Chinese Herbal Medicines 12 (2020) 247–256

Contents lists available at ScienceDirect

Chinese Herbal Medicines

journal homepage: www.elsevier.com/locate/chmed

Original Article

Comparing chloroplast genomes of traditional Chinese herbs *Schisandra sphenanthera* and *S. chinensis*

Xue-ping Wei [a,b,1], Hui-juan Li [a,1], Peng Che [a], Hao-jie Guo [a], Ben-gang Zhang [a,b], Hai-tao Liu [a,b], Yao-dong Qi [a,b,*]

[a] *Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicine, Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China*

[b] *Engineering Research Center of Tradition Chinese Medicine Resource, Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China*

ARTICLE INFO

Article history:
Received 26 June 2019
Revised 17 September 2019
Accepted 26 September 2019
Available online 22 June 2020

Keywords:
chloroplast genome
Nan-Wuweizi
Schisandra chinensis (Turcz.) Baill
Schisandra sphenanthera Rehd. et Wils.
Wuweizi

ABSTRACT

Objective: *Schisandra sphenanthera* and *S. chinensis* are the two important medicinal plants that have long been used under the names of "Nan-Wuweizi" and "Wuweizi", respectively. The misuse of "Nan-Wuweizi" and "Wuweizi" in herbal medical products calls for an accurate method to distinguish these herbs. Chloroplast (cp) genomes have been widely used in species delimitation and phylogeny due to their uniparental inheritance and lower substitution rates than that of the nuclear genomes. To develop more efficient DNA markers for distinguishing *S. sphenanthera*, *S. chinensis*, and the related species, we sequenced the cp genome of *S. sphenanthera* and compared it to that of *S. chinensis*.

Methods: The cp genome of *S. sphenanthera* was sequenced at the Illumina HiSeq platform, and the reference-guided mapping of contigs was obtained with a *de novo* assembly procedure. Then, comparative analyses of the cp genomes of *S. sphenanthera* and *S. chinensis* were carried out.

Results: The cp genome of *S. sphenanthera* was 146 853 bp in length and consisted of a large single copy (LSC) region of 95 627 bp, a small single copy (SSC) region of 18 292 bp, and a pair of inverted repeats (IR) of 16 467 bp. GC content was 39.6%. A total of 126 functional genes were predicted, of which 113 genes were unique, including 79 protein-coding genes, 30 transfer RNA (tRNA) genes, and four ribosomal RNA (rRNA) genes. Five tRNA, four protein-coding genes, and all rRNA were duplicated in the IR regions. There were 18 intron-containing genes, including six tRNA genes and 12 protein-coding genes. In addition, 45 SSRs were detected. The whole cp genome of *S. sphenanthera* was 123 bp longer than that of *S. chinensis*. A total of 474 SNPs and 97 InDels were identified. Five genetic regions with high levels of variation ($Pi > 0.015$), *trn*S-*trn*G, *ccs*A-*ndh*D, *psb*I-*trn*S, *trn*T-*psb*D and *ndh*F-*rpl*32 were revealed.

Conclusion: We reported the cp genome of *S. sphenanthera* and revealed the SNPs and InDels between the cp genomes of *S. sphenanthera* and *S. chinensis*. This study shed light on the species identification and further phylogenetic study within the genus of *Schisandra*.

1. Introduction

Schisandra Mich. (Schisandraceae) is a genus with disjunctive distribution between East Asia and North America. It consists of 23 species based on Saunders' treatment (Saunders, 2000). Only one species, *S. glabra*, is indigenous to North America, whereas all other species distribute in East and Southeast Asia, and Far Eastern Siberia (Saunders, 2000). The fruit of the species from China are typically used as folk herbs, especially *Schisandra sphenanthera* Rehd. et Wils. and *S. chinensis* (Turcz.) Baill, which are the source plants of "Nan-Wuweizi" (*Schisandrae Sphenantherae Fructus*) and "Wuweizi" (*Schisandrae Chinensis Fructus*) recorded in the Chinese Pharmacopoeia (National Pharmacopoeia Committee, 2015). Despite of the similar use in traditional Chinese medicine, including arresting discharge, replenishing *qi*, promoting fluid secretion, toning the kidney, and inducing sedation, there is difference between their chemical compositions (Lu & Chen, 2009). The medical efficacy of the two herbs has been considered different since the Ming Dynasty (1368–1644 CE). For example, *Enlightening Primer of Materia Medica* (*Ben Cao Meng Quan* in Chinese), an ancient

* Corresponding author.
E-mail address: ydqi@implad.ac.cn (Y.-d. Qi).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chmed.2019.09.009
1674-6384/

book of herbal medicine written by Jia-mo Chen from Ming Dynasty, pointed out that "Nan-Wuweizi" was used to treat wind-cold cough while "Wuweizi" was better for the treatment of consumptive damage. The different pharmacodynamics imply that they should not be treated as the same drug. However, the fruits of *Schisandra* are all red berries and nearly all of them, including *S. sphenanthera* and *S. chinensis*, were used as crude drugs. The diagnostic morphological characters, such as tepal color, number and shape of stamens, and gynoecium, typically are identified in living plants, thus it is difficult to differentiate the source plants of crude drug and traditional Chinese medicine products. Some previous studies attempted to resolve this problem using gene fragments. For example, a recent scheme using a combination of ITS + *trn*H-*psb*A + *mat*K + *rbc*L as the most ideal DNA barcode to determine the medicinal plants of Schisandraceae was proposed by Zhang et al. (2015). However, the results revealed that this combination of DNA barcodes was not suitable for most species of *Schisandra* due to its low variation.

Chloroplasts (cp) are the key plastid organelles in nearly all terrestrial plants and algae (Neuhaus & Emes, 2000). Cp genomes of the sequenced species usually have a quadripartite structure, which includes two identical copies of a large inverted repeat (IR) sequence separated by a small single-copy (SSC) and a large single-copy (LSC) region. Variations among the cp genomes involves the plastome size and structure (Howe, 2016). Cp genomes, with the features of uniparental inheritance and lower substitution rates than that of the nuclear genomes, have been widely used in species identification, phylogenetics, and genetic engineering in previous studies (e.g., Ravi et al., 2008; Parks et al., 2009; Daniell et al., 2016; Sabater, 2018).

Here we sequenced the cp genome of *S. sphenanthera* and compared it to the cp genome of *S. chinensis*, which have been reported in our previous study (Guo et al., 2017). The aim was to find more effective molecular markers to reveal more accurate interspecific relationships, and also to identify medicinal material within *Schisandra*.

2. Materials and methods

2.1. Plant materials

Fresh leaves of *Schisandra sphenanthera* were collected from Ankang, Shaanxi Province, China (Collection No.: SX2016101403; 109°0′932″E, 32°53′03″N). The voucher specimen was deposited in the herbarium of the Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College (IMD), China.

2.2. DNA extraction and sequencing

Total genomic DNA was extracted from silica gel-dried leaves using the Plant Genomic DNA Kit (Tiangen Biotech, Beijing, China) following the manufacturer's instructions. The Illumina HiSeq platform was used to sequence the *S. sphenanthera* genome with a paired-end (PE) 150 genomic library by Beijing Novogene Bioinformatics Technology Co., Ltd. (Beijing, China).

