RESEARCH ARTICLE

Complete chloroplast genomes of two *Siraitia* Merrill species: Comparative analysis, positive selection and novel molecular marker development

Hongwu Shi[1]☯, Meng Yang[1]☯, Changming Mo[2], Wenjuan Xie[3], Chang Liu[1], Bin Wu[1]*, Xiaojun Ma[1]*

1 Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, **2** Guangxi Crop Genetic Improvement and Biotechnology Laboratory, Nanning, China, **3** Guilin Medical University, Guilin, China

☯ These authors contributed equally to this work.
* bwu@implad.ac.cn (BW); mayixuan10@163.com (XM)

OPEN ACCESS

Citation: Shi H, Yang M, Mo C, Xie W, Liu C, Wu B, et al. (2019) Complete chloroplast genomes of two *Siraitia* Merrill species: Comparative analysis, positive selection and novel molecular marker development. PLoS ONE 14(12): e0226865. https://doi.org/10.1371/journal.pone.0226865

Editor: Shashi Kumar, International Centre for Genetic Engineering and Biotechnology, INDIA

Received: June 1, 2019

Accepted: December 5, 2019

Published: December 20, 2019

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0226865

Data Availability Statement: The chloroplast genome sequencing data were deposited in GenBank under accession numbers MK755853 and MK755854, respectively.

Abstract

Siraitia grosvenorii fruit, known as Luo-Han-Guo, has been used as a traditional Chinese medicine for many years, and mogrosides are its primary active ingredients. Unfortunately, *Siraitia siamensis*, its wild relative, might be misused due to its indistinguishable appearance, not only threatening the reliability of the medication but also partly exacerbating wild resource scarcity. Therefore, high-resolution genetic markers must be developed to discriminate between these species. Here, the complete chloroplast genomes of *S. grosvenorii* and *S. siamensis* were assembled and analyzed for the first time; they were 158,757 and 159,190 bp in length, respectively, and possessed conserved quadripartite circular structures. Both contained 134 annotated genes, including 8 rRNA, 37 tRNA and 89 protein-coding genes. Twenty divergences (Pi > 0.03) were found in the intergenic regions. Nine protein-coding genes, *accD*, *atpA*, *atpE*, *atpF*, *clpP*, *ndhF*, *psbH*, *rbcL*, and *rpoC2*, underwent selection within Cucurbitaceae. Phylogenetic relationship analysis indicated that these two species originated from the same ancestor. Finally, four pairs of molecular markers were developed to distinguish the two species. The results of this study will be beneficial for taxonomic research, identification and conservation of *Siraitia* Merrill wild resources in the future.

Introduction

Siraitia plants are important perennial vines belonging to the fourth most economically important plant family, Cucurbitaceae, and the genus has been widely cultivated as economic crops in southern China and northern Thailand [1]. Among these crops, *S. grosvenorii*, a traditional Chinese medicinal plant native to Guangxi, China, has been cultivated for approximately 200 years. The fruit of *S. grosvenorii*, Luo-Han-Guo, has been used as a traditional

Funding: This work was supported by Beijing Municipal Natural Science Foundation (5172028), the National Natural Science Foundation of China (81373914, 81573521), and CAMS Innovation Fund for Medical Sciences (CIFMS) (No. 2017-I2M-1-013).

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

Chinese medicine for the treatment of lung congestion, cold, and sore throat [2,3]. The primary active ingredients of *S. grosvenorii* are mogrosides, which are a class of cucurbitane-type triterpenoids, including mogrosides IV, V and VI. Modern pharmacology studies have shown that mogrosides have antidiabetic, antioxidative and anti-inflammatory effects [4,5]. Mogrosides are also natural zero-calorie sweeteners and have been used as sugar substitutes; mogroside V is 250 times sweeter than sucrose [6]. Compared to *S. grosvenorii*, *S. siamensis* has advantages in disease resistance and fruit set percentage [7]. Siamenoside I, a kind of mogroside, was separated from the fruit of *S. siamensis* and is approximately 560 times sweeter than sucrose [8], and is about 1.4 fold sweeter than aspartame. In recent years, people have paid more attention to developing and utilizing *Siraitia* germplasm resources because of the importance of mogrosides in sweetener development.

Many studies have focused on the improvement of cultivated varieties and excavation of the potential medicinal value of medicinal plants [9–11], as well as identification of the wild species [12,13]. Among *Siraitia* plant materials, only *S. grosvenorii* fruit has been stipulated to have medicinal use and is listed in the latest edition of the Chinese Pharmacopoeia [14]. Thus, indiscriminate use of wild relatives, such as *S. siamensis*, might cause Luo-Han-Guo's poor therapeutic effect. On the other hand, both *S. grosvenorii* and *S. siamensis* are dioecious and have a low natural pollination rate, leading to few fruits, although seed traits are important indicators for identifying these two species [15]. Most of the *Siraitia* plants origin privately from wild resources without professional identification and named with ordinary variety names [16]. These phenomena suggest that the cultivation of *Siraitia* species has been immethodical and nonstandard on some level. Moreover, the lack of an effective approach for distinguishing among *Siraitia* species has hindered genetic diversity studies and at least partly led to the gradual loss of some varieties. Thus, high-resolution molecular markers are urgently needed to solve these problems.

Universal molecular markers, such as *ITS*, *rbcL* and *psbA*, are widely used for identifying some species rapidly and accurately [17–19], but they cannot distinguish wild relatives. The chloroplast is a vital and semiautonomous plant cell organelle and has essential roles in photosynthesis and carbon fixation [20,21]. Although most plant chloroplast genomes display highly conserved structures, some structural rearrangements, including inverted repeat (IR) loss, gene loss and indels, are the result of adaptation to their environments [22]; thus, several highly variable regions could be developed as markers for species identification [23], such as indel and single nucleotide polymorphism (SNP) markers for *Panax ginseng* subspecies [24] and indel markers for *Ipomoea nil* and *Ipomoea purpurea* [25]. In addition, abundant closely related species could be identified by combining several markers, such as two indel markers for the identification of three *Aconitum* species [26].

In this study, the complete chloroplast genomes of *S. grosvenorii* and *S. siamensis* were assembled and analyzed, providing the first two sequences in *Siraitia* species. Comparative analyses revealed that the IR regions and coding sequence regions are highly conserved, and several higher-variation regions were primarily located in intergenic regions. Phylogenetic relationship analysis supported the position of two species in the basal lineage of Cucurbitaceae. The identification of nine protein-coding genes in several sites undergoing positive selection contributes to further investigation on the adaptive evolution of plants in ecosystems. Finally, four novel molecular markers (GSPC-F/R, GSPR-F/R, GSPB-F/R and GSPY-F/R) were developed to distinguish the two species. Overall, the sequencing and analysis of two species of chloroplast genomes in *Siraitia* will be beneficial for enhancing medicinal safety and for the species identification and conservation of wild *Siraitia* species, and it will provide new insight for the understanding of plant adaptive evolution in ecosystems.

Materials and methods

Plant materials, DNA extraction, and sequencing

The fresh leaves of two-year-old *S. grosvenorii* and *S. siamensis* plants were collected from Guangxi Medicinal Botanical Garden of the Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College (Nanning, China), then frozen at -80˚C until further use. Total DNA was extracted from approximately 100 mg samples using a plant genomic DNA kit (DP305) (Tiangen Biotech Co., Ltd., Beijing, China), DNA quality was assessed in a Nanodrop 2000 spectrophotometer (Thermo Scientific), and DNA integrity was evaluated using a 1.0% (w/v) agarose gel. DNA samples from each species were used to prepare two separate libraries with an average insert size of 500 bp and sequenced using an Illumina HiSeq 4000 (Illumina Inc., San Diego, CA, USA) with a standard protocol.

Chloroplast genome assembly

First, the low-quality reads sequenced from all the samples were filtered by Trimmomatic software [27]. Then, the trimmed reads, including nuclear and organelle genome data, were used to assemble the chloroplast genome. All chloroplast genomes of plants assessed in the National Center for Biotechnology Information (NCBI) were used to search against Illumina paired-end reads using SRA-BLASTN with an E-value cutoff of 1e-5 [28]. Clean reads with high homology were considered plastome reads and used for downstream genome assembly. SPAdes (v3.10.1) and CLC Genomics Workbench (v7) were used for the de novo genome assembly, the SPAdes using for the assembling that the parameters were set as "-k 21,33,55,77,99,127 –careful" [29]. The contigs obtained were identified by Gepard and spanned the entire plastome [30]. All the identified contigs were assembled using the SeqMan module of DNASTAR (v11.0) [31]. Then, three scaffolds, including the large single-copy (LSC), the IR, and small single-copy (SSC) regions, were obtained. The specific de-novo genome assembler NOVOPlasty was also used to reassemble the two species chloroplast genome for verification [32].To verify the assembly accuracy, the four boundaries between the single-copy (SC) regions and IR regions of the assembled sequence were confirmed by PCR amplification and Sanger sequencing, and the sequences of the primers are listed in S1 Table.

Genome annotation, repeats and simple sequence repeats (SSRs) analyses

The online program Dual Organellar GenoMe Annotator (DOGMA, http://dogma.ccbb.utexas.edu/) and the Chloroplast Genome Annotation, Visualization, Analysis, and GenBank Submission (CPGAVAS) were used to annotate the two genomes [33,34]. The protein-coding sequences were verified by Blastp against the GenBank database. The tRNA genes were identified by tRNAscan-SE and DOGMA [33,35]. Then, manual corrections of the positions of the start and stop codons and the intron/exon boundaries were performed based on the entries in the plastome database using the Apollo program (v.1.11.8) [36,37]. The circular genomic maps were drawn using OrganellarGenomeDRAW (v1.2) with the default setting and checked manually [38]. The newly generated complete chloroplast genome sequences were submitted to GenBank.

The software CodonW (1.4.4) was used to investigate the distribution of codon usage with the relative synonymous codon usage (RSCU) ratio [39]. The codon usage frequency and GC content of both species were calculated using the programs Cusp and Compseq in EMBOSS (v.6.3.1) [40,41]. Repeats, including forward, palindromic, reverse, and complement, were identified by REPuter with the following setting parameters: 3 for Hamming distance and 30 for minimal repeat size [42]. SSRs were detected by MISA software with the parameters set as reported previously, and the cutoffs of the unit numbers for mono-, di-, tri-, tetra-, penta-, and hexa- nucleotides were 8, 4, 4, 3, 3, and 3, respectively [43].

