The complete chloroplast genome provides insight into the polymorphism and adaptive evolution of *Garcinia paucinervis*

Yifei Wang\(^a\), Bo Zhao\(^a\), Zhaocen Lu\(^b\), Yancai Shi\(^b\) and Jingjian Li\(^a\)

\(^a\)Department of Pharmacognosy, College of Pharmacy, Guilin Medical University, Guilin, China; \(^b\)Department of Characteristic Economic Plant Research Center, Guangxi Institute of Botany, The Chinese Academy of Sciences, Guilin, China

**ABSTRACT**

*Garcinia paucinervis* is an evergreen tree with high medicinal value. Due to its vulnerable reproductive capacity coupled with excessive logging by humans, *G. paucinervis* has become an endangered species. In order to protect this species effectively, we drew the complete chloroplast genome of *G. paucinervis* and performed a series of comparative analyses on *G. paucinervis* and its neighbouring species. The chloroplast genome size of *G. paucinervis* is 157702 bp. In the chloroplast genome, we identified 130 genes, 56 RNA editing sites in the protein-coding genes, as well as 241 simple sequence repeats (SSRs) and 49 complex repetitive sequences. Comparative analysis identified some high divergent sequences in the intergenic spacers, which can be used as candidate markers for phylogenetic study. From an adaptive evolution point of view, a branch-site model analysis identified positively selected sites in 5 genes, most of which are involved in ribosome biogenesis, protein synthesis and other developmental processes. The detected codon substitutions may be associated with the evolution of *G. paucinervis* to adapt to the extreme habitat in karst. The ML tree revealed that *G. mangostana* was a sister to *G. gummi-gutta*, and they formed a diverging lineage to *G. paucinervis* in *Garcinia*. The above information is of great significance for us to understand the evolution of *G. paucinervis* chloroplast and lays the foundations for future studies in species conservation.

**INTRODUCTION**

*Garcinia paucinervis* (Garcinaceae) is a precious timber tree species which grows only in the karst mountain in southwestern China and northern Vietnam. This tree exhibits strong and heavy wood material characteristics, which make it suitable for house construction, furniture manufacture and ship-building [1]. In addition to the economic value, the branches, leaves and bark of *G. paucinervis* were traditionally used as medicine for heat-clearing, detoxification and detumescence [2]. At present, due to its vulnerable reproductive capacity coupled with excessive logging by humans, *G. paucinervis* has become an endangered species and has been listed as a class II national key protected species in China [3]. In order to protect this species effectively, researchers have paid a lot attention to formulate strategies for the cultivation and conservation of *G. paucinervis* [4]. However, there are still few studies regarding the evolution and phylogeny of *G. paucinervis* on the molecular level, which we believe to be an important yet underappreciated direction for the conservation of this species.

The chloroplast is a plant-specific organelle and the main place for photosynthesis. In angiosperms, the chloroplast genome is one of the three DNA genomes, which belongs to maternal inheritance and has highly conservative characteristics. The chloroplast genome of traditional angiosperms ranges from 120 and 160 kb in length [5, 6], and is a circular structure composed of four regions (i.e. a pair of inverted repeats (IRs), a large single-copy region (LSC) and a small single-copy region (SSC)). The contraction and expansion of the IR region is the main reason for the mutation of the chloroplast genome, but in general, the chloroplast genome is conserved in gene content and...
organization, this characteristic make it an ideal genetic resource for plant evolutionary studies [7]. Recent studies demonstrated that the chloroplast genome has the potential to resolve phylogenetic relationships at the species level, infer divergence dating and elucidate the adaptive evolution in specific species [8]. Especially, the molecular markers derived from the chloroplast genome can be used to determine the gene flow, genetic structure and genetic diversity among different populations [9]. Such valuable information can facilitate the efficient formulation of strategies for the conservation of plant genetic resources, and ultimately minimize the extinction risk of endangered plants.

In this study, we assembled the complete chloroplast genome of *G. paucinervis* and characterized its genomic structure and features including the gene content, repeat sequence and IR expansions/contractions, etc. Combining the previously reported chloroplast genomes of related species, we analyzed the sequence divergence, phylogenetic relationship and adaptive evolution. These results will provide insights into the RNA editing patterns, positively selected sites and evolutionary history of *G. paucinervis*, and the highly divergent regions detected could be employed to develop potential markers for plant identification. This information will also help us to develop an effective strategy for the conservation of *G. paucinervis*.

**Materials and methods**

**Plant materials and DNA extraction**

Fresh leaves of *G. paucinervis* were collected from Guangxi Institute of Botany, Chinese Academy of Sciences, Guilin, China. The total genomic DNA was extracted according to the modified CTAB method [10]. Since the *G. paucinervis* leaf contains high level of polysaccharides and polyphenols, we used double dosage of polyvinylpyrrolidone and β-mercaptoethanol during the early stages of DNA extraction. Finally, the DNA was quantified using a NanoDrop 2000 spectrophotometer.

**Genome sequencing and assembling**

The DNA of sufficient quality was fragmented by mechanical breakage, and then a series of treatments such as fragment purification and terminal repair were carried out. The fragments of 350 bp in length were selected via agarose gel electrophoresis and then amplified by polymerase chain reaction (PCR) to construct a sequencing library [11]. A paired-end library was constructed using Illumina NovaSeq platform following the manufacturer’s protocol (Illumina, San Diego, California, USA). After quality assessment, the chloroplast genome related reads were filtered by mapping all the raw reads to the reference chloroplast genome of *Garcinia mangostana* (NC_036341). The filtered reads were assembled into contigs using SPAdes software (version: 3.10.1; parameter:-k 127) [12]. All the contigs were aligned to the reference chloroplast genome by blastn (version: BLAST 2.2.30+; parameter: -value 1e-5) [13]. Based on the alignment result, the gap between contigs was filled by Gapcloser included in the SOAP package v1.12 [14]. Finally, the complete chloroplast genome sequence of *G. paucinervis* was obtained.