2.3. Genome assembly and annotation

Clean reads were obtained after filtering out reads of low quality, such as the reads with more than 50% bases which mass value less than 5 or the reads with more than 10% "N". The BLAST+ program (Camacho et al., 2009) was used to capture *S. sphenanthera* cp reads by comparing them with the cp sequence of *S. chinensis* (Accession No. KY111264) as a reference. The cp reads were assembled by SOAPdenovo2 (Luo et al., 2012). SSPACE-STANDARD-3.0 was used to extend cp contigs and build scaffolds (Boetzer et al., 2011). More cp DNA reads were extracted from total DNA reads with the scaffolds as a reference in order to obtain a complete cp genome sequence. The cp genome sequence was confirmed by mapping the total DNA reads in Geneious 10.0.2 (https://www.geneious.com/).

Chloroplast genome annotation was performed with CPGAVAS (Liu et al., 2012), and the annotation result was manually inspected by Apollo (Lee et al., 2009). The tRNA gene boundaries were corroborated by using tRNAscan-SE 1.21 (Schattner et al., 2005). The cp genome map was drawn using the OGDRAW program (Lohse et al., 2013). The complete chloroplast genome sequence was deposited in GenBank, receiving the accession number MK193856.

2.4. Repeat sequence and SSRs analyses

We investigated the distribution of SSRs located in the cp genome of *S. sphenanthera* and verified all of the repeats manually. The tandem repeat structure was detected using Tandem Repeats Finder (TRF) v4.04 (http://tandem.bu.edu/trf/trf.html) with the default parameters (Benson, 1999). Short dispersed repeats (SDRs) were identified by REPuter (http://tandem.bu.edu/trf/trf.html) including forward and palindromic repeats with the minimal repeat size of 30 bp and hamming distance = 3. Low complexity and nested repeats were removed manually. The MISA program (http://tandem.bu.edu/trf/trf.html) was used to exploit potential SSRs. Motif sizes of 10, 5, 5, 3, 3, and 3 were set as the minimum repeats for mono-, di-, tri-, tetra-, penta-, and hexa- nucleotides, respectively.

2.5. Comparative genome analysis

The mVISTA program was used to compare the cp genome of *S. sphenanthera, S. chinensis*, *Illicium oligandrum* Merr. et Chun and other three basal angiosperm species in a Shuffle-LAGAN mode (Frazer et al., 2004). Cp genomes of *S. chinensis* (KY111264), *I. oligandrum* (EF380354), *Nymphaea alba* L. (AJ627251), *Amborella trichopoda* Baill. (AJ506156) and *Trithuria inconspicua* Cheeseman (HE963749) were downloaded from GenBank. *S. sphenanthera* was set as the reference. IR expansion/contraction was also analyzed. In addition, DnaSP v.5 (Librado & Rozas, 2009) was used to obtain the SNPs and InDels between *S. sphenanthera* and *S. chinensis*.

2.6. Sliding window analysis of cp genomes

Sequences of *S. sphenanthera* and *S. chinensis* were aligned and adjusted manually using BioEdit v.7.1.11 (Hall et al., 2011). Then the DnaSP v.5 (Librado & Rozas, 2009) was used to conduct a sliding window analysis to assess the variability of these two cp genomes with 200-bp step size and 600-bp window length.

3. Results

3.1. Chloroplast genome assembly, organization, and gene content of S. sphenanthera

The complete cp genome of *S. sphenanthera* was 146 853 bp in length with the typical circular quadripartite structure of angiosperm cp genomes (Fig. 1). The LSC (95 627 bp) and SSC regions (18 292 bp) were separated by a pair of inverted repeat regions (IRA and IRB, 16 467 bp) (Fig. 1, Table 1). The GC content of the complete cp genome, LSC, SSC, and IR region of *S. sphenanthera* were 39.60%, 38.40%, 35%, and 45.6%, respectively. The relatively

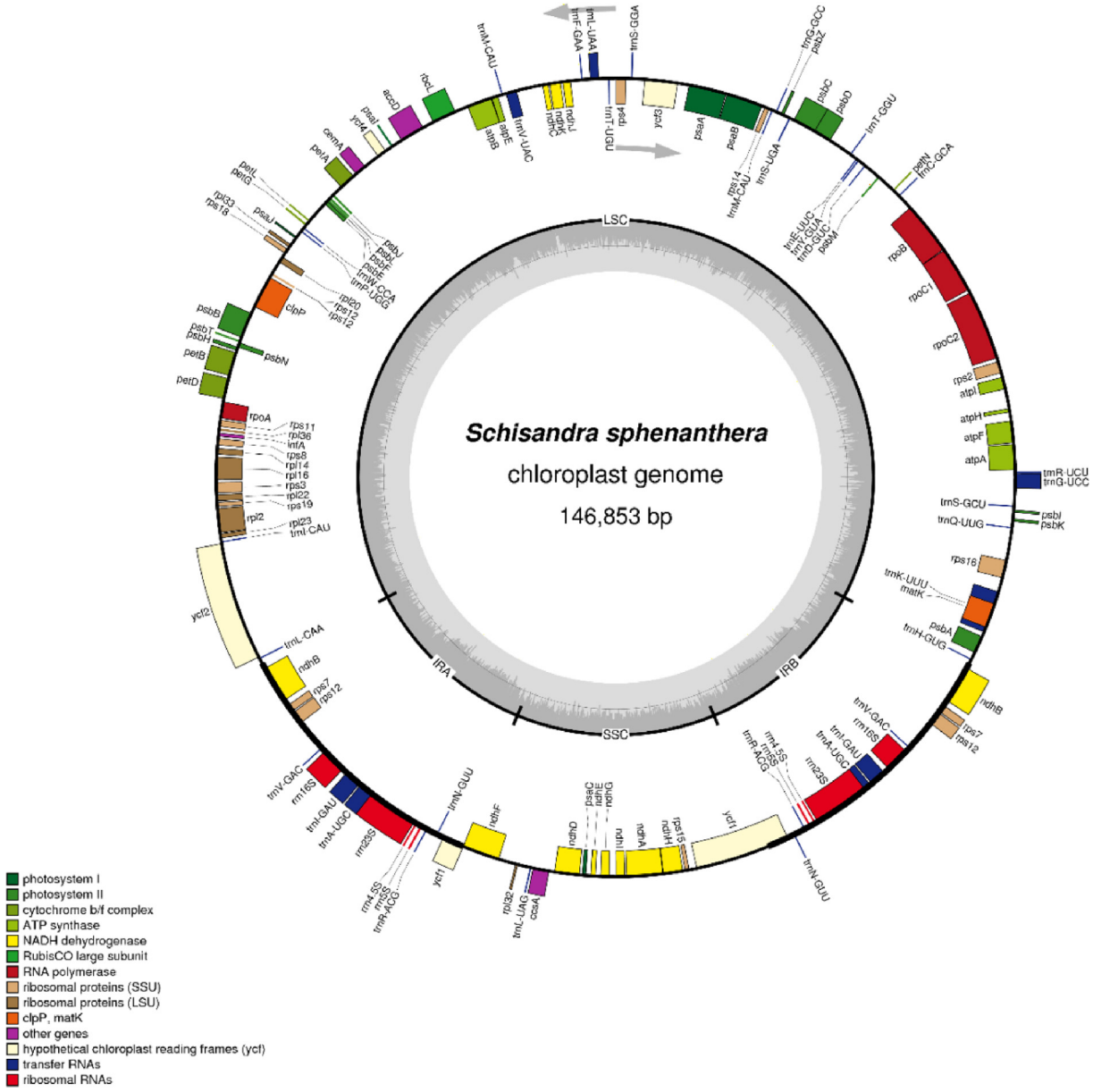

Fig. 1. Chloroplast genome map of *S. sphenanthera*. The gray arrow represents the direction in which the genes are translated. Genes shown outside of the circle are transcribed clockwise, while those inside are counterclockwise. Large single copy (LSC), small single copy (SSC), and inverted repeats (IRA, IRB) are indicated. The darker gray represents GC content in the inner circle, conversely the lighter one represents AT content.