Comparative genomic and selective pressure analyses

The mVISTA program in Shuffle-LAGAN mode with default parameters was used to compare the five complete chloroplast genomes using *S. grosvenorii* chloroplast genomes as a reference [44,45]. The sequence divergence of the chloroplast genomes was analyzed with a sliding window using DnaSP (v5.10) [46], and the step size was set to 200 bp with a 600 bp window length. Moreover, a total of 104 intergenic regions and 77 exons were manually extracted among four species, and the corresponding sequences aligned using ClustalW2 (v2.0.12) [47] were used to calculate the nucleotide variability (Pi) using DnaSP (v5.10). Selective pressure was analyzed for consensus protein-coding genes among twelve genomes from Cucurbitaceae species. EasyCodeML software with the site model was performed to calculate the nonsynonymous (*Ka*) and synonymous (*Ks*) substitution ratios and likelihood ratio tests (LRTs). The values of both *Ka*/*Ks* (ω) and the LRTs were coupled to evaluate the selection on amino acid sites [48].

Phylogenetic analyses

A total of 30 chloroplast genomes, including 28 from Cucurbitaceae, were used for the phylogenetic analyses, including *Nicotiana tabacum* and *Arabidopsis thaliana* as outgroups. In this study, 28 chloroplast genome sequences were downloaded from NCBI GenBank (S2 Table). The software MAFFT was used to generate alignments of 64 consensus protein-coding gene sequences [49], and then the alignments were manually adjusted by BioEdit [50]. Maximum likelihood (ML) analysis was carried out based on the Tamura-Nei model using a heuristic search for initial trees that were the most appropriate by Modeltest 3.7 [51]. Maximum parsimony (MP) analysis was conducted using PAUP (v4.0a) [52]. Bootstrap analysis was performed with 1000 replicates. Finally, the reconstructed trees were visualized using Figtree (v1.4.3) (http://tree.bio.ed.ac.uk/software/figtree/).

Molecular marker development and validation

The molecular regions between the two *Siraitia* species were examined by the alignment and comparison of mVISTA similarities. The primers for the molecular markers were designed using the software Primer Premier 5.0 [53]. The accuracy of the molecular markers was verified using PCR amplification, and PCR was conducted with the following program: initial denaturation at 94˚C for 2 minutes; followed by 35 cycles of amplification at 94˚C for 20 seconds, 56˚C for 20 seconds, and 72˚C for 2 minutes; and final extension at 72˚C for 10 minutes. The PCR products were separated with 1.0% (w/v) agarose gel for 20 minutes at 180 volts. Then, the DNA fragments were purified and sequenced.

Results and discussion

General features of the chloroplast genomes

The chloroplast genome structures of the two species are similar to those of other Cucurbitaceae species, which display a single circular molecule with a typical quadripartite structure. The complete chloroplast genome of *S. grosvenorii* was 158,757 bp in length and consists of a pair of IRs (IRa and IRb, each 26,288 bp in length) separated by a LSC (87,625 bp) region and a SSC (18,556 bp) region (Fig 1, S1 Fig and S3 Table). The complete chloroplast genome of *S. siamensis* possessed the same structure. The IRa and IRb were 26,289 bp each, the LSC was 88,069 bp, and the SSC was 18,543 bp. In total, the length of the whole genome was 159,190 bp. (Fig 1, S3 Table). The length of each region is similar to those of most plant chloroplast genomes reported previously [54]. The sequencing data of *S. grosvenorii* and *S. siamensis* were deposited in GenBank under the accession numbers MK755853 and MK755854, respectively.

Further analysis results revealed that the two species had approximately 36.8% GC content (S3 Table), which was distributed unevenly across the whole chloroplast genome. In contrast to the LSC regions, the GC contents in the IR regions displayed a higher value across the whole chloroplast genome, 42.8% in both *S. grosvenorii* and *S. siamensis*, possibly resulting from rRNA genes (*rrn16*, *rrn23*, *rrn4.5* and *rrn5*, 55.3% GC content in both cases) with high GC content located in the IR regions [55]; the higher GC content in the IR region was regarded as an indicator of species affinity [56]. Moreover, the LSC regions had a GC content of 34.6% in both species, and the lowest values of 30.5% and 30.7% were found in the SSC regions in *S. grosvenorii* and *S. siamensis*, respectively. In addition, the GC contents of the protein-coding regions were 37.8% in *S. grosvenorii* and 37.9% in *S. siamensis*, and the percentages of GC content for the first, second, and third codon positions were 45.6%, 38.0% and 29.8% in

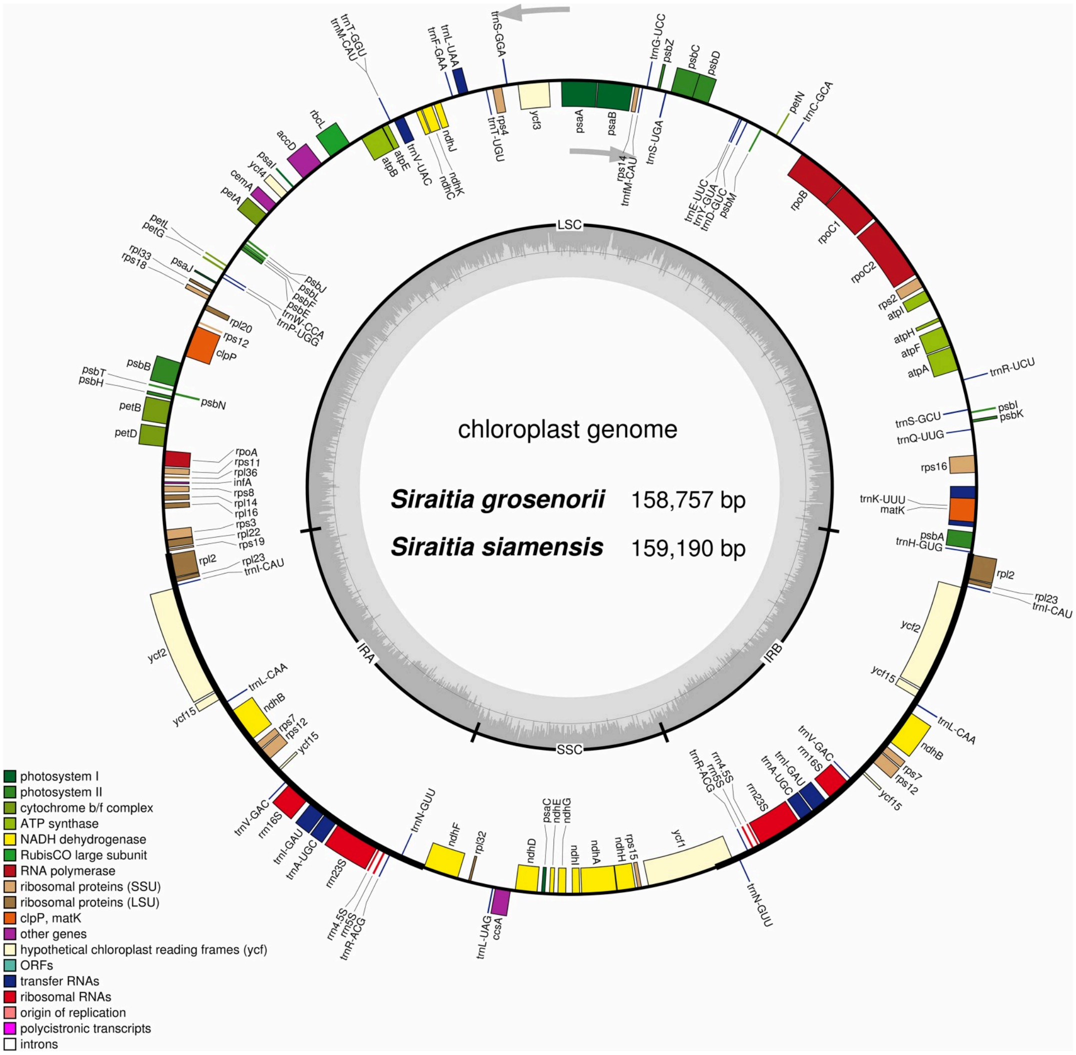

Fig 1. Circular Gene map of the complete chloroplast genomes of *S. grosvenorii* and *S. siamensis*. The quadripartite structure includes two copies of an IR region (IRa and IRb) that separated by (LSC) and SSC regions. Genes drawn in the circle are the transcribed clockwise, and those on the outside are transcribed counter-clockwise. The darker gray area in the inner circle show the GC content, whereas the lighter corresponds to AT content. Different genes groups are colored.

https://doi.org/10.1371/journal.pone.0226865.g001

S. grosvenorii and 45.6%, 62.0% and 29.9% in *S. siamensis*, respectively (S3 Table). A bias toward using thymine (T) and adenine (A) in the third codon position has also been observed in other land plant plastomes [57–59]. The skewed GC distribution across the whole genome might be associated with the position of the origin and terminals for gene replication [60–62].

The two species exhibited the same gene content and arrangement in the chloroplast genome. A total of 134 genes were identified, including 89 protein-coding genes, 37 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes. Eight protein-coding genes, seven tRNA genes, and four rRNA genes were duplicated in the IR regions in both species (Fig 1); *ycf15-orf* was duplicated, and its start codon was GTG. The data revealed that 23 genes contain introns in both genomes, 21 with one intron and the rest with two introns, and the genes *clpP* and *ycf3* both contained three exons (S4 and S5 Tables). The gene *clpP* was related to energy transformation [63], and *ycf3* was necessary for the stable accumulation of photosystem I complexes [64]. Introns found in functional genes play a significant role in the regulation of gene expression, which can trigger desirable biological traits at particular times [65,66]. In addition, the *rps12* gene contained three exons and one intron because of trans-splicing, which resulted in a 5' end exon located in the LSC region, whereas the remaining exons were located in the IRs (S5 Table). Therefore, *rps12* was duplicated in the IR region. Furthermore, the results of gene location analysis revealed twelve genes with partial overlaps in their sequences: *trnK-UUU*/*matK*, *psbD*/*psbC*, *trnM-CAU*/*trnT-GGU*, *atpE*/*atpB*, *rps3*/*rpl22*, and *trnP-UGG*/*trnP-GGG*.

The basic characteristics of the chloroplast genomes of two *Siraitia* species and six other species from Cucurbitaceae are shown in S6 Table. Comparative analysis showed that the lengths of the eight genomes ranged from 155,293 bp (*Cucumis sativus*) to 159,190 bp (*S. siamensis*), and the overall GC content percentage of *C. sativus* (37.2%) was higher than that of any other genome (36.7%-37.1%). However, little difference could be found in gene number, gene type or GC content between *S. grosvenorii* and *S. siamensis*, which suggests that we should focus on other areas to find variation, such as intergenic spacers.