**Genome annotation and sequence characterization**

The chloroplast genome was annotated using the online program GeSeq, a special annotation software for organelles [15]. The coding sequences (CDS) and rRNAs were identified based on BLAST homology searches, followed by manual correction for the questionable annotation. Meanwhile, all tRNA genes were confirmed using the tRNAscan-SE online search server. The circular genome map was drawn by the OGDRAW program [15] and the whole sequence was submitted to GenBank (MT501656). The GC content was calculated for the whole chloroplast genome, and the codon usage frequencies and the relative synonymous codon usage (RSCU) were analyzed for all the CDS using MEGA5.1. In addition, potential RNA editing sites in protein-coding genes were predicted using the PREP suite with a cut-off value of 0.8 [16]. To identify simple sequence repeats (SSRs), the Perl script MISA was used with the filter thresholds set at repeat units ≥ 8 for mononucleotide, repeat units ≥ 4 for dinucleotides and trinucleotides, and repeat units ≥ 3 for tetranucleotides, pentanucleotides and hexanucleotides. All of the SSRs found were manually checked, and the redundant results were discarded. For the identification of complex repetitive sequences (i.e. forward, reverse, complement and palindromic), REPuter online software was used with parameters set as follows: hamming distance equal to 3, the minimum repeat size ≥ 30 bp and the sequence consistency of ≥ 90% [17].

**Sequence divergence and positive selected analysis**

To explore the divergence hotspot regions that facilitate the identification of closely related species, the
mVISTA online software was used to compare the chloroplast genome of *G. paucinervi* with those of *G. mangostana*, *G. gummi-gutta*, *C. cochinchenense* and *M. ferrea*. This comparative analysis was carried out in shuffle-LAGAN mode with the annotation of *G. paucinervi* as reference, and the sequence alignment was visualized in an mVISTA plot [18]. To better visualize the expansion/contraction event in IR region, IRscope online software was used to compare the junction regions of IR-LSC and IR-SSC in the chloroplast genome of these five species [19]. *G. paucinervi* is a special plant endemic to karst environment. To detect whether plastid genes in *G. paucinervi* were under positive selection, non-synonymous (dN) and synonymous (dS) substitution rates as well as the dN/dS (ω) values of protein-coding genes were calculated in this study. All of the CDS sequences were extracted from *G. paucinervi* and another 15 species in Malpighiales, and then the single-copy CDS sequences shared by all the species were selected and aligned using the codon model. After alignments, a branch-site model and Bayesian Empirical Bayes (BEB) method were employed to test the potential positive selection of a single codon in *G. paucinervi* chloroplast genome using PAML v4.8 [20]. Selective pressure was measured by the ratio (ω): ω > 1 indicates positive selection, ω = 1 indicates neutral selection, whereas ω < 1 indicates negative selection. In the test of the branch-site model, the branch under test for positive selection was set as the foreground branch (*G. paucinervi*, *G. mangostana* and *G. gummi-gutta*), while the other branches were set as background branches (all the other tested species). Log likelihood values of branch-site models as well as the statistical significance were calculated using likelihood ratio tests (LRT). Furthermore, BEB method was used to test the posterior probability of amino acid sites to identify the positively selected sites. Finally, the amino acid sequences of positively selected genes were displayed using the PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred).

**Phylogenetic analysis**

For the phylogenetic analysis, 17 species representing different orders in the rosids lineage were recruited to construct the phylogenetic tree. Seventy-eight protein-coding genes were extracted from the chloroplast genome of the analyzed species, and then aligned using MAFFT (version 7.222) with the default parameters [21]. We concatenated these gene sequences into a supermatrix and constructed the phylogenetic tree using Maximum Likelihood (ML) method by RAxML 7.0.4 [22] with 1000 replicates under the GTR + CAT model. *Vitis vinifera* was designated as the out-group.

**Results**

**Chloroplast genome organization of *G. paucinervi***

Like traditional angiosperms, the chloroplast genome of *G. paucinervi* consists of four regions: LSC, SSC and a pair of IRs (Figure 1). The total length of the *G. paucinervi* chloroplast genome is 157702 bp. The length of the LSC region is 85989 bp, which is the longest among the four regions; the second longest is the IR region, which is 26988 bp; and the smallest is the SSC region with a length of 17737 bp (Table 1). In the whole chloroplast genome, adenine (A), thymine (T), guanine (G) and cytosine (C) account for 31.61%, 32.18%, 17.76% and 18.45%, respectively. In addition, the GC content in the IR region is 42.30%, and it is 33.70% in the LSC region, while the GC content in the SSC region is 30.40%. We also noted that the GC content of rRNA and tRNA genes is 55.6% and 53.1%, whereas the GC content of the CDS is just 36.8%. Because all the rRNA genes are only distributed in the IR regions, this is one of the reasons that contribute to the highest GC content in the IR region.

A total of 130 genes were identified in the chloroplast genome of *G. paucinervi*, including 85 protein-coding genes, 37 tRNA genes and 8 rRNA genes. Among them, there are 94 unique genes, whereas a total of 18 genes are duplicated in the chloroplast genome, including 6 CDS genes, 8 tRNA genes and 4 rRNA genes. Each gene in the chloroplast genome has a specific function [23]. There are 44 genes involved in photosynthesis, which include six genes for ATP synthase, eleven genes for NADH dehydrogenase, six genes for the cytochrome b/f complex, five genes for photosystem I, fifteen genes for photosystem II, and one gene for the large chain of Rubisco (Table 2). In addition, there are 29 genes encoding tRNA, 12 genes encoding small ribosomal proteins (SSU), 8 genes encoding large ribosomal proteins (LSU), 4 genes encoding rRNA and 4 genes encoding RNA polymerase in chloroplast genome. We also identified four hypothetical chloroplast reading frames (ycf), two genes related to enzyme synthesis (matK and clpP) and three other genes (cemA, accD and ccsA) in the chloroplast genome.

There are 18 genes with introns in the chloroplast genome of *G. paucinervi* (Table 3). Among these genes, there are 5 tRNA genes and 13 protein-coding
genes. We found that rps12 locates its first exon in the LSC region and two other exons in the IRs. This phenomenon was also found in other angiosperms [24]. The comparative analysis results showed that the trnK-UUU gene has the longest intron, which is 2552 bp in length; rpoC1 has the longest exon and its length is 1634 bp. According to statistics, except the clpP gene, which has two introns, the other 17 genes have only one intron.

We compared the chloroplast genome of G. paucinervis with that of four neighbouring species (G. mangostana, G. gummi-gutta, C. cochinchinense, M. ferrea) and found that among the five complete chloroplast genomes, the longest is M. ferrea, which is 161470 bp in length, and it is longer than that of G. paucinervis by 3850 bp. The size of the chloroplast genome among these plants is M. ferrea > G. mangostana > G. paucinervis > C. cochinchinense > G. gummi-gutta. We noted that the total GC content of the five species is similar, but M. ferrea still has the largest GC content, which is 36.42%. Then we analyzed the gene number of chloroplast genome of five species, and the results showed that there is little difference in the number of CDS, the number of tRNA genes and the number of rRNA genes (Table 1).