Table 1
Characteristics of cp genomes of *Schisandra sphenanthera* and *S. chinensis*.

Names	*S. sphenanthera*		*S. chinensis*	
	Length (bp)/percent (%)	GC content/%	Length (bp)/percent (%)	GC content/%
Total	146 853	39.6	146 730	39.7
LSC	95627/65.1	38.4	95538/65.1	38.6
SSC	18292/12.5	35.0	18270/12.5	35.0
IR	16467/11.2	45.6	16461/11.2	45.7
CDS	72917/49.7	39.4	72837/49.6	39.4
1st position	24304/16.5	46.8	24279/16.5	46.8
2nd position	24305/16.5	39.3	24279/16.5	49.3
3rd position	24306/16.5	32.1	24279/16.5	32.1

high GC content in the IR regions also occurs in most other plants because of the high GC content of the four ribosomal RNA (rRNA) genes (Cheng et al., 2017). GC content was unevenly distributed throughout the entire cp genome, which could be related to the divergence of the conserved property between IR and SC regions (Yang et al., 2014).

Table 2
Gene contents in cp genome of *S. sphenanthera*.

Functions	Family names	Codes	List of genes
Genes for photosynthesis	Subunits of ATP synthase	*atp*	*atp*A, *atp*B, *atp*E, *atp*F+, *atp*H, *atp*I
	Subunits of NADH-dehydrogenase	*ndh*	*ndh*A+, *ndh*B*+, *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of cytochrome *b*/f complex	*pet*	*pet*A, *pet*B+, *pet*D+, *pet*G, *pet*L, *pet*N
	Subunits of photosystem I	*psa*	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunits of photosystem II	*psb*	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z
	Subunit of rubisco	*rbc*	*rbc*L
Other genes	Subunit of Acetyl-CoA-carboxylase	*acc*	*acc*D
	c-type cytochrom synthesis gene	*ccs*	*ccs*A
	Envelop membrane protein	*cem*	*cem*A
	Protease	*clp*	*clp*P+
	Translational initiation factor	*inf*	*inf*A
	Maturase	*mat*	*mat*K
Self replication	Large subunit of ribosome	*rpl*	*rpl*16+, *rpl*2+, *rpl*14, *rpl*20, *rpl*22, *rpl*23, *rpl*32, *rpl*33, *rpl*36
	DNA dependent RNA polymerase	*rpo*	*rpo*A, *rpo*B, *rpo*C1+, *rpo*C2
	Small subunit of ribosome	*rps*	*rps*18, *rps*15, *rps*16+, *rps*7*, *rps*12*+, *rps*3, *rps*2, *rps*11, *rps*4, *rps*19, *rps*8, *rps*14
	rRNA Genes	*rrn*	*rrn16S**, *rrn23S**, *rrn4.5S**, *rrn5S**
	tRNA Genes	*trn*	*trn*A-UGC*+, *trn*C-GCA, *trn*D-GUC, *trn*E-UUC, *trn*F-GAA, *trn*G-GCC, *trn*G-UCC+, *trn*H-GUG, *trn*I-CAU, *trn*I-GAU*+, *trn*K-UUU+, *trn*L-CAA, *trn*L-UAA+, *trn*L-UAG, *trn*fM-CAU, *trn*M-CAU, *trn*N-GUU*, *trn*P-UGG, *trn*Q-UUG, *trn*R-ACG*, *trn*R-UCU, *trn*S-GCU, *trn*S-GGA, *trn*S-UGA, *trn*T-GGU, *trn*T-UGU, *trn*V-GAC*, *trn*V-UAC+, *trn*W-CCA, *trn*Y-GUA
Genes of unkown function	Conserved open reading frames	*ycf*	*ycf*1*, *ycf*2, *ycf*3+, *ycf*4

Notes: The *rps*12 gene was divided, 5′-*rps*12 was located in the LSC region, and 3′-*rps*12 was located in the IR region.
* Gene located in the IR regions; + Intron-containing gene.

A total of 126 functional genes were predicted in the cp genome of *S. sphenanthera*, of which 113 genes were unique, including 79 protein-coding genes, 30 transfer RNA (tRNA) genes and four rRNA genes (Fig. 1, Table 2). Five tRNA, four protein-coding genes, and all rRNA replication events occurred in the IR regions. The *ycf*15 is present in many angiosperm species but absent in *S. sphenanthera* (Goremykin et al., 2003; Li et al., 2014).

Introns play an important role in the regulation of alternative gene splicing but have no significance in the structure of the translation products. Introns always accumulate more mutations without the pressure of natural selection (Graveley, 2001). Altogether 18 intron-containing genes were found, including six tRNA genes (*trn*K-UUU, *trn*G-UCC, *trn*L-UAA, *trn*V-UAC, *trn*I-GAU, *trn*A-UGC) and 12 protein-coding genes (*rps*16, *atp*F, *rpo*C1, *ycf*3, *clp*P, *pet*B, *pet*D, *rpl*16, *rpl*2, *ndh*B, *ndh*A, *rps*12) (Table 3). Both *ycf*3 and *clp*P contained two introns. In contrast, *rps*12 had no intron but three exons, which indicated the *rps*12 may have an RNA trans-splicing structure. This gene located within the boundaries of the LSC and the IR regions, with its 5′ exon locating in an LSC region and two 3′ exons locating in IR regions. Only one intron was found in the remaining 15 genes, among which *trn*K-UUU has the longest intron (2486 bp) and *trn*L-UAA has the shortest one (482 bp).

The total length of the protein-coding regions (CDS) genes was 72 917 bp, accounting for 49.7% of the whole cp genome of *S. sphenanthera*. The frequency of codon usage was calculated and summarized in Table 1 and Table 4. A total of 10.1% of all codons (2463) encoded leucine, and 1.2% (290) encoded cysteine. Leucine and cysteine were the most and least prevalent amino acids, respectively. The high AT content at the third codon position reflected a codon usage bias of A or T. With the exception of *trn*L-CAA, all of the types of preferred synonymous codons (RSCU >1) ended with A or U. The GC content of the third codon position was lowest (32.1%) in the CDS, which suggested the codon usage bias may have developed during the evolution course. The codon usage bias was generally found in cp genomes of plant and reflected a genomic bias towards a higher A + T content (Clegg et al., 1994).