IR/SC boundaries and IR contraction and expansion

The contraction and expansion of the IR regions account for common evolutionary events and are a major cause of differences in chloroplast genome size, and evaluating them could shed some light on the evolution of some taxa [67,68]. A detailed comparison of the IR/SC boundary regions of twelve species is shown in Fig 2. *A. thaliana* and *N. tabacum* were set as outgroups, and the rest were Cucurbitaceae. From the comparison, we noticed that the lengths of *ycf1* pseudogenes of both *Siraitia* species were both 1,181 bp, almost as long as those of the other five Cucurbitaceae species, except *Lagenaria siceraria* (28 bp) and *Momordica charantia* (29 bp). In addition, the *ndhF* gene, located in the SSC, reaches 134–135 bp across the IRb/SSC boundary in both *L. siceraria* and *M. charantia*, while the corresponding region was 7–12 bp in most other Cucurbitaceae species. Moreover, the gene *rps19* is located in the LSC region with 265–277 bp across the LSC/IRb boundary in *Trichosanthes kirilowii*, *Hemsleya lijiangensis* and *Gynostemma laxiflorum*, whereas in other species, it was located away from the LSC/IRb by 68–208 bp. The *rpl2* gene, duplicated in the IRs in most species, was present as only one copy in the IRb region in *M. charantia*, *Cucurbita pepo* and *Citrullus lanatus*, which might result from the location of the LSC/IRb boundary. All of these phenomena were related to the contraction/expansion of two IR regions in the complete chloroplast genomes.

Codon usage

RSCU is a measure of nonuniform synonymous codon usage in coding sequences in which values above 1 indicate that codons are used more frequently than expected [69,70]. All the

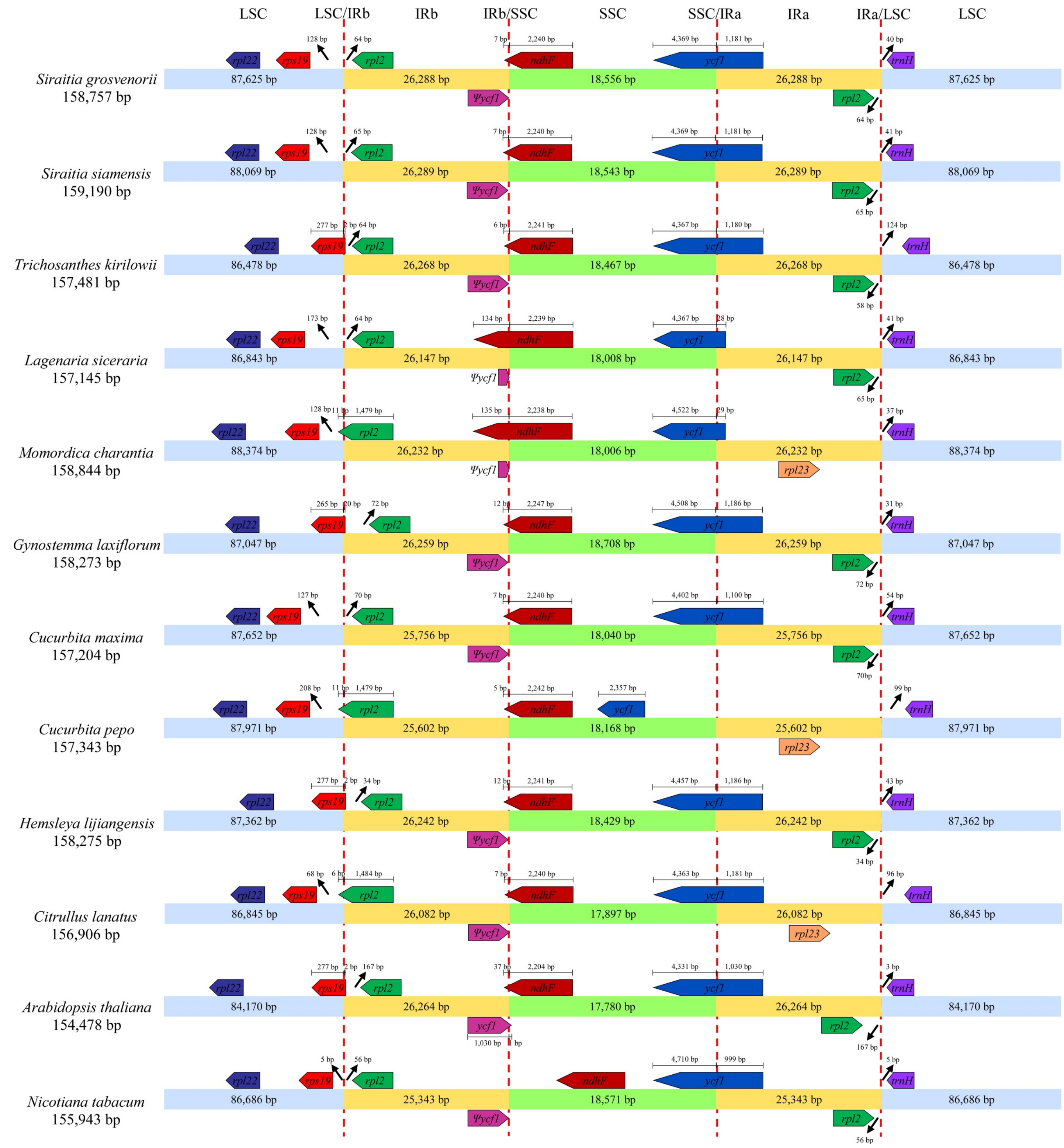

Fig 2. Comparison of border distance between adjacent genes and junctions of the LSC, SSC and IRs regions among twelve chloroplast genomes. Numbers with arrow above the gene features mean the distance between the ends of genes and the borders sites. The figure is not to scale with respect to sequence length.

https://doi.org/10.1371/journal.pone.0226865.g002

protein-coding genes were encoded with 26,509 and 26,508 codons in the *S*. *grosvenorii* and *S*. *siamensis* chloroplast genomes, respectively. Detailed codon analysis revealed that the two cases had similar codon constitutions and RSCU values (S7 Table). Among these codons, leucine (Leu) and cysteine (Cys) were, respectively, the highest (10.48%) and lowest (1.13%) prevalence amino acid codons in the two *Siraitia* species. In addition, most of the amino acid codons showed preferences, with exception of methionine (Met) and tryptophan (Trp), which both had RSCU values of 1. The chloroplast genomes of the two *Siraitia* species both had 30 biased codons with RSCU >1, and the third positions of the biased codons were A/U except for Leu (UUG); otherwise, the codons with high frequency (>30%) and fraction were Asp-GAU, Glu-GAA, Ile-AUU, Lys-AAA, Leu-UUA, Asn-AAU, and Tyr-UAU, and the bias toward these seven codons was consistent with the low content of GC in the third codon position. Fig 3 shows that the RSCU value increased with the quantity of codons coding for a specific amino acid. A strong AT bias in codon usage is common in sequences with strong codon preference and was also found in most other land plant chloroplast genomes [71,72].

Repeat structure and SSRs analyses

Repeat units, which are distributed in chloroplast genomes with high frequency, play an important role in genome evolution [73–75]. S2(A) Fig shows repeat structures that were longer than 30 bases in eight species. The repeats of the *S*. *grosvenorii* chloroplast genome consist of 19 forward, 24 palindromic, two reverse, and one complement. By contrast, a slightly different number of repeats was found in *S*. *siamensis*, which contained 17 forward, 20 palindromic, one reverse and no complement. In the eight-species comparison, the number of repeats for *C*. *lanatus* (31) was the lowest, and the highest number of repeats was 49 in *M*. *charantia*, *G*. *laxiflorum*, and *Cucurbita maxima*.

On the other hand, SSRs, which are also known as microsatellites and are distributed abundantly across genomes, are tracts of repetitive DNA with certain motifs, ranging from 1–6 or more base pairs, that are repeated typically 5–50 times [76,77]. SSRs are widely used as molecular markers for species identification, analysis of phylogenetic relationships and population genetics because of their high polymorphism rates and stable reproducibility [78,79]. Here, a total of 252 and 253 SSRs were identified by MISA software within the chloroplast genomes of *S*. *grosvenorii* and *S*. *siamensis*, respectively (S2(B) Fig), and mononucleotide repeats were largest in number, 57 and 56, respectively. Moreover, S8 Table shows that A/T mononucleotide repeats (97.2% and 97.8%, respectively) were the most common, and for dinucleotide repeats, AT/ TA (68.4% and 67.8%, respectively) was the majority. However, only one repeat unit (AAT/TTA) was found among trinucleotide repeats. In addition, AT-rich repetitive motifs were high in the remaining SSR types. The SSRs within the chloroplast genomes of both species mainly comprised AT-rich repetitive motifs, consistent with the fact that AT content (63.2%) was very high (GC content was 36.8%) in both cases. Furthermore, these results were also consistent with previous reports that proportions of short poly-A or poly-T repeats were higher than those of poly-G or poly-C within most SSRs in many plant chloroplast genomes [80,81]. Distribution of the SSRs loci in the chloroplast genome of *S*. *grosvenorii* and *S*. *siamensis* were exhibited in S9 Table.

Sequence divergence and nucleotide diversity

Complete chloroplast genomes are often used to analyze plant taxonomy, phylogenetic relationships, and genetic diversity [82]. In this study, two *Siraitia* species were compared with other three species in Cucurbitaceae using mVISTA software, and *S*. *grosvenorii* was set as the

