* genes were located at the border area of IRb and SSC in the cpgenomes of *S. miltiorrhiza* [42], *S. miltiorrhiza f. alba*, and *T. grandis*, and the area of SSC in the cpgenome of *E. brevicornu*. Most of the *ycf1* genes were located at the border area of SSC and IRa in the *Salvia* species and *T. grandis*, while they were also located at the border area of SSC and IRb in the cpgenomes of *C. spicatus, E. brevicornu, G. longituba*, and *T. grandis*. Besides, the *rpl2*, *trnH*, and *rpl22* genes were located at the IR and LSC border areas, except for the cpgenome of *G. longituba*, respectively. Lastly, the *psbA* genes were located in the cpgenome of *Salvia* species, *T. grandis*, and *C. spicatus* (Fig. 2).

Identification of hypervariable regions

We selected five Lamiales species: *C. spicatus, S. miltiorrhiza, S. miltiorrhiza f. alba, T. grandis*, and *G. longituba*, to analyze the genetic distance among the intergenic spacer regions (IGS) [43]. The result showed that 30 out of 58 selected IGS had K2p values varying from 6.084 to 29.242 (Figure S4). Among them, five IGS had higher K2p values above the value of 18.0, namely, which were *ndhG-ndhI* (29.242), *accD-psaI* (22.442), *rps15-ycf1* (19.488), *rpl20-clpP* (18.322) and *cysA-ndhD* (18.091). These regions are suitable candidates for the development of molecular markers in the future.
Phylogenetic analysis

For phylogenetic analysis, we extracted a total of 85 different CDS from 15 species (Table S1 and S8). Among them, 21 CDS were present in all these 15 species, including rpl14, rpl36, rps3, rps4, rps7, rps11, rps12, rps14, rps15, rps18, rps19, rpl2, rpl16, rpl20, rpl32, marK, psbM, rpl33, rps2, rps8 and ycf2. In particular, the sequences of rpl14, rpl36 were identical in all the species, demonstrating their extreme conservation. The genes psbZ, rpoC2, psbL, and ycf4 were present in all the other 14 species except C. deserticola, G. longituba, T. grandi, Rehum nobile, and C. corylifolium. The gene ycf15 was present in 12 species except C. corylifolium, Dipacys asper, and E. brevicornu. The genes rpoC, ycf68, and lhbA were annotated in T. grandis, E. ulmoides and G. longituba, respectively. In the species of C. corylifolium, genes of orf42 and orf56 are distinct (Table S8).

The phylogenetic tree showed that all 13 species were clustered together besides the two outgroup species, E. brevicornu, and D. polystachya. The Maximum likelihood (ML) bootstrap values were fairly high. The values for seven nodes are 100%, with values ≥ 87% for another four nodes. The 6 Lamiales species, E. ulmoides, and D. asper, were clustered together [44]. In contrast, the 4 Polygonaceae species and C. corylifolium from the Fabaceae family were clustered together. C. deserticola forms its own branch because this is possible considering it is a parasitic plant. Nevertheless, the two plants of E. brevicornu and D. polystachya were more distantly related to the others (Fig. 3 and Table S8).

Discussion

We sequenced and assembled a C. spicatus chloroplast genome for the first time, characterized its structure, and identified many unique features. A total of 131 protein-coding genes were predicted, including 87 protein-coding genes, 36 tRNA genes, and 8 rRNA genes. Five hypervariable
regions have been identified. The IR border analysis revealed a complicated IR contraction and expansion pattern, suggesting that the IR border areas undergo continuous recombinations. Codon usage analysis identified significant preference. In summary, these unique features can be used to understand the evolution process and functions of *C. spicatus* chloroplast.