3.2. Repeat structure and SSRs analyses

The simple sequence repeats (SSRs), also called microsatellites, were highly variable nucleotide arrays composed of 1–6 nucleotide repeat units in tandem (Chen et al., 2006). We obtained a total of 45 SSRs in the cp genome of *S. sphenanthera*. These SSRs included 30 mononucleotide repeats, six dinucleotide repeats, one trinucleotide repeat, five tetranucleotide repeats, and three pentanucleotide repeats. Hexanucleotide repeat was not found (Fig. 2). All mononucleotide SSRs were composed of A (12) and T (17) except one composed of G. Five dinucleotide repeats were composed of A/T motifs and one composed of a T/C motif. The repeat number of mononucleotide motifs ranged from 10 to 13. The largest SSRs were tandem repeats and 23 bp in length. It composed of tetranucleotide repeats of AAAT for three times, and mononucleotide repeats of A for 11 times. SSRs composed of G or C were less frequent than those of A or T, which might have been related to the greater stability of G-C making it difficult to change within the genome. Thirty-seven SSRs (82.22%) were located in the LSC region, six SSRs (13.33%) were located in the SSC region and two SSRs (4.44%) were located in the IR regions. Furthermore, 38 SSRs (84.44%) located in the noncoding regions, among which 31 were in the intergenic regions and seven in introns. Only seven SSRs (15.56%) were found in the gene coding regions (*psb*I, *rpo*C2, *rpo*B, *psb*C, *cem*A, *ycf*2). Consequently, SSRs were more abundant in noncoding regions than that in the CDS base on the above statistics.

Short dispersed repeats (SDRs), with lengths longer than 30 bp, were considered to be one of the major factors promoting the rearrangements of cp genome (Qian et al., 2013). Thirty-three pairs repeats ranging from 30 to 149 bp in length were found in the cp genome of *S. sphenanthera* (Fig. 3). Four forward, 10 palindromic, and 19 tandem repeats were identified. The length of all the forward repeats were 30–60 bp and palindromic repeats were 30–41 bp. Two tandem repeat motifs located in *ycf*2 of the LSC region, one of which contained the longest repeat with the length of 149 bp while the other 18 tandem repeats exhibited only 30 to

Table 3
Genes with introns and length of exons and introns in cp genome of *S. sphenanthera*.

No.	Genes	Location	Exon I/bp	Intron I/bp	Exon II/bp	Intro II/bp	Exon III/bp
1	*trn*K-UUU	LSC	39	2486	34		
2	*rps*16	LSC	41	819	220		
3	*trn*G-UCC	LSC	33	760	61		
4	*atp*F	LSC	148	757	416		
5	*rpoC1*	LSC	456	711	1608		
6	*ycf*3	LSC	127	727	230	756	153
7	*trn*L-UAA	LSC	37	482	50		
8	*trn*V-UAC	LSC	41	571	55		
9	*clp*P	LSC	71	804	294	556	247
10	*pet*B	LSC	6	761	642		
11	*pet*D	LSC	8	690	526		
12	*rpl*16	LSC	9	945	402		
13	*rpl*2	LSC	394	659	431		
14	*ndh*B	IR	869	609	757		
15	*trn*I-GAU	IR	34	939	39		
16	*trn*A-UGC	IR	39	790	35		
17	*ndh*A	SSC	556	1061	542		
18	*rps*12	LSC-IR	114		232		26

Table 4
Codon-anticodon recognition pattern and relative synonymous codon usage (RSCU) for cp genome of *S. sphenanthera*.

Codon	Count	RSCU	tRNA	Codon	Count	RSCU	tRNA
UAU(Y)	697	1.57		UUA(L)	701	1.71	*trn*L-UAA
UAC(Y)	191	0.43	*trn*Y-GUA	UUG(L)	527	1.28	*trn*L-CAA
UGG(W)	431	1	*trn*W-CCA	CUU(L)	492	1.2	
GUU(V)	515	1.47		CUC(L)	197	0.48	
GUC(V)	172	0.49	*trn*V-GAC	CUA(L)	345	0.84	*trn*L-UAG
GUA(V)	486	1.38	*trn*V-UAC	CUG(L)	201	0.49	
GUG(V)	233	0.66		AAA(K)	830	1.41	*trn*K-UUU
ACU(T)	490	1.57		AAG(K)	350	0.59	
ACC(T)	243	0.78	*trn*T-GGU	AUU(I)	954	1.39	
ACA(T)	363	1.16	*trn*T-UGU	AUC(I)	460	0.67	*trn*I-GAU
ACG(T)	151	0.48		AUA(I)	649	0.94	*trn*I-CAU
UCU(S)	517	1.66		CAU(H)	477	1.52	
UCC(S)	305	0.98	*trn*S-GGA	CAC(H)	151	0.48	*trn*H-GUG
UCA(S)	397	1.27	*trn*S-UGA	GGU(G)	572	1.33	
UCG(S)	173	0.55	*trn*S-GCU	GGC(G)	178	0.41	*trn*G-GCC
AGU(S)	360	1.15		GGA(G)	674	1.56	*trn*G-UCC
AGC(S)	121	0.39		GGG(G)	300	0.7	
CGU(R)	323	1.25	*trn*R-ACG	UUU(F)	761	1.17	
CGC(R)	112	0.43		UUC(F)	538	0.83	*trn*F-GAA
CGA(R)	326	1.26		GAA(E)	913	1.44	*trn*E-UUC
CGG(R)	120	0.47		GAG(E)	355	0.56	
AGA(R)	482	1.87	*trn*R-UCU	GAU(D)	760	1.56	
AGG(R)	184	0.71		GAC(D)	213	0.44	*trn*D-GUC
CAA(Q)	607	1.46	*trn*Q-UUG	UGU(C)	218	1.5	
CAG(Q)	224	0.54		UGC(C)	72	0.5	*trn*C-GCA
CCU(P)	397	1.53		GCU(A)	607	1.76	
CCC(P)	226	0.87		GCC(A)	235	0.68	
CCA(P)	301	1.16	*trn*P-UGG	GCA(A)	400	1.16	*trn*A-UGC
CCG(P)	114	0.44		GCG(A)	139	0.4	
AAU(N)	828	1.51		UAA(*)	33	1.19	
AAC(N)	272	0.49	*trn*N-GUU	UAG(*)	23	0.83	
AUG(M)	591	1	*trn*fM-CAU *trn*M-CAU	UGA(*)	27	0.98	

Notes: RSCU: Relative Synonymous Codon Usage.

70 bp in length. The highest percentage of repeats, 22 of 33 pair repeats (two forward repeats; eight palindromic repeats; 12 tandem repeats) were completely distributed in the LSC region.

3.3. Comparative analysis of basal angiosperms

Comparative analysis of the cp genomes of six basal angiosperms was performed with the annotation of *S. sphenanthera* as the reference (Fig. 4; Table 5). The cp genome of *S. sphenanthera* was most similar to that of *S. chinensis* and the most different from that of the basal angiosperm *Trithuria inconspicua*. The average size of analyzed genomes was 155 024 bp. The divergence was higher in the non-coding regions than that of the coding regions. The variation pattern in the length of SC regions was consistent with that of IR regions among all the six species analyzed except *Illicium oligandrum*. The cp genomes of *Trithuria inconspicua* (165 389 bp) and *S. chinensis* (146 730 bp) were the longest and the shortest in length, respectively. The variations of the length of IR regions contributed most to the differences in genome size among species (Table 1).