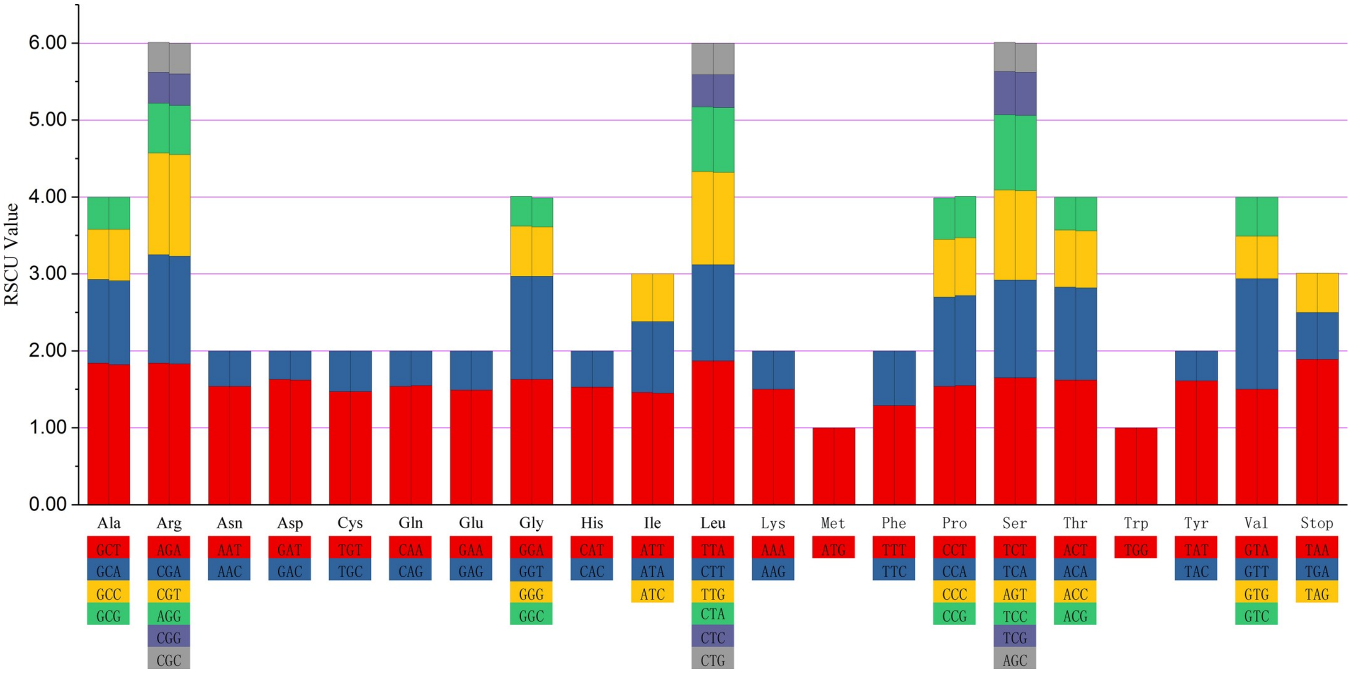

Fig 3. Codon usage for 20 amino acid and stop codons in all protein-coding genes. The columnar stacking diagram on the left and right of each amino acid display the codon usage within the chloroplast genome of *S. grosvenorii* and *S. siamensis*, respectively.

https://doi.org/10.1371/journal.pone.0226865.g003

reference. As shown in Fig 4, sequence divergence was similar for the whole sequences of the complete chloroplast genomes. In contrast to the other two Cucurbitaceae species, *M. charantia* was more similar to the two *Siraitia* species across the complete chloroplast genomes. The data plot revealed that the noncoding region was more divergent than its coding counterparts. The two IR regions were both less divergent than the single-copy regions, which might be the result of the four highly conserved rRNAs located in the IR regions [55].

In addition, the nucleotide diversity of 181 regions was analyzed using DnaSP software, including 77 protein-coding genes and 104 intergenic regions among four chloroplast genomes (*T. kirilowii*, *M. charantia*, and two *Siraitia* species). The results revealed that intergenic regions were more divergent than protein-coding genes (Fig 5). The average nucleotide variability (Pi) in the intergenic regions was 0.01772, almost twice as much as that in protein-coding genes (Pi = 0.00950). This is consistent with previous research in angiosperm chloroplast genomes [83]. The *trnL-ccsA* (Pi = 0.07302) and *petG-trnW* (Pi = 0.06780) regions were notably variable among the intergenic regions, as were the genes *ycf1* (Pi = 0.03048), *psaJ* (Pi = 0.02972) and *atpE* (Pi = 0.02447) among the protein-coding genes. *ycf1* is commonly used as a representative plant DNA barcoding region [84]. Several highest-level divergences ($Pi > 0.03$) were found in the intergenic regions and could be developed as specific molecular markers for species identification, including *trnR-atpA*, *ndhC-trnV*, *petG-trnW*, *rpl32-trnL*, *trnL-ccsA*, *ndhF-rpl32*, *psbZ-trnG*, *psbC-trnS*, *ndhG-ndhI*, *rps8-rpl14*, *ccsA-ndhD*, *trnT-trnL*, *psbK-psbI*, *psbA-trnK*, *rps15-ycf1*, *trnF-ndhJ*, *atpA-atpF*, *rps19-rpl2*, *psaC-ndhE*, and *petD-rpoA*. It has been reported that divergent noncoding regions allow the discrimination of potential molecular markers and DNA barcodes [85]. Furthermore, sliding window analysis was also performed (S3 Fig). The average value of Pi was 0.00221 between the two *Siraitia* species, and higher variability was found in the LSC and SSC regions.

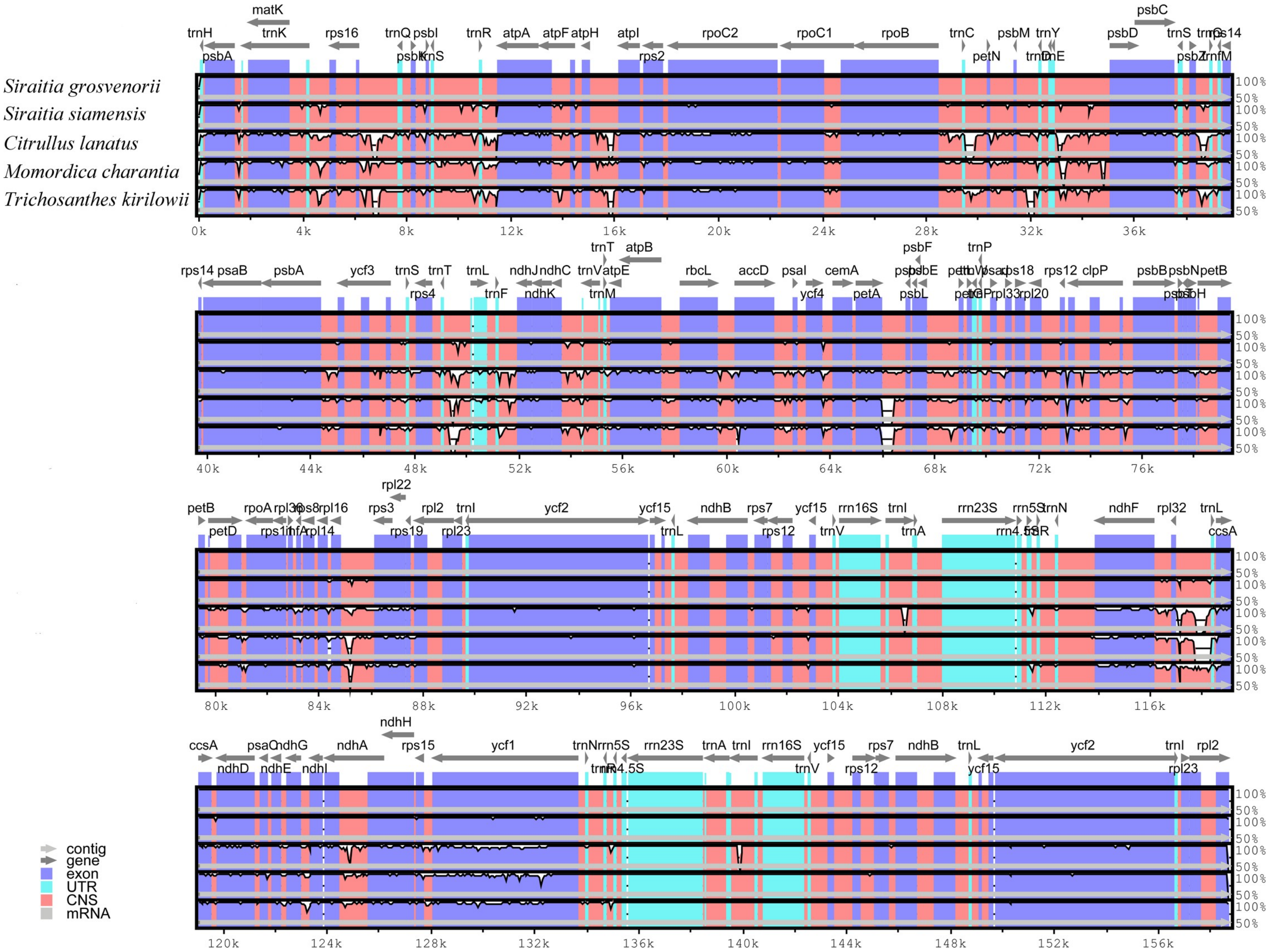

Fig 4. Structure comparison of five chloroplast genomes using mVISTA program. Gray arrows and thick black lines above the alignment indicate genes with their orientation and the position of the IRs, respectively. Purple bars, blue bars, pink bars, gray bars and white peaks represent exon, Untranslated Region (UTRs), Conserved Noncoding Sequences (CNS), mRNA and genomes differences, respectively. A cut-off of 70% identity was used for the plots, and the Y-scale represents the percent identity ranging from 50% to 100%.

https://doi.org/10.1371/journal.pone.0226865.g004

Phylogenetic relationships of six genera within Cucurbitaceae

Chloroplast genomes are significant genomic resources for the reconstruction of accurate and high-resolution phylogenetic relationships and taxonomic status in angiosperms [86]. Complete chloroplast genomes and protein-coding genes have been widely employed to determine phylogenetic relationships at almost every taxonomic level [87]. In this study, to identify the phylogenetic positions of the two *Siraitia* species within the Cucurbitaceae family, we aligned 64 protein-coding sequences from 30 chloroplast genomes; *A*. *thaliana* and *N*. *tabacum* were set as outgroups, and the alignment length was 62,522 bp. The ML and MP trees displayed similar phylogenetic topologies (Fig 6). All nodes in the ML tree and MP tree were strongly supported by high bootstrap values: 22 of 27 nodes with 100% bootstrap values were found in the MP tree and 21 of 27 in the ML tree. In addition, the results illustrated that two *Siraitia* species were the most closely related species to *M*. *kirilowii*, and these two taxa were grouped with two