Figure 1. Circular map of the chloroplast genome of G. paucinervis. Genes belonging to different functional groups are highlighted in different colours. Genes on the outside of the circle are transcribed counter-clockwise, while genes on the inside are transcribed clockwise. The darker grey in the inner circle corresponds to GC content, and the lighter grey corresponds to AT content.
Table 1. Summary of the chloroplast genomes of Gaucinervis and its neighbouring species.

| Characteristics | G. paucinervis | G. mangostana | G. gummi-gutta | M. ferea | C. cochinchinense |
|-----------------|---------------|---------------|---------------|---------|------------------|
| Category Genome |               |               |               |         |                  |
| Length (bp)     | 157702        | 158179        | 156202        | 161470  | 157103           |
| A (%)           | 31.61%        | 31.66%        | 31.63%        | 31.48%  | 31.50%           |
| C (%)           | 18.45%        | 18.36%        | 18.41%        | 18.55%  | 18.36%           |
| G (%)           | 17.76%        | 17.71%        | 17.74%        | 18.78%  | 17.85%           |
| T (%)           | 32.18%        | 32.27%        | 32.20%        | 32.10%  | 32.29%           |
| GC (%)          | 36.21%        | 36.07%        | 36.15%        | 36.42%  | 36.21%           |
| Category LSC    |               |               |               |         |                  |
| Length (bp)     | 85989         | 86458         | 84998         | 88760   | 85664            |
| GC (%)          | 33.70%        | 33.50%        | 33.50%        | 34.10%  | 33.90%           |
| Length (%)      | 54.53%        | 54.66%        | 54.42%        | 54.97%  | 54.53%           |
| Category SSC    |               |               |               |         |                  |
| Length (bp)     | 17737         | 17073         | 17088         | 17482   | 18895            |
| GC (%)          | 30.40%        | 30.10%        | 37.10%        | 30.60%  | 34.90%           |
| Length (%)      | 11.25%        | 11.19%        | 10.94%        | 10.83%  | 12.03%           |
| Category IR     |               |               |               |         |                  |
| Length (bp)     | 26988         | 27009         | 27058         | 27614   | 26272            |
| GC (%)          | 42.30%        | 42.10%        | 42.10%        | 42.10%  | 42.10%           |
| Length (%)      | 17.11%        | 17.07%        | 17.32%        | 17.10%  | 16.72%           |
| Category Genes  |               |               |               |         |                  |
| Total           | 130           | 130           | 129           | 131     | 129              |
| CDS             | 85            | 85            | 85            | 86      | 84               |
| tRNA            | 37            | 37            | 36            | 37      | 37               |
| rRNA            | 8             | 8             | 8             | 8       | 8                |

Table 2. Genes in the chloroplast genome of Gaucinervis.

| Gene group | Gene name                                      |
|------------|-----------------------------------------------|
| Ribosomal RNA genes | rRNA-UGC, rRNA-GCA, rRNA-UUC, rRNA-GCC, rRNA-GUG, rRNA-CAU, rRNA-GAU, rRNA-UUU, rRNA-CAA, rRNA-UAA, rRNA-UAG, rRNA-UGG, rRNA-UUG, rRNA-UCU, rRNA-GCU, rRNA-GGA, rRNA-UGA, rRNA-GGU, rRNA-UGU, rRNA-UCG, rRNA-CAU |
| Transfer RNA genes | psaA, psaB, psaC, psaD, psaE, psaF, psaG, psaH, psaI |
| Photosystem I | psaA, psaB, psaC, psaD, psaE, psaF, psaG, psaH, psaI |
| Photosystem II | psaA, psaB, psaC, psaD, psaE, psaF, psaG, psaH, psaI |
| Cytochrome b/f complex | petA, petB, petC, petD, petE, petG, petN |
| ATP synthase | atpA, atpB, atpC, atpD, atpE, atpF, atpH, atpI |
| Rubisco large subunit | rbcL |
| NADH dehydrogenase | ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK |
| ATP-dependent protease | ctpP |
| Ribosomal protein genes (larger subunit) | rps12, rps14, rps16, rps20, rps22, rps23, rps33, rps36 |
| Ribosomal protein genes (smaller subunit) | rps12, rps14, rps16, rps20, rps22, rps23, rps33, rps36 |
| RNA polymerase | rpoA, rpoB, rpoC1, rpoC2 |
| Maturase | matK |
| hypothetical chloroplast reading frames (ycf) | ycf1, ycf2, ycf3, ycf4 |
| other genes | cemA, accD, ccsA |

Codon usage and RNA editing in protein coding genes

Codon is a sequence of three nucleotides which together form a unit of genetic code in DNA or RNA molecule. Although one codon can only correspond to a specific amino acid, a specific amino acid can correspond to more than one codon [25]. In this study, we found that there are 23337 codons (including stop codons) in the chloroplast genome of Gaucinervis and 20 amino acids are translated. According to statistics, leucine is the most abundant amino acid and it takes up 2456 codons, followed by isoleucine with 2054 codons. The least prevalent amino acid is cysteine, which only takes up 268 codons. Subsequently, we used MEGA5 software to analyze the relative synonymous codon usage (RSCU) in the Gaucinervis chloroplast genome. In principle, when the RSCU value of a codon is > 1, it is considered to be a relatively frequently used codon. In contrast, an RSCU value < 1 means that the codon is a relatively less used codon. The RSCU value equal to 1 indicates there is no preference for codon use [26]. We found that there are 64 types in these 23337 codons. Among them, 30 types of codons have RSCU values > 1, and these codons all end with A or U except UUG (Table 4). There are 32 types of codons whose RSCU values are < 1, and most of these codons end with G or C. In addition, the RSCU values of AUG and UGG are equal to 1,
which suggests that the two types of codons have no bias in expression.

Generally, the initiation codon of the protein coding gene is ATG, but we found some differences when checking all the CDS. We found that the initiation codon of the \textit{ndhD} gene is ACG rather than ATG. At first we thought it is a special case of RNA editing, which can convert ACG into a standard ATG initiation codon. As expected, some previous studies have conducted related experiments on whether ACG could serve as the initiation codon, and their results demonstrated that the chloroplast translator accepts ACG use as an initiation codon [27]. Although the protein-coding genes seldom use ACG as the start codon, there are still many genes that use ACG as the initiation codon for translation [28]. Therefore, it is not a sequencing error in this case that the \textit{ndhD} gene uses ACG as the initiation codon.