3.4. IR contraction and expansion

The expansion and contraction of the boundaries between the IR and SC regions were primarily responsible for the size variations in genomes among the angiosperm lineages. We compared the

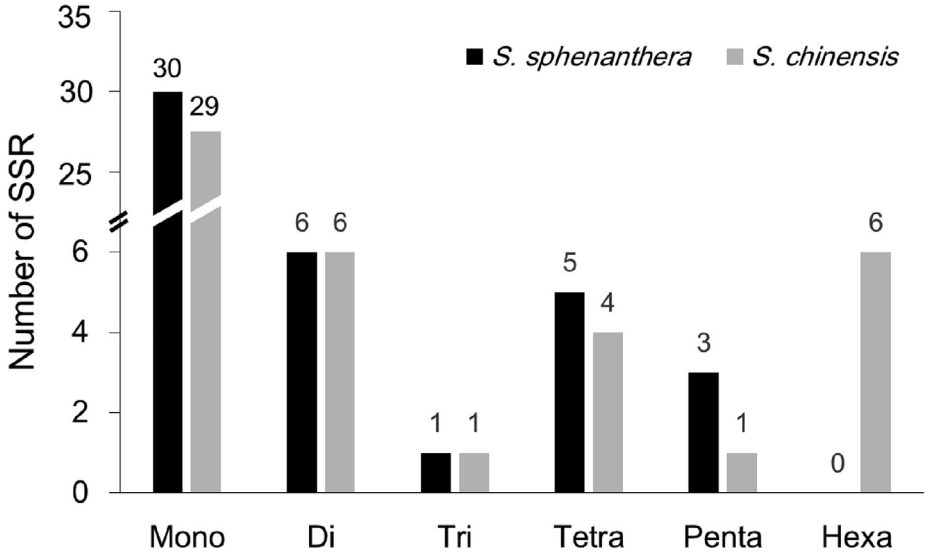

Fig. 2. Analysis of simple sequence repeats (SSRs) in *Schisandra sphenanthera* and *S. chinensis* cp genomes. Mono: Mono-nucleotide; Di: Di-nucleotide; Tri: Tri-nucleotide; Tetra: Tetra-nucleotide; Penta: Penta-nucleotide; Hexa: Hexa-nucleotide.

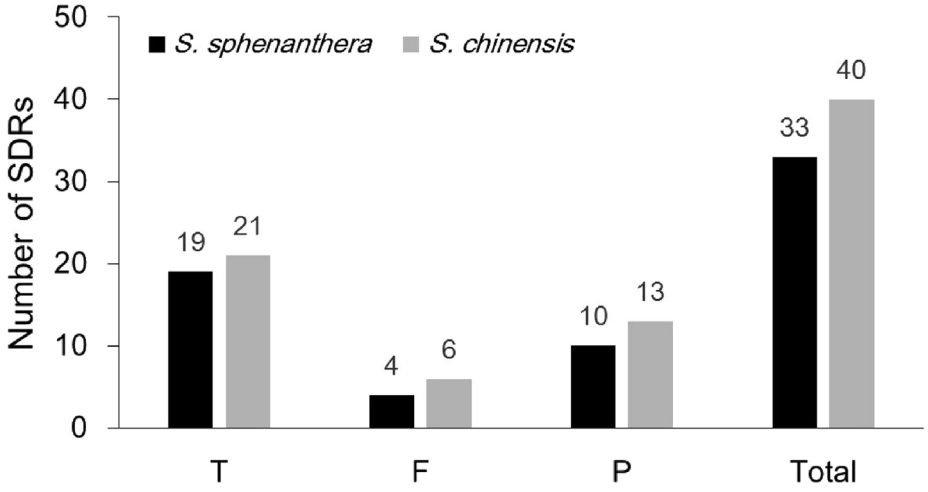

Fig. 3. Analysis of short dispersed repeats (SDRs) in cp genomes of *Schisandra sphenanthera* and *S. chinensis*. T: tandem repeats; F: forward repeats; P: palindromic repeats.

junction of the IR/SC boundaries and their adjacent genes among *S. sphenanthera*, *S. chinensis*, *Illicium oligandrum*, *Nymphaea alba*, *Trithuria inconspicua*, and *Amborella trichopoda* (Fig. 5).

The adjacent genes of IR/SC boundaries of the cp genome of *S. sphenanthera* were the same as those of *S. chinensis* and *Illicium oligandrum*. However, the distance from the genes to the boundaries were different. The adjacent genes of the LSC-IRA (*rps*19, *rpl*2) and IRB-LSC (*rpl*2, *trn*H) of *Nymphaea alba*, *Trithuria inconspicua,* and *Amborella trichopoda* were identical to each other, but different from *S. sphenanthera* (*trn*L, *ndh*B at LSC-IRA; *ndh*B, *trn*H at IRB-LSC). The gene located at the border of IRA-SSC and SSC-IRB in *Trithuria inconspicua* was *ndh*D, whereas it was *ycf*1 in the other species. Obvious contraction happened in the IRA regions, which created the IRA-LSC border between the *trn*L and *ndh*B in *S. sphenanthera*.

At the IRA-SSC border, *ndh*F shared some nucleotides with *ycf*1 in three species (33 bp in *S. sphenanthera*, 112 bp in *S. chinensis*, 11 bp in *Illicium oligandrum*). The complete *ycf*1 appeared to be a pseudogene in the IRA region because it created an incomplete duplication of the normal copy of *ycf*1 when it spanned across the IRB-SSC border.

At the IRB-SSC border, the IRB region expanded by 1283 bp toward *ycf*1 in *S. sphenanthera*, 1281 bp in *S. chinensis*, 413 bp in *Illicium oligandrum*, 155 bp in *Nymphaea alba*, and 1582 bp in *Amborella trichopoda*. The IRB-LSC border spanned across the *trn*H in the cp genome of *S. sphenanthera* and *S. chinensis*, whereas it was located entirely in the LSC region in other species. This result showed the clear contraction of the IRB region of *S. sphenanthera* and *S. chinensis* (Fig. 5).

3.5. Divergence hotspots of *S. sphenanthera* and *S. chinensis*

Basic informations indicated that the cp genomes of *S. sphenanthera* and *S. chinensis* have been well preserved during their long evolution. The cp genomes of *S. sphenanthera* and *S. chinensis* shared the same four-part circular structure and the former was only 123 bp longer than the latter. The GC contents of the cp genomes of these two species were similar. The number of unique genes, protein-coding genes, tRNA genes, and rRNA genes were also identical to each other, as well as the loss of *ycf*15 in *S. sphenanthera* also occurred in *S. chinensis*.