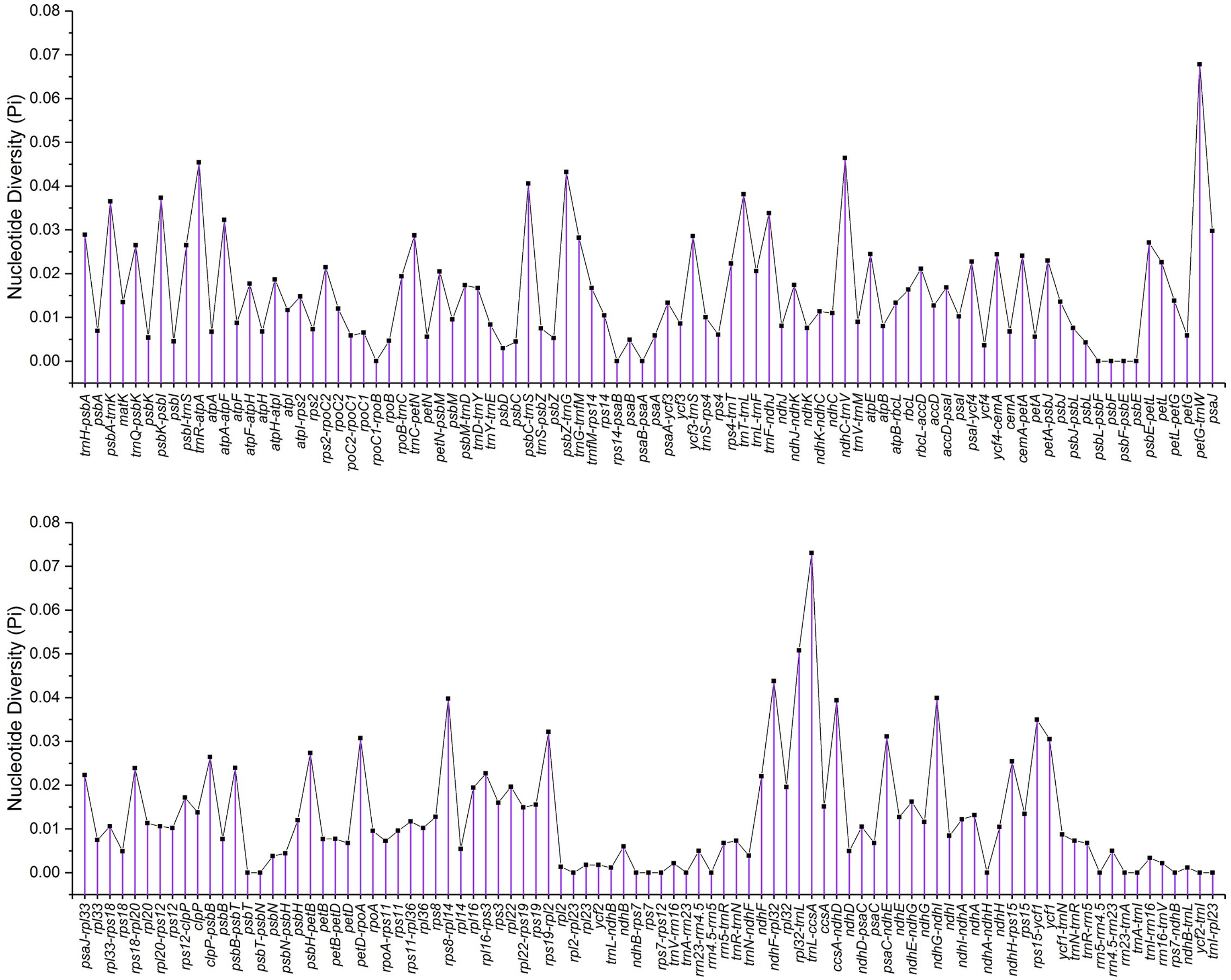

Fig 5. Comparison of nucleotide diversity (Pi) value for 77 protein-coding genes and 104 intergenic regions among six closely species.

https://doi.org/10.1371/journal.pone.0226865.g005

species from Sicyoeae, three species from Cucurbiteae, and nine from Benincaseae, which showed a nested evolutionary relationship in the MP and ML trees. Furthermore, all species were clustered into a lineage distinct from the outgroups and strongly supported the new classification system of Cucurbitaceae [54].

Selective pressure events

Synonymous substitutions (*Ks*) accumulate nearly neutrally, whereas nonsynonymous substitutions (*Ka*) are subjected to selective pressures of varying degree and direction (positive or negative); values of *Ka*/*Ks* (ω) above 1.0 indicate that the corresponding genes experience positive selection, while ω values ranging from 0.5 to 1.0 indicate relaxed selection [88]. In the current study, we performed a selective analysis of the exons of each protein-coding gene using site-specific models with four comparison models (M0 vs. M3, M1a vs. M2a, M7 vs. M8 and

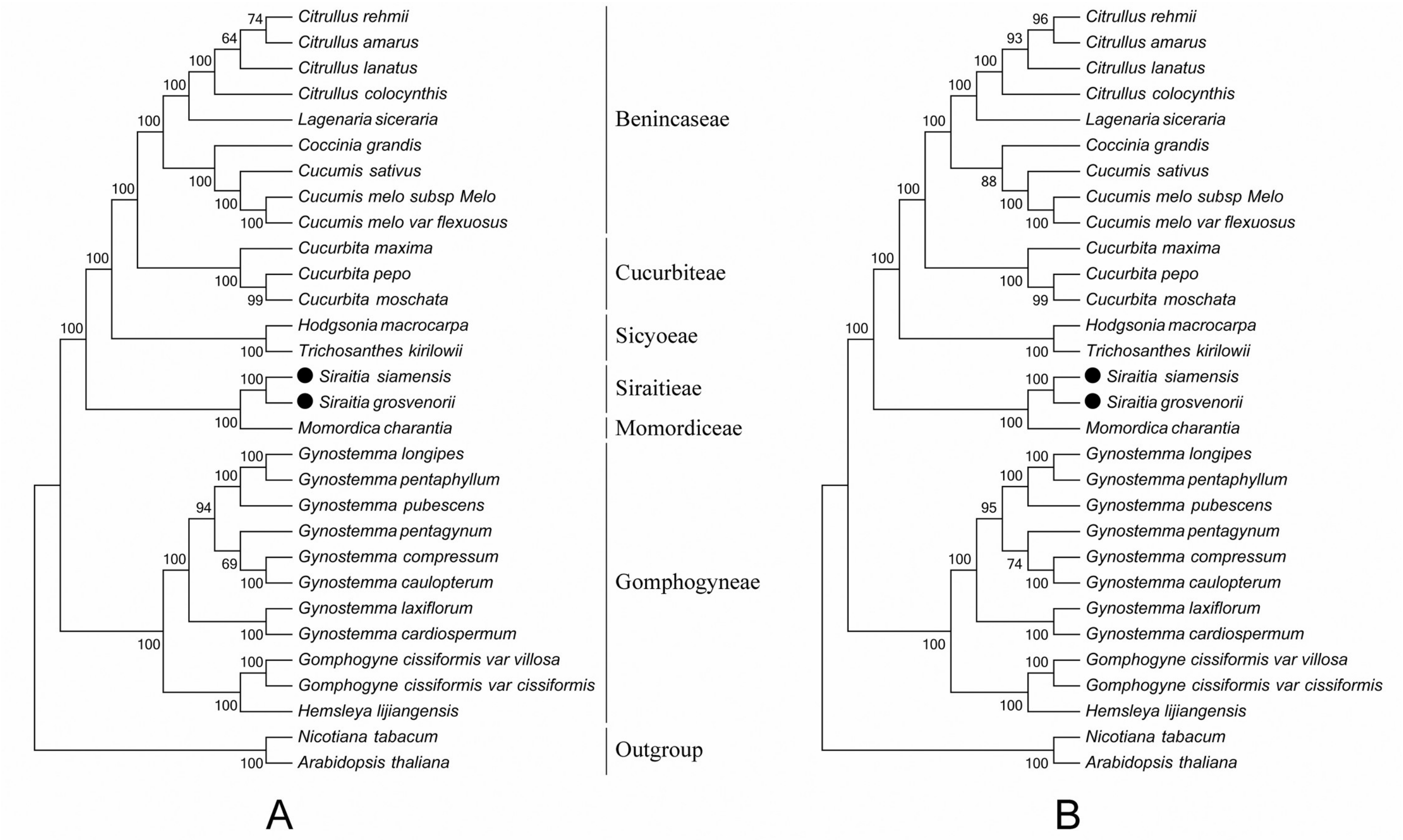

Fig 6. Molecular phylogenetic relationship of Siraitieae with 64 protein-coding genes of 26 cucurbitaceae species. The unrooted trees were constructed by (A) maximum parsimony (MP) and (B) maximum likelihood (ML) methods with bootstrap values ≥50%.

https://doi.org/10.1371/journal.pone.0226865.g006

M7a vs. M8a, LRT threshold $p \leq 0.05$) in EasyCodeML software as reported previously [48]. Among the eight models, M2a was the positive model, and p (p_0, p_1, p_2) were the proportions of purifying selection, neutral selection, and positive selection. A total of 58 consensus protein-coding genes from 12 closely related species were evaluated with respect to selective pressure. Nine genes (*accD*, *atpA*, *atpE*, *atpF*, *clpP*, *ndhF*, *psbH*, *rbcL*, *and rpoC2*) were found to have undergone positive selection, and the ω_2 values ranged from 2.17 to 11.44 in the M2a model (Table 1).

To determine which sites were subject to positive selection, naive empirical Bayes (NEB) and Bayes empirical Bayes (BEB) methods were used to analyze the location of consistent selective sites in the alignment of chloroplast genomes in the M7 vs. M8 model. The data analysis revealed that the gene *rpoC2* possesses 7 positive selective sites, followed by *atpF* (6), *rbcL* (4), *atpA* (2), *atpE* (2), *clpP* (2), *ndhF* (2), *accD* (1), and *psbH* (1). All positively selected sites in these nine genes are shown in Table 1.

Among the nine positively selected genes, *rpoC2* encodes the RNA polymerase β”. A comparison of *rpoC2* between fertile lines of sorghum and cytoplasmic male sterile lines showed that a 165 bp deletion was identified that encodes several protein motifs involved with transcription factors; this region might play an important role in the regulation of developmental pollination [89]. The *Siraitia* species are dioecious, so the finding that *rpoC2* evolved under positive selection might indicate that it is essential for sex differentiation. The chloroplast plays

Table 1. The results of positive selective pressure analysis in M7 vs.M8 model.

Gene name	Modle	np	LnL	ω2 (M2a)	LRTs (2ΔLnL)	LRT P-value	Positive sites
rpoC2	M8	26	-8422.120232	11.44045	43.1617	0	486 V*, 527 P**, 754 S*, 965 S**, 1024 F*, 1331 P**, 1353 F*
	M7	24	-8443.701061				
atpF	M8	26	-1276.119906	2.17251	6.8495	0.032557126	2 E*, 8 K**, 9 K*, 14 F*, 61 E*, 97 L*
	M7	24	-1279.544665				
rbcL	M8	26	-2716.450055	3.38881	38.4242	0.000000005	251 M*, 255 I**, 470 E*, 472 M*
	M7	24	-2735.662163				
atpA	M8	26	-2937.475433	5.46438	14.4739	0.00071952	258 S*, 484 N*
	M7	24	-2944.712359				
atpE	M8	26	-839.182861	7.38230	7.7637	0.020612244	36 D*, 130 G*
	M7	24	-843.064731				
ndhF	M8	26	-5697.671745	9.36793	15.6948	0.000390771	509 I*, 686 I*
	M7	24	-5705.519135				
clpP	M8	26	-1298.084585	3.91456	9.4032	0.009080555	11 V*, 51 Y*
	M7	24	-1302.786205				
accD	M8	26	-3041.598734	2.91222	9.2524	0.009791966	223 I*
	M7	24	-3046.224927				
psbH	M8	26	-387.753418	11.27276	10.1816	0.006153015	72 S**
	M7	24	-392.844231				

* $P < 0.05$

** $P < 0.01$

np represents the degree of freedom.

https://doi.org/10.1371/journal.pone.0226865.t001

important roles in photosynthesis and carbon fixation, and six genes (*atpF*, *rbcL*, *atpA*, *atpE*, *ndhF*, and *psbH*) with essential roles in photosynthesis were positively selected in this study. The *Siraitia* species are distributed in southeast Asia, so requirements for sufficient light might have exerted strong selective forces on the six genes during plant evolution. This phenomenon was also found in species within the *Urophysa* genus, which is distributed in southwest China [75]. The *clpP* gene, encoding the ATP-dependent clp protease, is likely involved in the transformation of chloroplast protein and might be essential for shoot development under clpP-mediated protein degradation [63,90,91]. The positive selection of the gene *clpP* in our study might be associated with the evolution of the vining character within Cucurbitaceae. As for the gene *accD*, it encodes the β–carboxyl transferase subunit of acetyl-CoA carboxylase [92]; it is a vital gene for leaf development and has effects on leaf longevity and seed yield [93]. Expression of the gene *accD* might indirectly affect the efficiency of photosynthesis. These nine genes have undergone positive selection, which might be the result of adaptation to their barren environment.