### Table 3. Summary of genes with introns in \textit{G. paucinervis} chloroplast genome.

| Gene        | Location | Exon I (bp) | Intron I (bp) | Exon II (bp) | Intron II (bp) | Exon III (bp) |
|-------------|----------|-------------|---------------|--------------|----------------|---------------|
| trnG-UCC    | LSC      | 23          | 717           | 48           |                |               |
| trnK-UUU    | LSC      | 35          | 2552          | 37           |                |               |
| trnL-UAA    | LSC      | 35          | 640           | 50           |                |               |
| trnV-UAC    | LSC      | 37          | 597           | 37           |                |               |
| atpF        | LSC      | 399         | 748           | 144          |                |               |
| clpP        | LSC      | 228         | 641           | 291          | 748            | 69            |
| petB        | LSC      | 6           | 841           | 642          |                |               |
| petD        | LSC      | 7           | 25            | 524          |                |               |
| rpl16       | LSC      | 399         | 43            | 9            |                |               |
| rpoC1       | LSC      | 1634        | 753           | 430          |                |               |
| rps16       | LSC      | 224         | 853           | 40           |                |               |
| ycf3        | LSC      | 387         | 729           | 126          |                |               |
| rps12       | LSC/IR   | 114         | 233           | 538          | 25             |               |
| trnA-UGC    | IR       | 38          | 801           | 35           |                |               |
| trnI-GAU    | IR       | 37          | 943           | 35           |                |               |
| ndhB        | IR       | 756         | 699           | 777          |                |               |
| rpl2        | IR       | 434         | 669           | 400          |                |               |
| ndhA        | SSC      | 534         | 1148          | 561          |                |               |

### Table 4. The codon–anticodon recognition pattern and codon usage in the chloroplast genome of \textit{G. paucinervis}.

| Codon | Amino acid | Count | RSCU | Codon | Amino acid | Count | RSCU |
|-------|------------|-------|------|-------|------------|-------|------|
| UUU (F) | Phe | 956 | 1.38 | UCU (S) | Ser | 490 | 1.67 |
| UUC (F) | Phe | 427 | 0.62 | UCC (S) | Ser | 261 | 0.91 |
| UUA (L) | Leu | 869 | 2.12 | UCA (S) | Ser | 363 | 1.23 |
| UUG (L) | Leu | 474 | 1.15 | UCG (S) | Ser | 166 | 0.57 |
| CUU (L) | Leu | 537 | 1.31 | AGU (S) | Ser | 348 | 1.18 |
| CUC (L) | Leu | 134 | 0.33 | AGC (S) | Ser | 126 | 0.44 |
| CUA (L) | Leu | 308 | 0.75 | CCC (P) | Pro | 395 | 1.67 |
| CUG (L) | Leu | 134 | 0.34 | CCC (P) | Pro | 166 | 0.69 |
| AUU (I) | Ile | 1043 | 1.52 | CCA (P) | Pro | 253 | 1.08 |
| AUG (M) | Met | 537 | 1 | ACC (T) | Thr | 498 | 1.69 |
| GUU (V) | Val | 474 | 1.49 | ACA (T) | Thr | 379 | 1.3 |
| GUC (V) | Val | 158 | 0.5 | AGC (T) | Thr | 111 | 0.37 |
| GUA (V) | Val | 466 | 1.46 | GCC (A) | Ala | 585 | 1.87 |
| GUG (V) | Val | 174 | 0.55 | GGC (A) | Ala | 197 | 0.63 |
| GCA (A) | Ala | 340 | 1.08 | UGU (C) | Cys | 205 | 1.51 |
| GCC (A) | Ala | 134 | 0.42 | UGC (C) | Cys | 63 | 0.49 |
| UAA (*) | stop codon | 39 | 1.6 | UGA (*) | stop codon | 16 | 0.62 |
| UAG (*) | stop codon | 24 | 0.78 | UGG (W) | Trp | 403 | 1 |
| CAU (H) | His | 427 | 1.53 | CGU (R) | Arg | 284 | 1.27 |
| CAC (H) | His | 126 | 0.47 | CGC (R) | Arg | 87 | 0.38 |
| CAA (Q) | Gln | 664 | 1.61 | CCA (R) | Arg | 332 | 1.49 |
| CAG (Q) | Gln | 166 | 0.39 | CGG (R) | Arg | 95 | 0.44 |
| AAU (N) | Asn | 916 | 1.59 | AGA (R) | Arg | 411 | 1.86 |
| AAC (N) | Asn | 237 | 0.41 | AGG (R) | Arg | 126 | 0.55 |
| AAA (K) | Lys | 995 | 1.58 | UAU (Y) | Tyr | 719 | 1.68 |
| AAG (K) | Lys | 284 | 0.44 | UAC (Y) | Tyr | 142 | 0.32 |
| GAU (D) | Asp | 719 | 1.6 | GGU (G) | Gly | 498 | 1.27 |
| GAC (D) | Asp | 182 | 0.4 | GGC (G) | Gly | 190 | 0.48 |
| GAA (E) | Glu | 995 | 1.57 | GGA (G) | Gly | 632 | 1.6 |
| GAG (E) | Glu | 269 | 0.43 | GGG (G) | Gly | 261 | 0.66 |

Note: RSCU is Relative synonymous codon usage.
To examine whether there are any RNA editing sites in other genes, we used the PREP suite to perform RNA editing prediction for all the protein-coding genes in the *G. paucinervis* chloroplast genome by setting the first codon position of the first nucleotide. Consequently, a total of 56 RNA editing sites are identified in the *G. paucinervis* chloroplast genome (Supplemental Table S1), and these sites were changed from C-to-U. The *ndhB* gene has 11 RNA editing sites, and this is the gene that contains the largest number of RNA editing sites among all the protein-coding genes. The second is the *ndhD* gene, which contains 10 RNA editing sites. Further analysis found that most of the RNA editing events occurred to change polar amino acids to non-polar amino acids. However, each of the genes *matK*, *petD*, *ndhD* and *psbE*, has an RNA editing site where conversion changed from non-polar amino acids to polar amino acids.