We have chosen 15 species having common medical values for phylogenetic analysis. Most of them have been used to treat renal and urinary diseases, such as acute and chronic nephritis, cystitis, urinary calculi, lithanguria, nephropelitistis, diuresis, premature ejaculation, soreness of the waist, and skelalgia. As reported, Fructus Psoraleae from dry mature fruits of *C. corylifolium* has been used clinically to treat kidney yang deficiency-induced impotence due to its strong estrogen-like activity [45]. Long-term administration of *D. polygonoides* can induce changes of AST and ALT activities in renal and hepatic functions [46]. A polysaccharide from the roots of *D. asper* can regulate the renal complication through inhibition of AGE accumulation and RAGE expression in streptozotocin STZ-treated diabetic rats [47]. Comparison of raw product and Wine Steam-Processed Product on biological activities from *C. deserticola* showed that they could restore the level of sex hormone in the kidney-yang deficiency model [48]. Emodin is the main activity compound from *Rheum* genus of Polygonaceae family. It is reported that emodin can effectively treat proliferative glomerulonephritis and delay chronic renal failure [49]. The likely mechanism involves the inhibition of the production of inflammatory factors such as IL-1, IL-6, TNF and MCIL-6. Lastly, *Glechoma, Salvia* genus, *Tectona grandis*, and Herba *Epimedi* contain terpenoids, flavonoids, steroids, alkaloids, and organic acids with anti-oxidation and antiinflammatory biological activities such as treatment of gallbladder, diuresis, dissolved stones and male infertility [50–53]. Their similar clinical effects suggest that they might contain similar chemical components. This study was initiated to explore the therapeutic potential of the these plant species fully.

Results obtained from this study have several applications. Firstly, the chloroplast transgenic expression system can efficiently express exogenous proteins. Several oral vaccine antigens and therapeutic proteins have been overexpressed using this system [54]. As a result, the cpgenome generated from this study can be used to express endogenous and exogenous genes. Several oral vaccine antigens and therapeutic proteins have been over-expressed using this system [54]. As a result, the cpgenome generated from this study can be used to express endogenous and exogenous genes. In particular, we can improve the expression of genes related to the biosynthesis of secondary metabolites from *C. spicatus*. Secondly, we have established the phylogenetic relationships of these 15 medicinal plant species that are potentially effective on kidney disease. In the future, we can profile the chemical compositions from these species. We might identify additional therapeutic areas for these species beyond their traditional uses through the combined analysis of the phylogenetic relationship and the chemical profiles.
Conclusions

In the present study, we sequenced and assembled the C. spicatus cp genome for the first time using Illumina sequencing technology. We have identified several unique features of the cp genome, including codon usage preference, IR contraction and expansion, hypervariable regions, etc. The unique characteristics of the cp genome provide valuable resources for subsequent genetic analysis and bioprospecting.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11033-022-07135-4.

Acknowledgements We want to thank Mr. Yang Ni, Jingling Li, Miss Jingwen Yue, and Mr. Junchen Zhou, who have provided support in analyzing the data.

Authors contribution CL conceived the study; MJ and LQW collected C. spicatus samples, extracted DNA for next-generation sequencing; DQ, SSH, SYL assembled, validated the assembled genome, performed data analysis, and drafted the manuscript; LQW, HMC, CBJ, HDG, BW and YL reviewed the manuscript critically. All authors have read and agreed on the contents of the manuscript.

Funding This work was supported by funds from the National Science &Technology Fundamental Resources Investigation Program of China [2018FY100703], Chinese Academy of Medical Sciences, Innovation Funds for Medical Sciences (CIFMS) [2021–12M-1-022], National Science Foundation [81872966], Qinghai Provincial Key Laboratory of Phytochemistry of Qinghai Tibet Plateau [2020-ZJ-Y20], the Key Scientific Research Projects of Hunan Education Department [No. 20A447], Hunan technological innovation guidance project [2018SK52001]. The funders were not involved in the study design, data collection, and analysis, decision to publish, or manuscript preparation.

Data availability The cp genome sequence data of C. spicatus that support the findings of this study are openly available in NCBI at GenBank database with accession number MZ063774. (https://www.ncbi.nlm.nih.gov). The associated BioProject, SRA, and Bio-Sample numbers are PRJNA723363, SRR14350365 and SAMN18814878, respectively.