Molecular markers for distinguishing *S. grosvenorii* and *S. siamensis*

In this study, several notably variable regions were found in the comparison of the two chloroplast genomes. To develop high-resolution molecular markers for the identification of these two species, the specific divergent regions, including *ndhC-trnV-UAC*, *trnR-UCU-atpA*, *rpoB-trnC-GCA* and the gene *ycf1*, were chosen as molecular marker regions, and specific primers were designed against the conserved regions (Table 2). The primers, named GSPC-F/R, GSPR-F/R and GSPB-F/R, were used for amplifying the three intergenic regions and produced

Table 2. Primer identification for molecular markers.

Primer name	Primer sequence (5' to 3')	position
GSPC-F	GATGAACCAAATCAAGTGGC	ndhC-trnV-UAC
GSPC-R	CAGAAGCAGGACGATAGAGA	
GSPR-F	GGTTCAAATCCTATTGGACG	trnR-UCU-atpA
GSPR-R	GGCAAGAGGTCAACGATTAC	
GSPB-F	CTGTTTCCTACTCACACGAG	rpoB-trnC-GCA
GSPB-R	GGATTGGCTCTATCTCTTCG	
GSPY-F	GACGACTTGCTTTAGCGTTG	ycf1
GSPY-R	GGACTAAACAGGAACAAGAG	

https://doi.org/10.1371/journal.pone.0226865.t002

different-length fragments in *S*. *grosvenorii* and *S*. *siamensis*, respectively. The gene *ycf1*, with high divergence, was also chosen for the development of molecular markers, and several SNP sites were found after amplification with GSPY-F/R. (Fig 7). The four molecular markers developed in this study will contribute to the identification of *Siraitia* species and facilitate efficient utilization and conservation of wild *Siraitia* resources.

The lack of an effective approach made the position of different species in the genus unreliable until the *Siraitia* Merrill was acknowledged in 1980 [15]. The phylogenetic analysis in this study showed that the two *Siraitia* species were the most closely related species to *M*. *kirilowii*, which explains the reason that these species were placed into *Momordica* L. with the name of *Momordica grosvenorii* Swingle initially [94]. To the best of our knowledge, only seven species within the genus *Siraitia* have been confirmed based on morphological characteristics [15], although some varieties may remain undiscovered. Unfortunately, the increasing demand for high production of these species has brought the wild resources to the verge of depletion. The comparative analysis of chloroplast genomes in this study revealed that several highly variable intergenic regions will contribute to the development of specific markers to support the conservation of wild *Siraitia* varieties. Among the seven species, *S*. *grosvenorii*, which is recorded as both a medicinal and an edible species [4], has been widely cultivated as an important commercial crop. *S*. *siamensis* is known to have better disease resistance and setting percentage and considerable siamenoside I content. Different germplasm resources should be distinguished accurately and used for different purposes. Not only do adulterants of raw medicinal materials threaten the safety and reliability of food and medicine, but they also exacerbate the scarcity of wild resources in similar species. Our development of novel molecular markers will be of great use in the species authentication and conservation of *Siraitia* wild resources in the future.

Conclusions

In the current study, two complete chloroplast genomes of the *Siraitia* genus were assembled and analyzed for the first time. In a comparison with other species within Cucurbitaceae, several highly divergent noncoding regions were identified that would be beneficial for developing high-resolution molecular markers. Phylogenetic relationship analysis supported that *S*. *grosvenorii* and *S*. *siamensis* originated from the same ancestor, consistent with previous studies. Furthermore, 9 protein-coding genes were found to undergo selection, which might be the result of adaptation to the environment. Finally, molecular markers (GSPC-F/R, GSPR-F/R, GSPB-F/R and GSPY-F/R) were developed to distinguish the two species. The results in this study will be beneficial for taxonomic research, species identification, and conservation of the genetic diversity of *Siraitia* wild resources in the future.

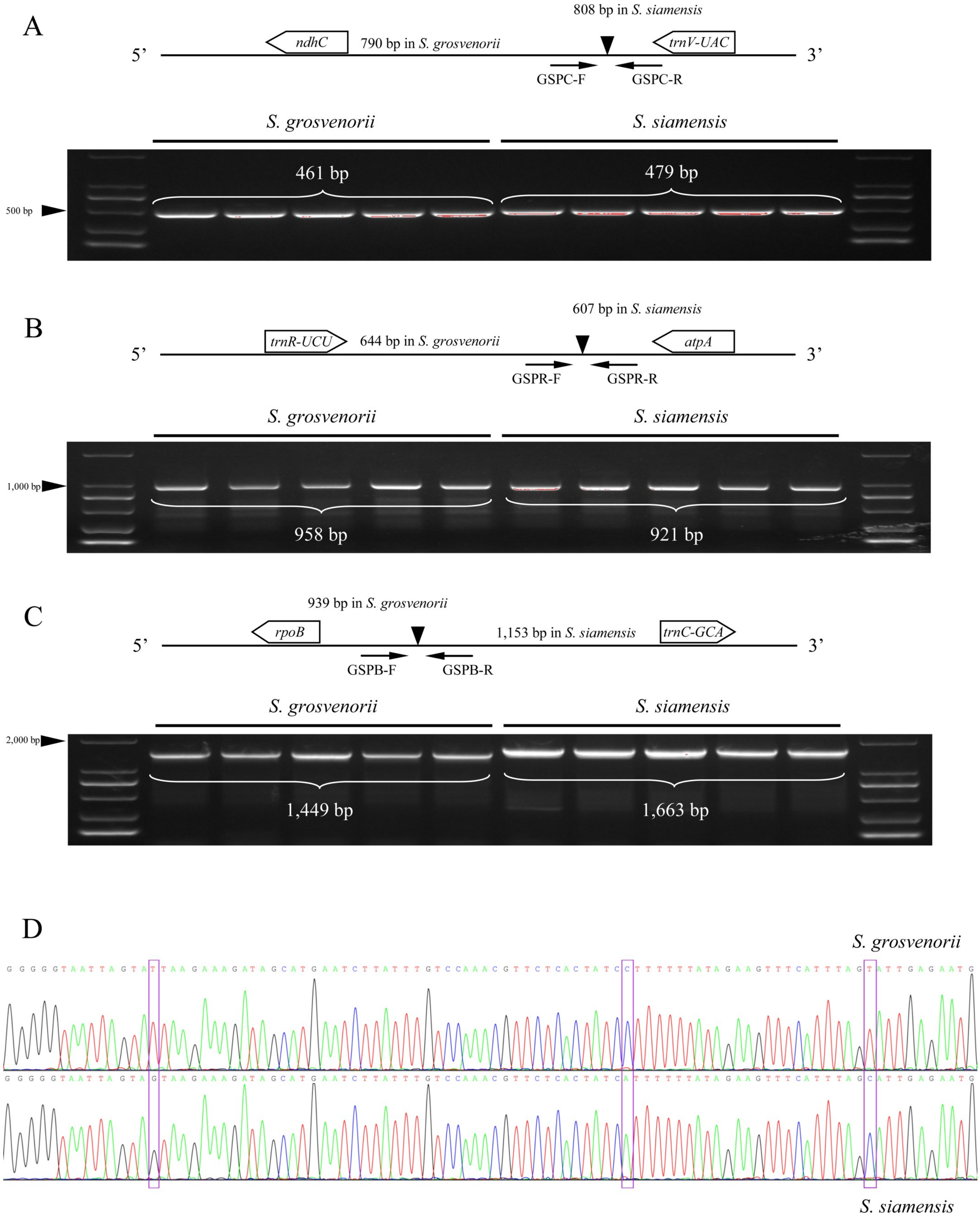

Fig 7. Schematic diagram displayed the four novel molecular markers. The indel makers in the intergenic spacers, including GSPC-F/R, GSPR-F/R, and SCPB-F/R, and ycf1 SNP were verified in *Siraitia* species with five individuals. (A) Sequencing results showed that PCR amplification by GSPC-F/R between the

ndhC and *trnV-UAC* in *S. grosvenorii* and *S. siamensis* were 790 bp and 808 bp, respectively. The difference can also be found by GSPR-F/R (B), and SCPB-F/R (C) between the two *Siraitia* species. (D) GSPY-F/R marker for *ycf1* SNP sites were validated in each case, which were effective to discriminate the two *Siraitia* species.

https://doi.org/10.1371/journal.pone.0226865.g007

Supporting information

S1 Fig. Chloroplast genomes of *S. grosvenorii* and *S. siamensis* in linear form.
(TIF)

S2 Fig. Distribution of the number of different type repeats and SSRs. (A) Repeat sequence in eight chloroplast genomes. Repeat sequences were identified by REPuter with length ≥30bp and sequence identified ≥90%. F, P, R, and C are the abbreviation of repeat type F (forward), P (palindrome), R (reverse) and C (complement), respectively. Different length repeat sequences are colored correspondingly. (B) Analysis of simple sequence repeat (SSRs) in chloroplast genomes of five species.
(TIF)

S3 Fig. Sliding window analysis of the whole chloroplast genome. Window length: 600 sites, Step size: 200 sites. X-axis: position of the midpoint of a window; Y-axis: nucleotide diversity (π) of each window. (A) Pi among *S. grosvenorii* and *S. siamensis*; (B) Pi among six species of Cucurbitaceae.
(TIF)

S1 Table. Primer sequence at the boundaries between single cope and IR regions.
(DOCX)

S2 Table. List of chloroplast genome sequence used in the study.
(DOCX)

S3 Table. Base composition in the chloroplast genomes of *S. grosvenorii* and *S. siamensis*.
(DOCX)

S4 Table. Genes contained in the chloroplast genomes of *S. grosvenorii* and *S. siamensis*.
(DOCX)

S5 Table. Location information of genes with introns in the chloroplast genome of *S. grosvenorii* and *S. siamensis*.
(DOCX)

S6 Table. Comparisons among the chloroplast genome characteristics of *S. grosvenorii* and *S. siamensis*, and other six Cucurbitaceae species.
(DOCX)

S7 Table. Codon usage and codon-anticodon recognition in all protein-coding genes of the chloroplast genomes of two *Siraitia* species.
(DOCX)

S8 Table. Types and amounts of SSRs in the *S. grosvenorii* and *S. siamensis* chloroplast genomes.
(DOCX)

S9 Table. Distribution of the SSRs loci in the chloroplast genome of *S. grosvenorii* and *S. siamensis*.
(XLSX)

Acknowledgments

The authors acknowledge Mr. Jianguo Zhou, Miss. Hui Zhang and Miss. Mei Jiang in Institute of Medicinal Plant Development for their kind suggestions in the raw data process. We also thank the Guangxi Medicinal Botanical Garden of the Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College (Nanning, China) for supplying the plant materials as a generous gift.