**Repeat sequence analysis**

SSRs are tandemly repeated DNA sequences widely distributed in eukaryote genomes [29]. Nowadays, SSRs have become one of the robust molecular markers for plant genetic analysis. Through statistical analysis, a total of 241 SSRs were found in the chloroplast genome of *G. paucinervis* (Figure 2A). Among the 241 simple sequence repeats, there are 177 mononucleotide repeats, 48 dinucleotide repeats, 5 trinucleotide repeats, 10 tetranucleotide repeats and one pentanucleotide, but no hexanucleotides were found. Among all the mononucleotide repeats, the amount of A/T is much higher than that of G/C, and it is also the highest among all SSRs, accounting for 69.71% of all the SSRs. This result is consistent with the fact that the complete chloroplast genome of *G. paucinervis* is rich in A/T. In all SSRs, the most frequent repeats are A/T, followed by AT/AT, AG/CT, G/C, AAAT/ATTT, AAT/ATT and AAAG/CTTT. We also found that there are...
four SSRs that appear only once. They are AC/GT, AATT/ACAG/CTGT and ACTAT/AGTAT. Subsequently, we conducted a comparative analysis of SSRs in *G. paucinervis*, *G. mangostana*, *G. gummi-gutta*, *C. cochinchinense* and *M. ferrea*. The results showed that *C. cochinchinense* has the largest number of SSR, which is 275. *M. ferrea* has the least number of SSR among the five species (Figure 2B). In addition, *G. mangostana* has the largest number of mononucleotide repeats, while *C. cochinchinense* has the largest number of dinucleotide repeats. Both *G. mangostana* and *M. ferrea* have seven trinucleotide repeats, and they also have the most trinucleotide repeats among the five species. *C. cochinchinense* and *M. ferrea* have the largest number of tetranucleotide and pentanucleotides repeats, respectively. Notably, ten hexanucleotide repeats were found in the *C. cochinchinense* chloroplast genome, but not in the other four species. The identified SSRs could be used to evaluate the genetic diversity and population structure pattern of these species.

In addition to SSRs, we also explored the long complex repetitive sequences including forward repeats, reverse repeats, palindromic repeats and complement repeats in the *G. paucinervis* chloroplast genome. Repetitive regions are considered to play an important role in genome recombination and rearrangement, which could lead to the variation of species. A total of 49 repetitive structures were found in the *G. paucinervis* chloroplast genome, including 17 forward repeats, 26 palindromes, five reverse repeats and one complement (Figure 3). Among them, there are 29 repetitive sequences (15 palindromic repeats, 8 forward repeats, 5 reverse repeats and one complement repeat) in the LSC region, 5 repetitive sequences (three palindromic repeats and two forward repeats) in the SSC region, and 15 repetitive sequences (8 palindromic repeats and 7 forward repeats) in the two IR regions. We also analyzed the repetitive sequences of four neighbouring species (*G. mangostana*, *G. gummi-gutta*, *C. cochinchinense* and *M. ferrea*), and identified a total of 49, 49, 49, 49 repeat structures in these species, including 19, 19, 25, 20 forward repeats, 27, 27, 24, 28 palindromic repeats, 3, 3, 0, 1 reverse repeats and 0, 0, 0, 0 complement repeats, respectively.

**Sequence divergence among different species**

The IR expansion and contraction have contributed to the size difference of chloroplast genome and are considered as a major factor affecting plant evolution. We compared the IR boundaries of five species using IRscope online software. There are obvious differences in the SC/IR boundaries among the five chloroplast genomes (Figure 4). In combination with the annotation files of chloroplast genomes, we found that there are four genes located at or adjacent to the boundaries of LSC/IR and SSC/IR, including *ndhF*, *ycf1*, *rps19* and *trnH*, and these genes vary in size and location among different species. Although all of the chloroplast genomes of these species have two *rps19* genes, only *M. ferrea* conserved its *rps19* gene entirely in IR regions, and separated from the IR/LSC region by 55 bp. The IR region of other four species contracted into *rps19* gene, resulting in a pseudogene fragment.
with lengths of 219-220 bp at the IRa/LSC border. In addition, although the rps19 gene of G. mangostana also crossed the IRb/LSC junction, some deletion of gene fragments has occurred in this gene obviously, resulting in approximately 50 bp shorter length than that in other species. Similarly to rps19, the SSC/IRa border was located in the ycf1 gene, creating a long ycf1 pseudogene fragment in all of the species. The length of the ycf1 pseudogene was 1419 bp in G. paucinervis, 1421 bp in G. mangostana and G. gummi-gutta, 1908 bp in M. ferrea and 1298 bp in C. cochinchinense. Besides, the ycf1 pseudogene and the ndhF gene overlapped in G. gummi-gutta chloroplast genome by 1 bp, while in the other four species, the ndhF gene was located entirely at the SSC region. We also noted that the trnH gene crossed the LSC/IRa border and overlapped with the rps19 pseudogene in G. paucinervis, G. mangostana and C. cochinchinense. Besides, the ycf1 pseudogene and the ndhF gene was located entirely at the SSC region. We also noted that the trnH gene crossed the LSC/IRa border and overlapped with the rps19 pseudogene in G. paucinervis, G. mangostana and C. cochinchinense, while it was fully located in the IRa region in M. ferrea 137 bp away from the IRa/LSC border. As indicated above, we can conclude that the variations of SC/IR boundaries have led to the length variation of the whole chloroplast genome.

Highly variable regions in coding and non-coding regions together constitute potential DNA barcodes or molecular markers, which could provide basis for phylogenetic analysis and species identification [24]. To identify the high divergence sequences, we used mVISTA online software to analyze the variation in chloroplast genome sequences of five species. The alignment showed that the five chloroplast genome sequences were relatively conservative, but there are some regions that exhibited high sequence divergence (Figure 5). The sequences in the LSC and SSC regions were more divergent than the sequences in the IR regions. This result is due to the conservative characteristics of rRNA genes, which are all distributed in the IR region. In addition, the sequences of the non-coding regions are also more divergent than that of the coding region. The most divergent non-coding regions among these chloroplast genomes were trnH-psbA, trnT-psbD, psaA-ycf3, ndhC-trnV and ndhD-ccsA.