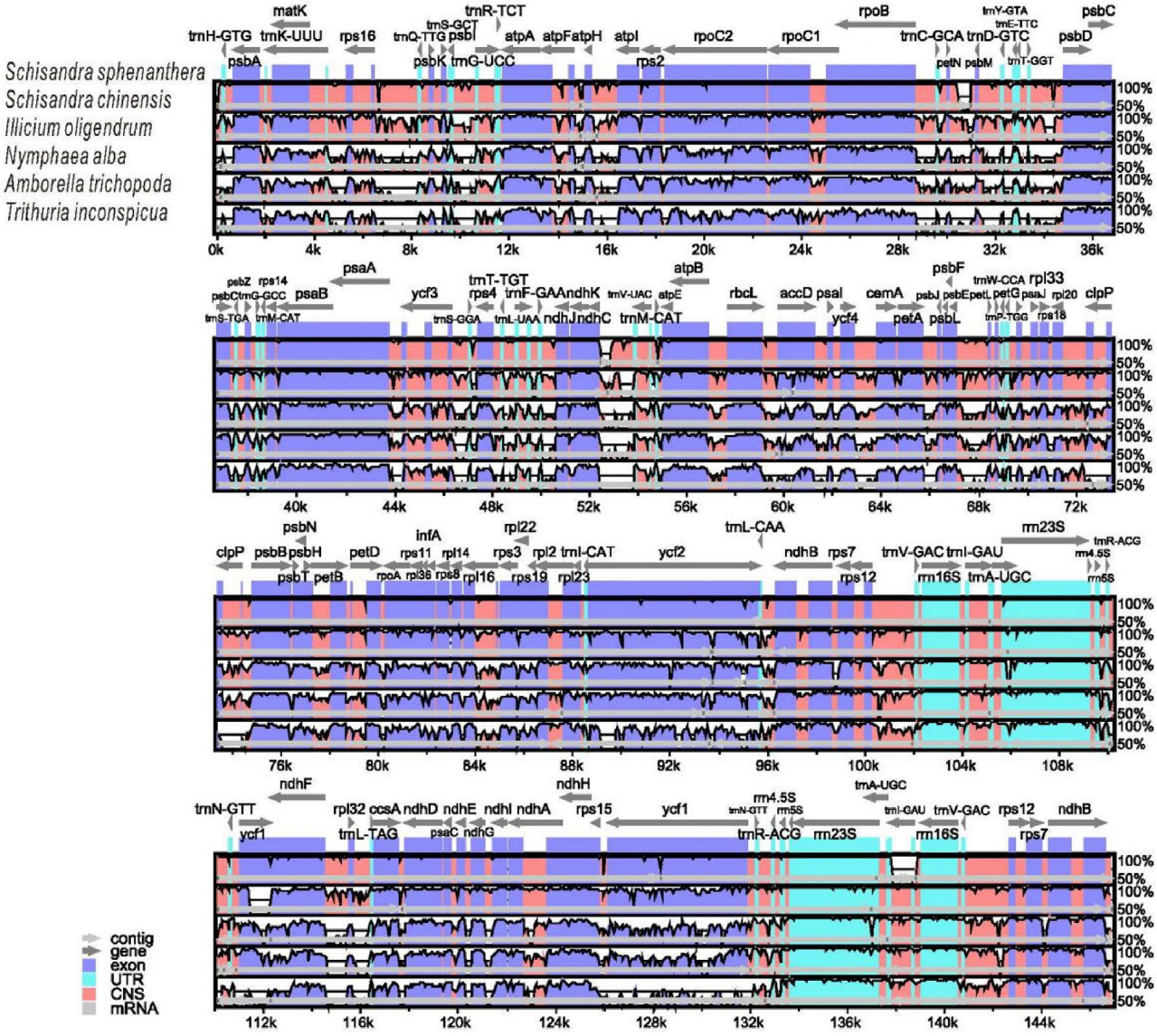

Fig. 4. Sequence identity plots between six sequenced chloroplast genomes, with *Schisandra sphenanthera* as a reference. The vertical scale indicates the identity percentage (50–100%). The horizontal axis corresponds to the coordinates within the chloroplast genome. Annotated genes are displayed along the top.

Table 5
Cp genomes size comparison of six basal angiosperms.

Species	Length/bp			
	Total	LSC	SSC	IR
Schisandra sphenanthera	146 853	95 627	18 292	16 467
Schisandra chinensis	146 730	97 351	20 305	15 058
Illicium oligandrum	148 553	97 144	20 267	15 571
Amborella trichopoda	162 686	90 970	18 414	26 651
Nymphaea alba	159 930	90 014	19 562	25 177
Trithuria inconspicua	165 389	84 468	6354	37 284
Average	155 024	92 596	17 199	22 702

Numbers of SNPs and InDels were analyzed between *S. sphenanthera* and *S. chinensis*. As a result, 474 SNPs and 97 InDels were identified (Table S1), among which 363, 12, and 99 SNPs located in the LSC, IRA/B and SSC region, respectively; 80, five, and 12 InDels located in the LSC IRA/B and SSC region, respectively. The first two longest InDels distributed in the LSC region. The longest InDel (474 bp) was present in *S. sphenanthera* but absent in *S. chinensis*. In contrast, the second InDel (435 bp) existed in *S. chinensis* but absent in *S. sphenanthera*. These SNPs and InDels are all potential markers to distinguish these two species.

The sliding window analysis showed the nucleotide variability values between *S. sphenanthera* and *S. chinensis* differed from 0 to 0.03000 with a mean of 0.00364, which suggested a high sequence similarity. Five regions with much higher variation ($Pi > 0.015$), *trn*S-*trn*G, *ccs*A-*ndh*D, *psb*I-*trn*S, *trn*T-*psb*D and *ndh*F-*rpl*32, were recognized (Fig. 6). Three of these loci were found in the LSC region,

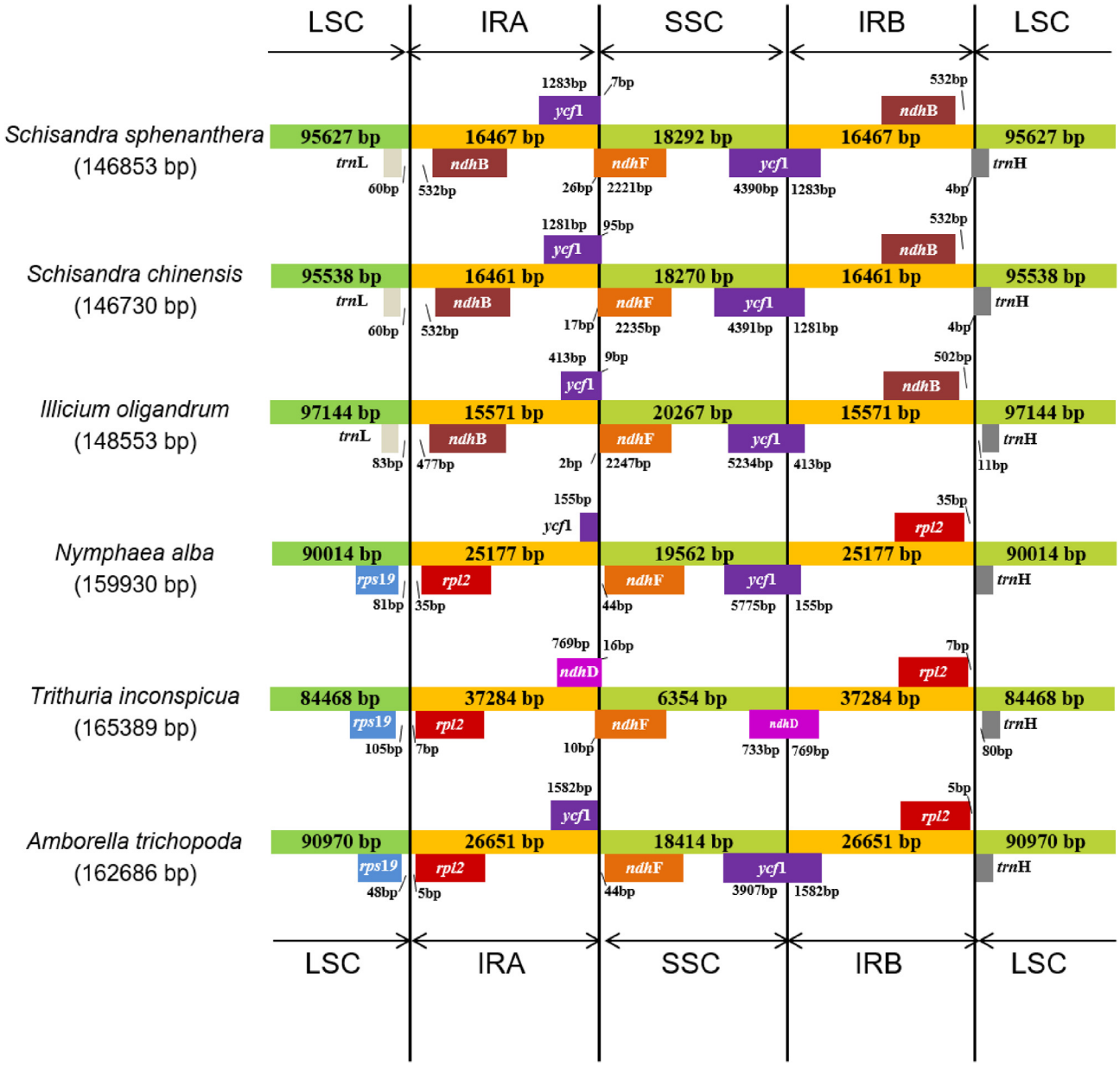

Fig. 5. Comparison at junction of IR/SC boundaries.

two were in the SSC region, but none located in the IR regions. According to the analysis, it's obvious that the SC regions had strikingly higher divergence compared to the IR regions.