Declarations

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.

References

1. Acta Phytotax (1974) Sin. 12:233
2. Xie ZY (1996) National herbal compendium, 2nd edition, In: Kuang LJ (ed) People’s Medical Publishing House: Beijing, p 572
3. Jia MR, Li XW (2005) Chinese Ethnic Materia Medica, 1st ed, In: Zhao YX (ed) China Medical Science and Technology Press, Beijing, p 166
4. Zhao AH, Zhao QS, Li RT, Sun HD (2004) Chemical constituents from Clerodendranthus spicatus. Acta Bot Yunnanica 26(5):563–568
5. Zheng QX, Sun ZC, Zhang XP, Yuan JQ, Wu HF, Yang JS, Xu XD (2012) Clerodendranolic acid, a new phenolic acid from Clerodendranthus spicatus. Molecules 17(11):13656–13661
6. Wang M, Liang JY, Chen XY (2007) Water-soluble constituents of Clerodendranthus spicatus. Chin J Nat Med 5(1):27–30
7. Chen YY, Tan CH, Tan JJ, Zhao XM, Jiang SH, Zhu DY (2009) Two new diterpenoid glucosides from Clerodendranthus spicatus. Helv Chim Acta 92:2802–2807. https://doi.org/10.1002/hjca.200900121
8. Zhong JY, Wu ZS (1984) Chemical constituents of Clerodendranthus spicatus. Acta Bot Yunnanica 6(3):344–345
9. Zou J, Zhu YD, Zhao WM (2008) Two new alkyl glycosides from Clerodendranthus spicatus. J Asian Nat Prod Res 10(7–8):603–607. https://doi.org/10.1080/10286020802133076
10. Chen Z, Yang CX, Mao XJ (2015) Progress in related pharmacological research on function of Clerodendranthus spicatus. Pharm Inform 4(2):23–30. https://doi.org/10.12677/PI.2015.42003
11. Zhang YY, Wu J, Li SP, Li NL, Huang YF, Lin QH (2021) Research progress on chemical constituents and pharmacological actions of Clerodendranthus spicatus. Acta Chin Med Pharmacol 49(1):112–120
12. Luo C, Yue XD, Wu FH, Gao X, Wang TX (2015) Genetic diversity in Clerodendranthus spicatus (Thunb.) based on ISSR markers. J Trop Biol 6(2):204–222. https://doi.org/10.15886/jtcb.2015.02.016
13. Robles P, Quesada V (2021) Organelle genetics in plants. Int J Mol Sci 22(4):2104. https://doi.org/10.3390/ijms22042104
14. Schmidt SB, Eisenhut M, Schneider A (2020) Chloroplast transition metal regulation for efficient photosynthesis. Trends Plant Sci 25(8):817–828. https://doi.org/10.1016/j.tplants.2020.03.003
15. Van Dingenen J, Blomme J, Gonzalez N, Inzé D (2016) Plants grow with a little help from their organelle friends. J Exp Bot 67(22):6267–6281. https://doi.org/10.1093/jxb/erw399
16. Xiong AS, Peng RH, Zhuang J, Gao F, Zhu B, Fu XY, Yue Y, Jin XF, Tian YS, Zhao W, Yao QH (2009) Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnol Adv 27(4):340–347. https://doi.org/10.1016/j.biotechadv.2009.01.012
17. Gray MW (1999) Evolution of organellar genomes. Curr Opin Genet Dev 9:678–687. https://doi.org/10.1016/S0959-437X(99)00330-1
18. Daniell H, Lin C-S, Ming Yu, Chang W-J (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 17(1):134. https://doi.org/10.1186/s13059-016-1004-2
19. Nigam D, Garcia-Ruiz H (2020) Variation profile of the orthotopic virus genome. Pathogens 9(7):521. https://doi.org/10.3390/pathogens9070521
20. Vieira L, Do N, Faoro H, Fraga HP, De F, Rogalski M, de Souza EM, de Oliveira Pedrosa FRO, Nodari, Guerra MP (2014) An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS ONE 9(1):e84792. https://doi.org/10.1371/journal.pone.0084792
21. Lei WJ, Ni DP, Wang YJ, Shao JD, Wang XC, Dan Y, Wang JS, Chen HM, Liu C (2016) Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. Sci Rep 6:21669. https://doi.org/10.1038/srep21669
22. Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordiero GM, Henry RJ (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9(3):328–333. https://doi.org/10.1111/j.1467-7652.2010.00558
23. Cronn R, Liston A, Parks M, Gernandt D, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using
Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36(19):e122. https://doi.org/10.1093/nar/gkn502

24. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2144–2148. https://doi.org/10.1093/bioinformatics/btu170

25. Dierckxsens N, Mardulpyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45(3):el4. https://doi.org/10.1093/nar/gkw955

26. Shi LC, Chen HM, Jiang M, Wang LQ, Wu X, Huang LF, Liu C (2019) CGPASV2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res 47(W1):65–73. https://doi.org/10.1093/nar/gkz345

27. Firtina C, Kim JS, Alser M, Cali DS, Ercument Cicek A, Alkan C, Mutlu O (2020) Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. Bioinformatics 36(12):3669–3679. https://doi.org/10.1093/bioinformatics/btaa179

28. Peden JF (1999) Analysis of codon usage. PhD Thesis, University of Nottingham, UK

29. Pugibio P, Bravo IG, Garcia-Valle S (2008) CAALcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3:38. https://doi.org/10.1186/1745-6150-3-38

30. Kurtz S, Jomuna A, Choudhuri V, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29(22):4633–4642. https://doi.org/10.1093/nar/gk663

31. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. https://doi.org/10.1093/nar/27.2.573

32. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA - Scalable and accurate assembly polishing algorithm. Bioinformatics 36(12):3669–3679. https://doi.org/10.1093/bioinformatics/btaa179

33. Amiryousefi A, Hyvo Nen J, Poczai P (2018) IRscope: an online and analyzer. Nucleic Acids Res 47(W1):65–73. https://doi.org/10.1093/nar/gkz345

34. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. https://doi.org/10.1093/nar/27.2.573

35. Zhang BB, Nie SQ, Liang JY, Wu FH, Feng F (2014) Phylogenetic analysis and development of molecular markers for five medicinal Alpinia species based on complete plastome sequences. BMC Plant Biol 21:431. https://doi.org/10.1186/s12870-021-03204-1

36. Qian J, Song YJ, Gao HH, Zhu YJ, Xu J, Pang XH, Yao H, Sun C, Li XE, Li CY, Liu JY, Xu HB, Chen SL (2013) The complete chloroplast genome sequence of the medicinal plant Salvia miltiorriza. PLoS ONE 8(2):e57607. https://doi.org/10.1371/journal.pone.0057607

37. Poczai P, Hyvonen J (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37(4):1897–1912. https://doi.org/10.1007/s11392-009-0626-3

38. Liang CL, Wang L, Lei J, Duan BZ, Ma WS, Xiao SM, Qi HJ, Wang Z, Liu YQ, Shen XF, Guo S, Hu HY, Xu J, Chen SL (2019) A Comparative analysis of the chloroplast genomes of four Salvia medicinal plants. Engineering 5:907–915. https://doi.org/10.1016/j.eng.2019.01.017

39. Cai XY, Zhang ZJ, Xiong JL, Yang M, Wang ZT (2021) Experimental and molecular docking studies of estrogen-like and anti-osteoporosis activity of compounds in Fructus Psoraleae. J Ethnopharmacol 276(43):114044. https://doi.org/10.1016/j.jep.2021.114044

40. Marsha-LynMcKoy, Grant K, Asemota H, Simon O, Omoruyi F (2015) Renal and hepatic function in hypercholesterolemic rats fed jamaican bitter yam ( Dioscorea polygoenoides). J Dietary Suppl 12(2):173–183. https://doi.org/10.3109/19390211.2014.952860