**Survey for positive selection in G. paucinervis chloroplast genome**

In the chloroplast genome, there are some divergences in protein-coding genes under the influence of natural environment. To examine whether those divergent sites were under positive selection, we used the branch-site model implemented in PAML [20]. Of the 76 protein-coding genes we tested, signals of positive selection were detected in 5 genes (matK, psbK, rps4, rps12 and rps19) with $p < 0.05$ and a high posterior probability in the BEB test (Table 5), indicating that these genes maybe play a key role in the adaptive evolution of G. paucinervis. In the branch-site model, one codon (223rd) in the matK gene was under positive selection with a BEB posterior probability larger than 0.99. The alignment results of the amino acid sequences revealed that although many amino acids vary among different members of different families, the 223rd codon encoding valine (V) was specific for G. paucinervis but instead encoded phenylalanine (F) in other families. The secondary structure of the matK protein predicted by PSIPRED server showed that this positively selected site was located in the $\alpha$-helix (Figure 6). In the rps12 gene, the site under positive selection was the codon encoding the 25th amino acid, which is located in the random coil and is valine.
(V) for *G. paucinervis*, while glycine (G) in the species from other families (Supplemental Figure S1). In addition, we observed that a positively selected site changed from glutamic acid (E) to threonine (T) for *G. paucinervis* in *rps19* (Supplemental Figure S2). Interestingly, *rps4* was found to harbour two sites under positive selection, which were the 22nd codon encoding alanine (A) and the 117th codon encoding phenylalanine (F) for *G. paucinervis*, and these sites were located in the random coil and β-sheet respectively (Supplemental Figure S3). In the species from other families, the 22nd codon and the 117th codon

**Table 5.** Identification of potential positive selection sites based on branch-site model.

| Gene   | Null Hypothesis | Alternative Hypothesis | Significance Test |
|--------|-----------------|------------------------|-------------------|
|        | InL             | Site class             | BEB               |
|        | Site class      | 2ΔI                    | NEB               |
|        | Site class      | p-Value                |                   |
| matK   | -10714.832783   | α = 1                  | 223F, 0.995**     |
|        | -10709.207412   | α = 1                  | 221I, 0.997**     |
|        | -0.625371**     | α1 = 1; α2 = 999.00000; α2b = 999.00000 | 223F, 1.000**     |
|        |                 | 0.000796**             |                   |
| psbK   | -775.854912     | α = 1                  | 41V, 0.986*       |
|        | -773.062359     | α = 1                  | 41V, 0.999**      |
|        | -2.792553       | α0 = 0.13469; α1 = 1;  | 0.018114          |
|        |                 | α2a = 91.32454; α2b = 91.32454 |                   |
| rps4   | -2551.113076    | α = 1                  | 22T, 0.973*       |
|        | -2548.369839    | α = 1                  | 22T, 0.967*       |
|        | -2.743237       | α0 = 0.06962; α1 = 1;  | 0.019164          |
|        |                 | α2a = 15.57568; α2b = 15.57568 |                   |
| rps12  | -1125.39382     | α = 1                  | 117L, 0.935       |
|        | -1121.751811    | α = 1                  | 117L, 0.956**     |
|        | -3.642009*      | α0 = 0.02024; α1 = 1;  | 0.006957**        |
|        |                 | α2a = 9.0458; α2b = 9.0458 |                   |
| rps19  | -1368.29067     | α = 1                  | 25G, 0.988*       |
|        | -1366.040923    | α = 1                  | 25G, 0.993**      |
|        | 2.249747        | α0 = 0.11413; α1 = 1;  | 0.033905          |
|        |                 | α2a = 12.20125; α2b = 12.20125 |                   |

Note: The degree of freedom (df) = 1 for the comparison of Null model and Alternative model was used in this study; InL: the log-likelihood difference between the two models; 2ΔI: twice the log-likelihood difference between the two models. BEB: Bayes Empirical Bayes; NEB: Naive Empirical Bayes. The positive selected sites were identified with posterior probability. For positive selected sites, posterior probabilities (PP) >90% (>95% indicated with asterisks, >99% indicated with double asterisks) are shown.
encoded threonine (T) and leucine (L), respectively. Furthermore, we also detected one codon site under positive selection in \textit{psbK}, which encodes a protein that directly participates in the photosystem. The observed site was located in the \(\alpha\)-helix, and encoded phenylalanine (F) specific for \textit{G. paucinervis} (Figure S4), whereas valine (V) was found in the species from other families. It is worth mentioning that most of the positively selected genes are involved in ribosome biogenesis, protein synthesis and other developmental processes.

**Phylogenetic analysis**

To study the phylogenetic position of \textit{G. paucinervis} within the rosids lineage, we used 78 protein-coding genes shared by the chloroplast genomes of three species in \textit{Garcinia} and 14 species from other orders in the rosids lineage to construct an ML tree with \textit{V. vinifera} as an outgroup. The resulting topology was in accordance with previous phylogeny research [30], suggesting that the phylogenetic tree is reliable. Importantly, the phylogenetic tree revealed that \textit{G. mangostana} was a sister to \textit{G. gummi-gutta}, and they contributed to a diverging lineage to \textit{G. paucinervis} in \textit{Garcinia}, which is a genus of Garcinaceae in the order Malpighiales (Figure 7). In addition, the phylogenetic tree continues to support a sister relationship between the Malpighiales and Celastrales, and they were positioned close to Oxalidales. These analyses substantially increase our understanding of the evolutionary relationship among species in the rosids lineage.

**Discussion**

The chloroplast genome is known to be highly conserved in gene content and gene order, but the sequence changes in the chloroplast genome, such as gene mutation, repeat sequence rearrangement and the contraction/expansion of IR region border, provide abundant taxonomic data. Up to now, there have been many studies carried out to evaluate the genetic diversity and genetic structure of the endangered plants using DNA markers derived from the chloroplast genome [31–33], demonstrating that the chloroplast genome could provide valuable genetic information for formulating efficient conservation
strategies for plant species. In this study, we characterized the complete chloroplast genome of *G. paucinervis*, an endangered evergreen tree (Figure 1). Combined with the previously reported chloroplast genomes of four species, we analyzed the differences between these species’ chloroplast genomes. Importantly, we mined 241 SSRs in the chloroplast genome of *G. paucinervis*, and most of these SSRs were also found in the comparative species but there was a difference in their length. We noted that the majority of di- to hexa-nucleotides are AT-rich in all species. This phenomenon is consistent with the SSR characteristics of chloroplast genomes in angiosperms [34]. All of these SSRs, together with the aforementioned complex repeat sequences, are informative sources for developing markers for genetic diversity studies of *G. paucinervis*.