4. Discussion

4.1. Cp genomes of S. sphenanthera and S. chinensis are relatively conservative compared with other species of basal angiosperms

Analyzing the cp genome of *S. sphenanthera*, *S. chinensis* and the other four species from basal angiosperm lineages showed that the cp genome of *S. sphenanthera* was most similar to that of *S. chinensis*, and the sizes of the cp genome of *S. sphenanthera* (146 853 bp) and *S. chinensis* (146 730 bp) were relative small when compared to the other four species. The largest and smallest cp genome analyzed in this study were *Trithuria inconspicua* and *S. chinensis* respectively, with the former approximately 18.7 kb larger than the latter. IR region contraction clearly occurred in *S. sphenanthera* and *S. chinensis* on the basis of the full analysis of the adjacent genes on the IR/SC borders and the distance from the genes to the borders. The expansion and contraction of IR boundaries could cause some genes to move into the IR regions or remain in the SC regions. The cp genome of the two species had the similar GC content and the same number of intron-containing genes, functional genes, and rRNAs genes to each other, and shared the lack of *ycf*15 gene. The *ycf*15 gene was first identified in the cp genome of *Nicotiana* (Shi et al., 2013). It has been reported that *ycf*15 may be used as a potential marker to distinguish *Colchicum* from *Gloriosa* since the deletion of *ycf*15 was thought to occur only in *Colchicum* (Nguyen et al., 2015). Previous studies have reported that the loss events also happened in some species of Pteridophyta and Gymnosperm (Kim et al., 2014; Li et al., 2016; Wakasugi et al., 1994).

The repeat regions of cp genomes play an important role in gene recombination and rearrangement (Smith, 2002). Chloroplast SSR

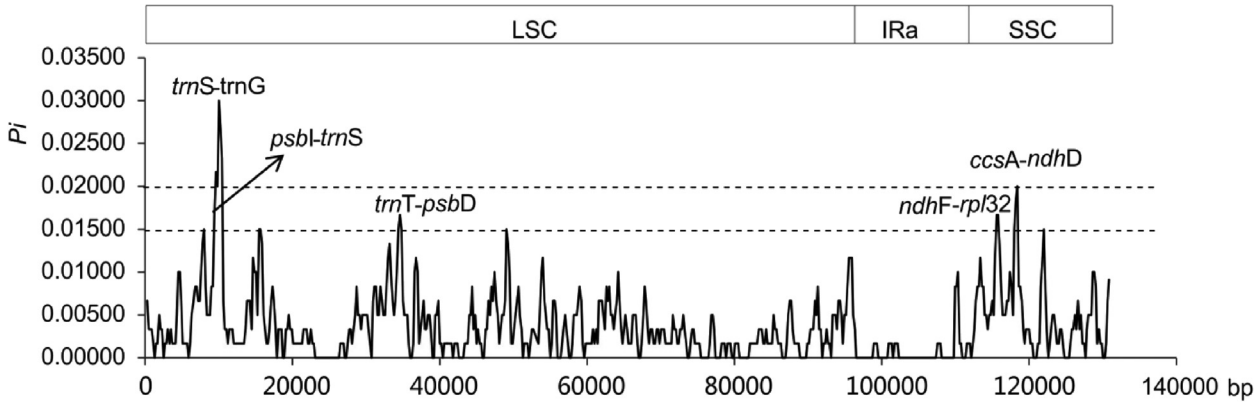

Fig. 6. Sliding window analysis of complete cp genomes of *S. sphenanthera* and *S. chinensis*. Window length: 600 bp, step size: 200 bp.

markers are more effective in genetic studies for population structure analyses as these short repeats are haploid and uniparental inherited (Echt et al., 1998). We obtained 45 SSRs in *S. sphenanthera* and 47 SSRs in *S. chinensis*. The number of SSRs in these two species was similar, but the repeat motifs differed slightly. The abundance of SDRs was related to the extent of gene rearrangement given the fact that most repeats always occurred near the rearrangement hotspots and might mediate these rearrangement events (Chumley et al., 2006; Haberle et al., 2008; Pombert et al., 2005). Short repeat motifs might influence inter-molecular recombination in plastid DNA and create diversity within the population (Kawata et al., 1997). These repeat motifs provided a source of information for additional population genetic studies of *S. sphenanthera*.

4.2. Genomic divergence and hotspot regions are the potential molecular markers

Comparative analyses of the basal angiosperms and the hotspot analysis showed that the sequence divergence in IR regions was lower than that in SC regions. The low sequence divergence in IR regions may be related to the copy correction between IR regions by gene conversion (Khakhlova & Bock, 2006). Both the cp genomes of *S. sphenanthera* and *S. chinensis* were conserved. However, we identified 474 SNPs and 97 InDels through sequence comparison. Empirical phylogenetic studies using SNPs have become commonplace across diverse taxa and studies (Leaché & Oaks, 2017). In addition, five regions with greater variability (*Pi* > 0.015), *trn*S-*trn*G, *ccs*A-*ndh*D, *psb*I-*trn*S, *trn*T-*psb*D, and *ndh*F-*rpl*32, were recognized through sliding window analysis.

To ensure the safety and accuracy of clinical medication of "Nan-Wuweizi" and "Wuweizi", rapid and efficient DNA barcodes were needed to distinguish their source plants. "Nan-Wuweizi" and "Wuweizi" were used in more than 10 and 90 kinds of different traditional herbal medical products, respectively (Committee, 2015). The chemical indicator of "Nan-Wuweizi" is schisantherin A ($C_{30}H_{32}O_9$) while that of "Wuweizi" is schisandrin ($C_{24}H_{32}O_7$). The price of "Wuweizi" is nearly six to nine times higher than that of "Nan-Wuweizi". At present, some "Nan-Wuweizi" were used as the counterfeits products of "Wuweizi". Besides, fruits of the other species of *Schisandra* were also misused as "Nan-Wuweizi" or "Wuweizi" because of their similar morphology. DNA degrades heavily in highly processed materials, making it difficult to conduct accurate molecular identification. This study provided multiple SNPs that can be used to design a short nucleotide signature for distinguishing the TCM "Nan-Wuweizi" and "Wuweizi". Furthermore, these SNPs and InDels, especially those from the five regions with high variation, have the greatest potential for clarifying the interspecific relationships accurately in the phylogenetic study and identifying the medicinal material within *Schisandra*.