Author Contributions

Conceptualization: Hongwu Shi, Chang Liu, Xiaojun Ma.

Data curation: Meng Yang, Bin Wu.

Formal analysis: Meng Yang, Bin Wu.

Funding acquisition: Xiaojun Ma.

Investigation: Changming Mo, Xiaojun Ma.

Methodology: Hongwu Shi, Meng Yang, Changming Mo, Wenjuan Xie, Chang Liu, Bin Wu.

Project administration: Changming Mo, Xiaojun Ma.

Resources: Changming Mo, Wenjuan Xie.

Software: Hongwu Shi, Meng Yang, Chang Liu, Bin Wu.

Supervision: Chang Liu, Bin Wu, Xiaojun Ma.

Validation: Chang Liu.

Visualization: Chang Liu, Xiaojun Ma.

Writing – original draft: Hongwu Shi, Meng Yang.

Writing – review & editing: Hongwu Shi, Bin Wu, Xiaojun Ma.

References

1. Jeffrey C (1980) A review of the Cucurbitaceae. Botanical Journal of the Linnean Society 81: 233–247. https://doi.org/10.1111/j.1095-8339.1980.tb01676.x
2. Tang Q, Ma X, Mo C, Wilson IW, Song C, Zhao H, et al. (2011) An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics 12: 343. https://doi.org/10.1186/1471-2164-12-343 PMID: 21729270
3. Philippe RN, De MM, Anderson J, Ajikumar PK (2014) Biotechnological production of natural zero-calorie sweeteners. Curr Opin Biotechnol 26: 155–161. https://doi.org/10.1016/j.copbio.2014.01.004 PMID: 24503452
4. Chun LI, Lin LM, Feng S, Wang ZM, Huo HR, Li D, et al. (2014) Chemistry and pharmacology of *Siraitia grosvenorii*: A review. Chinese Journal of Natural Medicines 12: 89–102. https://doi.org/10.1016/S1875-5364(14)60015-7 PMID: 24636058
5. Deng F, Liang X, Yang L, Liu Q, Liu H (2013) Analysis of Mogroside V in *Siraitia grosvenorii* with micelle-mediated cloud-Point extraction. Phytochemical Analysis 24: 381–385. https://doi.org/10.1002/pca.2420 PMID: 23349010
6. Itkin M, Davidovich-Rikanati R, Cohen S, Portnoy V, Doron-Faigenboim A, Oren E, et al. (2016) The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii*. Proceedings of the National Academy of Sciences of the United States of America 113: E7619–E7628. https://doi.org/10.1073/pnas.1604828113 PMID: 27821754
7. Zhong S, Lin W, Tang B (1993) Characteristics and cultivation of *Siraitia siamensis*. Guangxi Argricultural Sciences 5: 218–218.

8. Kasai R, Nie RL, Nashi K, Ohtani K, Zhou J, Tao GD, et al. (1989) Sweet cucurbitane glycosides from fruits of *Siraitia siamensis* (chi-zi luo-han-guo), a Chinese folk medicine. Agricultural and Biological Chemistry 53: 3347–3349. https://doi.org/10.1271/bbb1961.53.3347

9. Koc S, Isgor BS, Isgor YG, Shomali MN, Yildirim O (2015) The potential medicinal value of plants from Asteraceae family with antioxidant defense enzymes as biological targets. Pharmaceutical Biology 53: 746–751. https://doi.org/10.3109/13880209.2014.942788 PMID: 25339240

10. Misra RS, Sriram S, Govil JN, Pandey J, Shivakumar BG, Singh VK (2002) Medicinal value and export potential of tropical tuber crops. Crop Improvement Production Technology Trade and Commerce: 376–386.

11. Zheng Y, Zhang WJ, Wang XM (2013) Triptolide with potential medicinal value for diseases of the central nervous system. Cns Neuroscience and Therapeutics 19: 76–82. https://doi.org/10.1111/cns.12039 PMID: 23253124

12. Rivière-Dobigny T, Doan LP, Quang NL, Maillard JC, Michaux J (2010) Species identification, molecular sexing and genotyping using non-invasive approaches in two wild Bovids species: *Bos gaurus* and *Bos javanicus*. Zoo Biology 28: 127–136.

13. Caroline T, Segatto ALA, Júlia B, Bonatto SL, Freitas LB (2015) Genetic differentiation and hybrid identification using microsatellite markers in closely related wild species. AoB Plants 7: plv084. https://doi.org/10.1093/aobpla/plv084 PMID: 26187606

14. Pharmacopoeia of the People's of Republic China 2015 Edition. In: Commission CP, editor. Beijing: China Medical Science Press.

15. Lu AM, Zhang ZY (1984) The genus *Siraitia* Merr. in China. Guihaia 4: 27–33.

16. Ma XJ, Mo CM, Bai LH, Feng SX (2008) A New *Siraitia grosvenorii* Cultivar 'Yongqing 1'. Acta Horticulturae Sinica 35: 1855–1855. https://doi.org/10.16420/j.issn.0513-353x.2008.12.028

17. Trobajo R, Mann DG, Clavero E, Evans KM, Vanormelingen P, Mcgregor RC (2010) The use of partial cox1, rbcL and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). European Journal of Phycology 45: 413–425. https://doi.org/10.1080/09670262.2010.498586

18. Turenne CY, Sanche SE, Hoban DJ, Karlowsky JA, Kabani AM (1999) Rapid identification of fungi by using the ITS2 genetic region and an automated fluorescent capillary electrophoresis system. Journal of Clinical Microbiology 37: 1846–1851. PMID: 10325335

19. Liu C, Liang D, Gao T, Pang X, Song J, Yao H, et al. (2011) PTIGS-IdIt, a system for species identification by DNA sequences of the psbA-trnH intergenic spacer region. BMC Bioinformatics 12: S4. https://doi.org/10.1186/1471-2105-12-S13-S4 PMID: 22373238

20. Park J (2001) The Cell: A molecular approach, second edition. Yale Journal of Biology and Medicine 74: 361–365. https://doi.org/10.1016/0014-5793(78)81037-0

21. Jansen RK, Ruhlman TA (2012) Plastid genomes of seed plants. Genomics of Chloroplasts and Mitochondria 35: 103–126. https://doi.org/10.1007/978-94-007-2920-9_5

22. Etienne D, Sota F, Catherine FS, Mark B, Ian S (2011) Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. Molecular Biology and Evolution 28: 2077–2086. https://doi.org/10.1093/molbev/msr028 PMID: 21289370

23. Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7: 84. https://doi.org/10.1186/1741-7007-7-84 PMID: 19954512

24. Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, et al. (2015) Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. Plos One 10: e0117159. https://doi.org/10.1371/journal.pone.0117159 PMID: 26061692

25. Park I, Yang S, Kim WJ, Noh P, Lee HO, Moon BC (2018) The complete chloroplast genomes of Six *Ipomoea* species and indel marker development for the discrimination of authentic Pharbitidis Semen (seeds of *I. nil* or *I. purpurea*). Frontiers in Plant Science 9: 965. https://doi.org/10.3389/fpls.2018.00965 PMID: 30026751

26. Park I, Yang S, Choi G, Kim WJ, Moon BC (2017) The complete chloroplast genome sequences of *Aconitum pseudolaeve* and *Aconitum longecassidatum*, and development of molecular markers for distinguishing species in the *Aconitum* subgenus Lycoctonum. Molecules 22: 2012. https://doi.org/10.3390/molecules22112012 PMID: 29160852

27. Bolger AM, Marc L, Bjoern U (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

28. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

29. Anton B, Sergey N, Dmitry A, Gurevich AA, Mikhail D, Kulikov AS, et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

30. Krumsiek J, Rattei AT (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026–1028. https://doi.org/10.1093/bioinformatics/btm039 PMID: 17309896

31. Swindell SR, Plasterer TN (1997) SEQMAN. Contig assembly. Methods in Molecular Biology 70: 75–89. https://doi.org/10.1385/0-89603-358-9:75 PMID: 9089604

32. Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Research 2016: gkw955. https://doi.org/10.1093/nar/gkw955 PMID: 28204566

33. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255. https://doi.org/10.1093/bioinformatics/bth352 PMID: 15180927

34. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, et al. (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13: 715. https://doi.org/10.1186/1471-2164-13-715 PMID: 23256920

35. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25: 955–964. https://doi.org/10.1093/nar/25.5.955 PMID: 9023104

36. Liying C, Narayanan V, Alexander R, Kerr W, Jansen RK, Jim LM, et al. (2006) ChloroplastDB: the Chloroplast Genome Database. Nucleic Acids Research 34: D692–D696. https://doi.org/10.1093/nar/gkj055 PMID: 16381961

37. Ed L, Nomi HM, Gibson, Raymond C, Suzanna L (2009) Apollo: a community resource for genome annotation editing. Bioinformatics 25: 1836–1837. https://doi.org/10.1093/bioinformatics/btp314 PMID: 19439563

38. Dean L, Bjorn C (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research 32: 11–16. https://doi.org/10.1093/nar/gkh152 PMID: 14704338

39. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15: 1281–1295. https://doi.org/10.1093/nar/15.3.1281 PMID: 3547335

40. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2 PMID: 10827456

41. Itaya H (2013) GEMBASSY: an EMBOSS associated software package for comprehensive genome analyses. Source Code for Biology and Medicine 8: 17. https://doi.org/10.1186/1751-0473-8-17 PMID: 23987304

42. Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15: 426–427. https://doi.org/10.1093/bioinformatics/15.5.426 PMID: 10366664

43. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33: 2583–2585. https://doi.org/10.1093/bioinformatics/btx198 PMID: 28398459

44. Dubchak I, Ryaboy DV (2005) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. Methods in Molecular Biology 338: 69–89. https://doi.org/10.1385/1-59745-097-9:69

45. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046–1047. https://doi.org/10.1093/bioinformatics/16.11.1046 PMID: 11159318

46. Librado P, Rozas J (2009) Dnasp v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

47. Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Current Protocols Bioinformatics Chapter 2: Unit 2.3. https://doi.org/10.1002/0471250953.bi0203s00 PMID: 18792934

48. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW (2019) EasyCodeML: A visual tool for analysis of selection using CodeML. Ecology and Evolution: 1–8. https://doi.org/10.1038/s41559-018-0779-9

49. Kazutaka K, Kei-Ichi K, Hiroyuki T, Takashi M (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33: 511–518. https://doi.org/10.1093/nar/gki198 PMID: 15661851

50. Tippmann HF (2004) Analysis for free: Comparing programs for sequence analysis. Brief Bioinform 5: 82–87. https://doi.org/10.1093/bib/5.1.82 PMID: 15153308

51. Posada D (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818. https://doi.org/10.1093/bioinformatics/14.9.817 PMID: 9918953

52. Swofford DL (2002) PauP*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4.0b10 ed. MA, USA.

53. Lalitha S (2000) Primer Premier 5. Biotech software and internet report 1: 270–272. https://doi.org/10.1089/152791600459894

54. Zhang X, Zhou T, Kanwal N, Zhao Y, Bai G, Zhao G (2017) Completion of eight *Gynostemma* BL. (Cucurbitaceae) chloroplast genomes: characterization, comparative analysis, and phylogenetic relationships. Frontiers in Plant Science 8: 1583. https://doi.org/10.3389/fpls.2017.01583 PMID: 28955369

55. Qian J, Song JY, Gao HH, Zhu YJ, Xu J, Pang XH, et al. (2013) The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. Plos One 8: e57607. https://doi.org/10.1371/journal.pone.0057607 PMID: 23460883

56. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, et al. (2017) Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. Molecules 22: 1330. https://doi.org/10.3390/molecules22081330 PMID: 28800082

57. Clegg MT, Gaut BS, Learn GH, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. Proceedings of the National Academy of Sciences of the United States of America 91: 6795–6801. https://doi.org/10.1073/pnas.91.15.6795 PMID: 8041699

58. Jiang M, Chen H, He S, Wang L, Chen AJ, Liu C (2018) Sequencing, characterization, and comparative analyses of the plastome of *Caragana rosea* var. *rosea*. International Journal of Molecular Science 19: 1419. https://doi.org/10.3390/ijms19051419 PMID: 29747436

59. Zhou J, Chen X, Cui Y, Sun W, Li Y, Wang Y, et al. (2017) Molecular structure and phylogenetic analyses of complete chloroplast genomes of two *Aristolochia* medicinal species. International Journal of Molecular Science 18: 1839. https://doi.org/10.3390/ijms18091839 PMID: 28837061

60. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution 13: 660–665. https://doi.org/10.1093/oxfordjournals.molbev.a025626 PMID: 8676740

61. Necsulea A, Lobry J (2007) A new method for assessing the effect of replication on DNA base composition asymmetry. Molecular Biology and Evolution 24: 2169–2179. https://doi.org/10.1093/molbev/msm148 PMID: 17646257

62. Tillier ER, Collins RA (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. Journal of Molecular Evolution 50: 249–257. https://doi.org/10.1007/s002399910029 PMID: 10754068

63. Clarke AK, Gustafsson P, Lidholm JÅ (1994) Identification and expression of the chloroplast *clpP* gene in the conifer *Pinus contorta*. Plant Molecular Biology 26: 851–862. https://doi.org/10.1007/bf00028853 PMID: 7999999

64. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD (2014) The chloroplast *ycf3* and *ycf4* open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex. Embo Journal 16: 6095–6104. https://doi.org/10.1093/emboj/16.20.6095 PMID: 9321389

65. Rose AB (2008) Intron-mediated regulation of gene expression. Curr Top Microbiol Immunol 326: 277–290. https://doi.org/10.1007/978-3-540-76776-3_15 PMID: 18630758

66. Sugita M, Sugiura M (1996) Regulation of gene expression in chloroplasts of higher plants. Plant Molecular Biology 32: 315–326. https://doi.org/10.1007/bf00039388 PMID: 8980485

67. Boudreau E, Turmel M (1995) Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. Plant Molecular Biology 27: 351–364. https://doi.org/10.1007/bf00020189 PMID: 7888624

68. Nazareno AG, Carlsen M, Lohmann LG (2015) Complete chloroplast genome of *Tanaecium tetragonolobum*: The first Bignoniaceae plastome. Plos One 10: e0129930. https://doi.org/10.1371/journal.pone.0129930 PMID: 26103589

69. Comeron JM, Aguadé M (1998) An evaluation of measures of synonymous codon usage bias. Journal of Molecular Evolution 47: 268–274. https://doi.org/10.1007/pl00006384 PMID: 9732453

70. Lee S, Weon S, Lee S, Kang C (2010) Relative codon adaptation index, a sensitive measure of codon usage bias. Evolutionary Bioinformatics 2010: 47–55. https://doi.org/10.4137/EBO.S4608 PMID: 20535230

71. Zhou J, Cui Y, Chen X, Li Y, Xu Z, Duan B, et al. (2018) Complete chloroplast genomes of *Papaver rhoeas* and *Papaver orientale*: molecular structures, comparative analysis, and phylogenetic analysis. Molecules 23: 437. https://doi.org/10.3390/molecules23020437 PMID: 29462921

72. Guo S, Guo L, Zhao W, Xu J, Li Y, Zhang X, et al. (2018) Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. Molecules 23: 246. https://doi.org/10.3390/molecules23020246 PMID: 29373520

73. Liu W, Kong H, Zhou J, Fritsch PW, Hao G, Gong W (2018) Complete chloroplast genome of *Cercis chuniana* (Fabaceae) with structural and genetic comparison to six species in Caesalpinioideae. International Journal of Molecular Science 19: 1286. https://doi.org/10.3390/ijms19051286 PMID: 29693617

74. Dong W, Xu C, Cheng T, Lin K, Zhou S (2013) Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. Genome Biology & Evolution 5: 989–997. https://doi.org/10.1093/gbe/evt063 PMID: 23595020

75. Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD, et al. (2018) Comparative analysis of the chloroplast genomes of the Chinese endemic genus *Urophysa* and their contribution to chloroplast phylogeny and adaptive evolution. International Journal of Molecular Science 19: 1847. https://doi.org/10.3390/ijms19071847 PMID: 29932433

76. Guy-Franck R, Alix K, Bernard D (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiology and molecular biology reviews 72: 686–727. https://doi.org/10.1128/MMBR.00011-08 PMID: 19052325

77. Gulcher J (2012) Microsatellite markers for linkage and association studies. Cold Spring Harbor Protocols 2012: 425–432. https://doi.org/10.1101/pdb.top068510 PMID: 22474656

78. Kawabe A, Nukii H, Furihata H (2018) Exploring the history of chloroplast capture in *Arabis* using whole chloroplast genome sequencing. International Journal of Molecular Science 19: 602. https://doi.org/10.3390/ijms19020602 PMID: 29463014

79. Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, et al. (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. Theoretical and Applied Genetics 95: 163–173. https://doi.org/10.1007/s001220050544

80. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S (2015) Plant DNA barcoding: from gene to genome. Biol Rev Camb Philos Soc 90: 157–166. https://doi.org/10.1111/brv.12104 PMID: 24666563

81. Lu Q, Ye W, Lu R (2018) Phylogenomic and comparative analyses of complete plastomes of *Croomia* and *Stemona* (Stemonaceae). International Journal of Molecular Science 19: 2383. https://doi.org/10.3390/ijms19082383 PMID: 30104517

82. Khan A, Khan IA, Asif H, Azim MK (2010) Current trends in chloroplast genome research. African Journal of Biotechnology 9: 3494–3500. https://doi.org/10.1186/1471-2105-11-321

83. Zhang X, Zhou T, Yang J, Sun JJ, Ju MM, Zhao YM, et al. (2018) Comparative analyses of chloroplast genomes of Cucurbitaceae species: lights into selective pressures and phylogenetic relationships. Molecules 23: 2165. https://doi.org/10.3390/molecules23092165 PMID: 30154353

84. Group CPW (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America 106: 12794–12797. https://doi.org/10.1073/pnas.0905845106 PMID: 19666622

85. Lei W, Ni D, Wang Y, Shao J, Wang X, Yang D, et al. (2016) Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. Scientific Reports 6: 21669. https://doi.org/10.1038/srep21669 PMID: 26899134

86. Jansen RK, Raubeson LA, Boore JL, Depamphilis CW, Chumley TW, Haberle RC, et al. (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. Methods in Enzymology 395: 348–384. https://doi.org/10.1016/S0076-6879(05)95020-9 PMID: 15865976

87. Li Y, Zhou J, Chen X, Cui Y, Xu Z, Li Y, et al. (2017) Gene losses and partial deletion of small single-copy regions of the chloroplast genomes of two hemiparasitic *Taxillus* species. Scientific Reports 7: 12834. https://doi.org/10.1038/s41598-017-13401-4 PMID: 29026168

88. Ohta T (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. Journal of Molecular Evolution 40: 56–63. https://doi.org/10.1007/bf00166595 PMID: 7714912

89. Chen Z, Schertz KF, Mullet JE, Dubell A, Hart GE (1995) Characterization and expression of *rpoC2* in CMS and fertile lines of sorghum. Plant Molecular Biology 28: 799–809. https://doi.org/10.1007/bf00042066 PMID: 7640353

90. Hiroshi K, Pal M (2003) The plastid clpP1 protease gene is essential for plant development. Nature 425: 86–89. https://doi.org/10.1038/nature01909 PMID: 12955146

91. Maliga P (2004) Plastid transformation in higher plants. Annual Review of Plant Biology 55: 289–313. https://doi.org/10.1146/annurev.arplant.55.031903.141633 PMID: 15377222

92. Tseng CC, Sung TY, Li YC, Hsu SJ, Lin CL, Hsieh MH (2010) Editing of *accD* and *ndhF* chloroplast transcripts is partially affected in the *Arabidopsis vanilla* cream1 mutant. Plant Molecular Biology 73: 309–323. https://doi.org/10.1007/s11103-010-9616-5 PMID: 20143129

93. Madoka Y, Tomizawa KI, Mizoi J, Nishida I, Nagano Y, Sasaki Y (2002) Chloroplast transformation with modified accD operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. Plant and Cell Physiology 43: 1518–1525. https://doi.org/10.1093/pcp/pcf172 PMID: 12514249

94. Li J (1993) A revision of the genus *Siraitia* Merr. and two new genera of Cucurbitaceae. Acta Phytotax Sinica 31: 45–55.