41. Li T, Hua QJ, Li N, Cui Y, Zhao M (2019) Protective effect of apolysaccharide from Dipascus asper Wall on strep to zotocin(STZ)-induced diabetic nephropathy in rat. Int J Biol Macromol 133:1194–1200. https://doi.org/10.1016/j.ijbiomac.04.069

42. Zhang Y, Wang YW, Yang S, Xiao YF, Guan HB, Yue X, Wang XQ, Li XR (2019) The difference of chemical components and biological activities of the raw products slices and the wine steam-processed product from Cistanche deserticola. Evidence-Based Compl Alternat Med (3):1–10. https://doi.org/10.1155/2019/2167947

43. Ding Y, Huang ZH (2007) Progress in pharmacological effects of emodin, pharmacology and clinical of Chinese medicine. 23(5): 236–238. https://doi.org/10.13412/j.cnki.zyyl.2007.05.119

44. Zhang QJ, Yang XS, Zhu HY, Hao XJ (2006) Chemical constituents and their pharmacological research progress in medicinal plants of Glechoma Linn. Chin Tradit Herbal Drugs 37(6):950–952. https://doi.org/10.1016/j.jep.2021.06-1325

45. Zhang BB, Nie SQ, Liang JY, Wu FH, Feng F (2014) Researches on the chemical constituents and bioactivity of Salvia spp. Strait Pharm J 26(11):1–5

46. Ramachandran S, Rajasekaran A, Kumar KTM (2011) Antidiabetic, antiatherosclerotic and antioxidant potential of methanol extract of Tectona grandis flowers in streptozotocin induced diabetic rats. Asian Pac J Trop Med 4(8):624–631. https://doi.org/10.1016/S1674-7645(11)00160-0

47. He LJ, Jiang JJ, Chen H, Zhang L, Zheng XY, Xu RQ, Huang NL (2020) A review on pharmacological effects and clinical application of Herba Epimedii. Clin J Chin Med 12(2):17–20

48. Kumar S, Daniell H (2004) Engineering the chloroplast genome for hyperexpression of human therapeutic proteins and vaccine antigens. In: Balbás P, Lorence A (ed) Recombinant Gene Expression. Methods in Molecular Biology. Humana Press. https://doi.org/10.1385/1-59259-774-2:365.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Authors and Affiliations

Qing Du1,2,8 · Mei Jiang1,3 · Sihui Sun4 · Liqiang Wang5 · Shengyu Liu6 · Chuanbei Jiang7 · Haidong Gao7 · Haimei Chen1 · Yong Li4 · Bin Wang4 · Chang Liu1

Qing Du
171765300@qq.com

Mei Jiang
mjiang0502@163.com

Sihui Sun
sunshu0311@163.com

Liqiang Wang
lys832000@163.com

Shengyu Liu
shengyuliu558@163.com

Chuanbei Jiang
842545698@qq.com

Haidong Gao
gaohd@genepioneer.com

Haimei Chen
hmchen@implad.ac.cn

Yong Li
13975599139@126.com

1 Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, People’s Republic of China

2 College of Pharmacy, Qinghai Provincial Key Laboratory of Phytochemistry of Qinghai Tibet Plateau, Qinghai Minzu University, Xining, Qinghai 810007, People’s Republic of China

3 School of Pharmaceutical Sciences, Qilu University of Technology, Shandong Academy of Sciences, Jinan, Shandong 250353, People’s Republic of China

4 College of Pharmacy, Xiangnan University, Chenzhou, Hunan 423000, People’s Republic of China

5 College of Pharmacy, Heze University, Heze, Shandong 274015, People’s Republic of China

6 Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, People’s Republic of China

7 Genepioneer Biotechnologies Inc., Nanjing, Jiangsu 210023, People’s Republic of China

8 Fresh Sky-Right (Beijing) International Science and Technology Co. Ltd, Beijing 100187, People’s Republic of China