5. Conclusion

This study reported the cp genome of *S. sphenanthera* and performed the comparative genome analyses with *S. chinensis*. The cp genomes of the two species were highly similar in sequence and structure. Compared with *S. chinensis*, cp genome of *S. sphenanthera* was 123 bp longer in length, while their IR region, SSRs, and the border circumstances were similar. Ninety-seven InDels and 474 SNPs were identified between *S. chinensis* and *S. sphenanthera*, which could be the short nucleotide signature with the greatest potential to distinguish the processed crude drug or traditional Chinese medicine products. Five highly variable regions, *trn*S-*trn*G, *ccs*A-*ndh*D, *psb*I-*trn*S, *trn*T-*psb*D, and *ndh*F-*rpl*32, could be developed as DNA barcodes for accurate species identification. This study not only facilitated the biological identification of these two important traditional medicinal plants, but also provided plenty information for the further study of *Schisandra*.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (Grant No. 81703650) and the CAMS Initiative for Innovative Medicine (CAMS-I2M) (Grant No. 2016-I2M-2-003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chmed.2019.09.009.

References

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research, 27*, 573–580.
Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics, 27*, 578–579.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: Architecture and applications. *BMC Bioinformatics, 10*, 421.

Smith, T. C. (2002). Chloroplast evolution: Secondary symbiogenesis and multiple losses. *Current Biology, 12*, R62–R64.
Chen, C. X., Zhou, P., Choi, Y. A., Huang, S., & Gmitter, F. G. Jr., (2006). Mining and characterizing microsatellites from citrus ESTs. *Theoretical and Applied Genetics, 112*, 1248–1257.
Cheng, H., Li, J. F., Zhang, H., Cai, B. H., Gao, Z. H., Qiao, Y. S., et al. (2017). The complete chloroplast genome sequence of strawberry (*Fragaria*x *ananassa* Duch.) and comparison with related species of Rosaceae. *PeerJ, 5* e3919.
Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution, 23*, 2175–2190.
Clegg, M., Gaut, B. S., Learn, G., & Morton, B. R. (1994). Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America, 91*, 6795–6801.
Daniell, H., Lin, C. S., Yu, M., & Chang, W. J. (2016). Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biology, 17*, 134.
Echt, C. S., Deverno, L. L., Anzidei, M., & Vendramin, G. G. (1998). Chloroplast microsatellites reveal population genetic diversity in red pine, *Pinus resinosa* Ait. *Molecular Ecology, 7*, 307–316.
Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Research, 32*(Web Server issue), W273–W279.
Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S., & Hellwig, F. H. (2003). Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution, 20*, 1499–1505.
Graveley, B. R. (2001). Alternative splicing: Increasing diversity in the proteomic world. *Trends in Genetics, 17*, 100–107.
Guo, H. J., Liu, J. S., Luo, L., Wei, X. P., Zhang, J., Qi, Y. D., et al. (2017). Complete chloroplast genome sequences of *Schisandra chinensis*: Genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. *Science China Life Sciences, 60*, 1286–1290.
Haberle, R. C., Fourcade, H. M., Boore, J. L., & Jansen, R. K. (2008). Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *Journal of Molecular Evolution, 66*, 350–361.
Hall T., Biosciences I., Carlsbad, & Ca (2011). BioEdit: An important software for molecular biology. GERF Bulletin of Biosciences, 2: 60–61.
Howe C.J. (2016). Chloroplast Genome. In: eLS John Wiley & Sons Ltd.
Kawata, M., Harada, T., Shimamoto, Y., Oono, K., & Takaiwa, F. (1997). Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). *Current Genetics, 31*, 179–184.
Khakhlova, O., & Bock, R. (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *The Plant Journal, 46*, 85–94.
Kim, H. T., Chung, M. G., & Kim, K. J. (2014). Chloroplast genome evolution in early diverged leptosporangiate ferns. *Molecules and Cells, 37*(5), 372–382.
Leaché, A. D., & Oaks, J. R. (2017). The utility of Single Nucleotide Polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics, 48*, 69–84.
Lee, E., Harris, N., Gibson, M., Chetty, R., & Lewis, S. (2009). Apollo: A community resource for genome annotation editing. *Bioinformatics, 25*, 1836–1837.
Li, Q. S., Li, Y., Song, J. Y., Xu, H. B., Xu, J., Zhu, Y. J., et al. (2014). High-accuracy *de novo* assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytologist, 204*, 1041–1049.
Li, Z. H., Qian, Z. Q., Liu, Z. L., Deng, T. T., Zu, Y. M., Zhao, P., et al. (2016). The complete chloroplast genome of Armand pine *Pinus armandii*, an endemic conifer tree species to China. *Mitochondrial DNA, 27*, 2635–2636.
Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics, 25*, 1451–1452.
Liu, C., Shi, L. C., Zhu, Y. J., Chen, H. M., Zhang, J. H., Lin, X. H., 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.
Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). Organellar genome DRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research, 41*, W575–W581.
Lu, Y., & Chen, D. F. (2009). Analysis of *Schisandra chinensis* and *Schisandra sphenanthera*. *Journal of Chromatography A, 1216*, 1980–1990.
Luo, R. B., Liu, B. H., Xie, Y. L., Li, Z. Y., Huang, W. H., Yuan, J. Y., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *GigaScience, 1*, 18.
National Pharmacopoeia Committee. (2015). Chinese Pharmacopoeia. China Medical Science Press, Beijing, 66–67, 244 pp.
Neuhaus, H. E., & Emes, M. J. (2000). Nonphotosynthetic metabolism in plastids. *Annual Review of Plant Physiology and Plant Molecular Biology, 51*, 111–140.
Nguyen, P. A., Kim, J. S., & Kim, J. H. (2015). The complete chloroplast genome of colchicine plants (*Colchicum autumnale* L. and *Gloriosa superba* L.) and its application for identifying the genus. *Planta, 242*, 223–237.
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.
Pombert, J. F., Otis, C., Lemieux, C., & Turmel, M. (2005). The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. *Molecular Biology and Evolution, 22*, 1903–1918.
Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., et al. (2013). The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS ONE, 8* e57607.
Ravi, V., Khurana, J. P., Tyagi, A. K., & Khurana, P. (2008). An update on chloroplast genomes. *Plant Systematics and Evolution, 271*, 101–122.
Sabater, B. (2018). Evolution and function of the chloroplast. Current investigations and perspectives. *International Journal of Molecular Sciences, 19*, 3095.
Saunders, R. M. K. (2000). Monograph of *Schisandra* (Schisandraceae). *Systematic Botany Monographs, 58*, 1–146.
Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research, 33*, W686–W689.
Shi, C., Liu, Y., Huang, H., Xia, E. H., Zhang, H. B., & Gao, L. Z. (2013). Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: An exemplary study of *ycf*15 function and evolution in angiosperms. *PLoS ONE, 8* e59620.
Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., & Sugiura, M. (1994). Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thumbergii*. *Proceedings of the National Academy of Sciences of the United States of America, 91*, 9794–9798.
Yang, Y., Dang, Y. Y., Li, Q., Lu, J. J., Li, X. W., & Wang, Y. T. (2014). Complete chloroplast genome sequence of poisonous and medicinal plant *Datura stramonium*: Organizations and implications for genetic engineering. *PLoS ONE, 9* e110656.
Zhang, J., Chen, M., Dong, X., Lin, R., Fan, J., & Chen, Z. (2015). Evaluation of four commonly used DNA barcoding Loci for Chinese medicinal plants of the family Schisandraceae. *PLoS ONE, 10* e0125574.