RNA editing is a post transcriptional process, which involves the insertion, deletion or conversion of nucleic acid bases in the coding region of the transcribed RNA. The “gene” produced by RNA editing is also called cryptogene, which means that the structure of the product cannot be deduced from the genomic DNA sequence [35, 36]. RNA editing is thought to be an important event during the course of plant evolution, since it could introduce the start codon or stop codon in the transcribed RNA to modify the gene expression pattern, and it also increases the amount of coding information of peptide chain; this post-translational modification ultimately confers plants a flexible capacity to respond to the complex environment. Previous studies found that different plant species show a variant degrees of losses in RNA editing sites during evolution [37–39]. In *G. paucinervis*, a total of 56 RNA editing sites were identified, and all these sites were changes from C-to-U, resulting in the conversion of 12 amino acids (Supplemental Table S1); this result is consistent with previous studies that the C-to-U conversion is the most predominant form in chloroplasts [40, 41]. Although U-to-C changes and G-to-A changes were found in other plants [41], we did not detect any of these changes in the *G. paucinervis* chloroplast. It is possible that these kinds of RNA editing sites were gradually lost in the course of evolution. Further verification on RNA sequencing of chloroplasts will be needed to support this hypothesis. In addition, out of these amino acids, the serine (S) to leucine (L) change was the most common (16/56), followed by serine (S) to phenylalanine (F) change and proline (P) to leucine (L) change (Supplemental Table S1). The *ndhB* gene has 11 RNA editing sites, which is the gene that contains the largest number of RNA editing sites among the protein-coding genes. This result is also found in *Arabidopsis*, tobacco, tomato, rice and maize [42, 43]. The *ndhD* is the second most extensively edited gene, which contains 10 RNA editing sites. The *ndh* genes are known to act as electro-regulators to control the accumulation of reactive oxygen species in cyclic photosynthesis electron transporters [44]. Since *G. paucinervis* is a drought-tolerant species that grows in karst environments, the RNA editing process in *ndh* genes may help *G. paucinervis* to more nimbly adjust photosynthesis efficiency when suffering from drought stress. Though the function of RNA editing in the chloroplast has not yet been clearly elucidated, the identified RNA editing sites in our study could help to better understand the physiological behavior of *G. paucinervis*.

*G. paucinervis* is a plant endemic to tropical and sub-tropical karst environments. As we know, karst topography is characterized by very poor soil content, low water storage capacity, insufficient light, and nutrient deficiency; these severe conditions could impose a strong selective pressure on plant evolution [45]. To adapt to the extreme habitats, *G. paucinervis* has evolved with special morphological traits. For example, this plant harbours rigid leaves with thick cuticle to reduce water transpiration; this change may affect the function of some genes in the chloroplast related to photosynthesis and photorespiration [46]. In an attempt to identify whether there exists any positive selection in the chloroplast genes, we identified 5 genes (*matK, psbK, rps4, rps12 and rps19*) under positive selection using the branch-site model in PAML. One codon (223rd) in the *matK* gene showed a positive selection signal with a BEB posterior probability larger than 0.99, and this site corresponded to an amino acid located in the α-helix of the encoded protein. The gene *matK* encodes maturase K, which is involved in group II intron splicing and is essential for RNA processing. Previous studies have proposed that some genes, including *atpF, rpl2, rps12, trnA, trnI* and *trnK*, require *matK* for intron excision in their transcripts, and the protein products from these genes are essential for normal photosynthesis [47], indicating that *matK* is an indispensable post transcription splicing factor in the chloroplast. Due to its fast substitution rate, *matK* is also widely used to construct the diverse phylogenetic relationships in higher plants. In addition to *matK*, we also identified one site under positive selection in *psbK*, and this site also corresponded to an amino acid located in the α-helix. Interestingly, previous studies did not observe any positive selection signals in this gene in other plants,
so this is probably a recent adaptive evolutionary feature in *G. paucinervis*. Gene *psbK* encodes a protein that directly participates in photosynthesis. Its protein product is one of the components of the core complex of photosystem II (PSII) which uses light energy to abstract electrons from H2O, generating O2 and a proton gradient subsequently used for ATP formation. It consists of a core antenna complex that captures photons, and an electron transfer chain that converts photonic excitation into a charge separation [48]. The site identified here indicates that the photosystem of *G. paucinervis* was indeed affected by natural selective forces. *rps4*, *rps12* and *rps19* belong to ribosomal protein genes, which encode important proteins for the small ribosome subunit. Previous studies reveal that *rps* genes encode key players in the decoding center of chloroplast ribosomes and loss of function of a single *rps* gene could lead to the arrest of cell growth and chloroplast translation [49]. In this respect, the positive selection sites found in the three ribosomal protein genes in the *G. paucinervis* chloroplast genome could be regarded and further analyzed in a broader perspective. Overall, the positive selection sites found in the five genes reflect a complex adaptive strategy changes in the photosystem likely associated with the adaptation of *G. paucinervis* to the extreme habitats.

**Conclusions**

In this study, we characterized the complete chloroplast genome of *G. paucinervis* for the first time. The genome annotation, and comparative analysis showed that the chloroplast genome of *G. paucinervis* has a typical quadripartite structure like traditional angiosperms, and the GC content, gene number and codon usage features are similar to those of other species in Garcinaceae. The phylogenetic tree supports *G. paucinervis* belonging to *Garcinia* in the Clusiaceae family. These data will enlarge our understanding of the evolutionary history of Clusiaceae species. Notably, the SSRs explored here are informative sources for developing markers for genetic diversity studies of *G. paucinervis*. The adaptive evolution analysis detected signs of positive selection in five chloroplast genes (i.e. *matK*, *psbK*, *rps4*, *rps12* and *rps19*). The change of these adaptive sites may be associated with the evolution of *G. paucinervis* to adapt to the extreme habitats in karst enviroments. Hence, this information will help us to develop an effective strategy for the conservation of *G. paucinervis*.

**Acknowledgements**

We are sincerely grateful to Yusong Huang (Guangxi Institute of Botany, The Chinese Academy of Sciences, Guilin, China) for providing plant picture.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This work was supported by the National Science Foundation of China (31960276), the National Science Foundation of Guangxi (2018GXNSFAA281015), and the Guilin Medical University Start-up fund (20501019034).

**Data availability statement**

The data generated during this study is available at Genbank (accession number: MT501656)

**References**

[1] Hu G, Zhang ZH, Yang P, et al. Development of microsatellite markers in *Garcinia paucinervis* (Clusiaceae), an endangered species of karst habitats. Appl Plant Sci. 2017;5(1):1600131.

[2] Hemshekhar M, Sunita K, Santhosh MS, et al. An overview on genus *Garcinia*: phytochemical and therapeutic aspects. Phytochem Rev. 2011; 10(3): 325–351.

[3] Wang S, Xie Y. China Species Red List, 366. Beijing, China: Higher Education Press; 2004.

[4] Ovalle-Magallanes B, Eugenio-Pérez D, Pedraza-Chaverri J. Medicinal properties of mangosteen (*Garcinia mangostana* L.): A comprehensive update. Food Chem Toxicol. 2017;109(Pt 1):102–122.

[5] Nie L, Cui Y, Chen X, et al. Complete chloroplast genome sequence of the medicinal plant Arctiumlappa. Genome. 2020; 63(1):53–60.

[6] Sabater B. Evolution and function of the chloroplast. Current investigations and perspectives. JIMMS. 2018; 19(10):3095.

[7] Zha X, Wang X, Li J, et al. Complete chloroplast genome of *Sophora alopecuroides* (Papilionoideae): molecular structures, comparative genome analysis and phylogenetic analysis. J Genet. 2020;99(1):30.

[8] Cheon KS, Kim KA, Kwak M, et al. The complete chloroplast genome sequences of four Viola species (Violaceae) and comparative analyses with its congenic species. PLoS One. 2019; 14(3):e0214162

[9] Zhou T, Ruhsam M, Wang J, et al. The complete chloroplast genome of *Euphrasia regelli*, pseudogenization of ndh genes and the phylogenetic relationships within. Orobancheae Front Genet. 2019; 10: 444.

[10] Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987; 19:11–15.
[21] Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19(5):455–477.

[22] Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24(8):1586–1591.

[23] Mayor C, Brudno M, Schwartz JR, et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 2000; 16(11):1046–1047.

[24] Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 2009; 37(Web Server):W253–W259.

[25] Kurtz S, Choudhuri JV, Ohlebusch E, et al. REPuter: software to visualize the junction sites of chloroplast genomes. Bioinformatics. 2010; 26(11):1391–1392.

[26] Zhao K, Chu X. G-BLASTN: accelerating nucleotide alignment by graphics processors. Bioinformatics. 2014;30(10):1384–1391.

[27] Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20(2):265–272.

[28] Tillich M, Lehwalk P, Pellizzer T, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017; 45(W1):W6–W11.

[29] Mower JP. JP. Modeling sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria and chloroplasts. Plant Mol Biol. 1996; 26(8):617–626.

[30] Naomi HE, Staton ME, Dattilo AJ, et al. Population structure and genetic diversity within the endangered species Pitypsyis ruthii (Asteraceae). Front Plant Sci. 2018; 49:943.

[31] Bock R. Surprising features of secondary structure. J Mol Biol. 1998; 281(1):31–48.

[32] Suwabe K, Tsukazaki H, Iketani H, et al. Simple sequence repeat-based comparative genomics between Brassica rapa and Arabidopsis thaliana: the genetic origin of clubroot resistance. Genetics. 2006; 173(1):309–319.

[33] Li HT, Yi TS, Gao LM, et al. Origin of angiosperms and the puzzle of the Jurassic gap. Nat Plants. 2019;5(5): 461–470.

[34] Oresic M, Shalloway D. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol. 1998; 281(1):31–48.

[35] Sasaki T, Yukawa Y, Miyamoto T, et al. Identification of RNA editing sites in chloroplast transcripts from the maternal and paternal progenitors of tobacco (Nicotiana tabacum): Comparative analysis shows the involvement of distinct trans-factors for ndhB editing. Mol Biol Evol. 2003; 20(7):1028–1035.

[36] Mower JP. Prediction of sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria and chloroplasts. Mol Biol Evol. 2003; 20(7):1028–1035.

[37] Ebert D, Peakall R. Chloroplast simple sequence repeats (cpSSRs): Technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour. 2009; 9(3):673–690.

[38] Adams KL, Qiu YL, Stoutemyer M, et al. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A. 2002; 99(15):9905–9912.

[39] Yu J, Song J, Gao H, et al. colleagues. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PloS One. 2013; 8(2):e57607.

[40] Maier RM, Zelt P, Kossel H, et al. RNA editing in plant mitochondria and chloroplasts. Plant Mol Biol. 1996; 32(1-2):343–365.

[41] Pock R. Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing. Biochimie. 2000; 82(6-7):549–557.

[42] Tangphatsornruang S, Uthaipaisanwong P, Sangsrakru D, et al. Characterization of the complete chloroplast genome of Hevea brasiliensis reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene. 2011;475(2):104–112.

[43] Kahlu S, Aspinall S, Gray JC, et al. Sequence of the tomato chloroplast DNA and evolutionary comparison of Solanaceous plastid genomes. J Mol Evol. 2006; 63(2):194–207.

[44] Sasaki T, Yukawa Y, Tatsutani H, et al. Identification of RNA editing sites in chloroplast transcripts from the maternal and paternal progenitors of tobacco (Nicotiana tabacum): Comparative analysis shows the involvement of distinct trans-factors for ndhB editing. Mol Biol Evol. 2003; 20(7):1028–1035.
Ivanova Z, Sablok G, Daskalova E, et al. Chloroplast Genome Analysis of Resurrection Tertiary Relict Haberlea rhodopensis Highlights Genes Important for Desiccation Stress Response. Front Plant Sci. 2017; 8:204.

Xie DF, Yu Y, Deng YQ, et al. Comparative analysis of the chloroplast genomes of the Chinese endemic genus Urophysa and their contribution to chloroplast phylogeny and adaptive evolution. UMS. 2018; 19(7):1847.

Zhao B, Li JJ, Yuan RW, et al. Adaptive evolution of the rbcL gene in the genus Rheum (polygonaceae.). Biotechnol Biotec Eq. 2017; 31(3):493–498.

Barthet MM, Hilu KW. Expression of matK: functional and evolutionary implications. Am J Bot. 2007; 94(8): 1402–1412.

UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43(Database issue):D204-212. Available link: https://www.uniprot.org/uniprot/P60147.

Ramundo S, Rahire M, Schaad O, et al. Repression of essential chloroplast genes reveals new signaling pathways and regulatory feedback loops in chlamydomonas. Plant Cell. 2013;25(1):167–186.