

RESEARCH PAPER



Comprehensive analysis of complete chloroplast genome sequence of *Plantago asiatica* L. (Plantaginaceae)

Jing Wu^a, Jing Zhang^a, Xiaohu Guo^a, Nianjun Yu^{a,b,c}, Daiyin Peng^{a,b,c}, and Shihai Xing^{a,b,d}

^aCollege of Pharmacy, Anhui University of Chinese Medicine, Hefei, China; ^bInstitute of Traditional Chinese Medicine Resources Protection and Development, Anhui Academy of Chinese Medicine, Hefei, China; ^cMOE-Anhui Joint Collaborative Innovation Center for Quality Improvement of Anhui Genuine Chinese Medicinal Materials, Hefei, China; ^dAnhui Province Key Laboratory of Research & Development of Chinese Medicine, Hefei, China

ABSTRACT

Plantago asiatica L. is a representative individual species of Plantaginaceae, whose high reputation is owed to its edible and medicinal values. However, the phylogeny and genes of the *P. asiatica* chloroplast have not yet been well described. Here we report the findings of a comprehensive analysis of the *P. asiatica* chloroplast genome. The *P. asiatica* chloroplast genome is 164,992 bp, circular, and has a GC content of 37.98%. The circular genome contains 141 genes, including 8 rRNAs, 38 tRNAs, and 95 protein-coding genes. Seventy-two simple sequence repeats are detected. Comparative chloroplast genome analysis of six related species suggests that a higher similarity exists in the coding region than the non-coding region, and differences in the degree of preservation is smaller between *P. asiatica* and *Plantago depressa* than among others. Our phylogenetic analysis illustrates *P. asiatica* has a relatively close relationship with *P. depressa*, which was also divided into different clades with *Plantago ovata* and *Plantago lagopus* in the genus *Plantago*. This analysis of the *P. asiatica* chloroplast genome contributes to an improved deeply understanding of the evolutionary relationships among Plantaginaceae.

ARTICLE HISTORY

Received 01 November 2022
Revised 22 December 2022
Accepted 23 December 2022

KEYWORDS

Plantago asiatica;
plantaginaceae; chloroplast
genome; comparative
analysis; phylogenetic
analysis

Plantago asiatica Linnaeus 1753. is an annual or biennial food and herb belonging to the genus *Plantago* of Plantaginaceae¹ (Figure 1a-1b). The leaves of *P. asiatica* are well-known as vegetables and tea in China or even East Asia.² *P. asiatica* as well as *Plantago depressa* Willd. (Figure 1c) are traditionally used as Plantaginis herb, while their dry seeds are known as Plantaginis semen.³ *P. asiatica* is easily cultivated, providing rich resources at low prices. It possesses good prospects in diseases prevention and treatment. Also, modern pharmacological studies have shown that the Plantaginis herb and Plantaginis semen have cholesterol degradation, anti-inflammatory, and anti-oxidation activities and can be used to treat liver disease and obesity.⁴⁻⁷ In particular, *P. asiatica* seeds extract rich in phenylethanoid glycosides, flavonoids, iridoids and alkaloids have been reported to prevent heart diseases related to cardiac hypertrophy.⁸ Polysaccharides from *P. asiatica* can be exploited as functional foods attributed to their thermal, antioxidant and radical-scavenging characteristics.^{9,10} Cosmetic and health care applications have been proposed for ethanol extracts from *P. asiatica*.¹¹ Besides, this plant has strong tolerance to formaldehyde in the air and good formaldehyde removal ability, either.¹² Scholars have shed light on compounds identification and clinical functions of *P. asiatica*, which will contribute to maximizing its economic value.¹³ To our knowledge, no study has yet linked *P. asiatica* metabolites to genes in the chloroplast genome,

which can be cornerstones of digging key genes and high-quality germplasm resource.¹⁴

The chloroplast is a key organelle where photosynthesis occurs. Energy in the atmosphere can be converted to starch, pigments, or amino acids here, which are necessary for life activities.¹⁵ Chloroplast has a relatively independent genetic system called the complete chloroplast (Cp) genome. Genome sequencing has incomparable advantages in analyzing the phylogenetic relationship and further genetic diversity or evolution.¹⁶ It suits plant taxonomic and adaptive evolutionary studies, especially those involved in interspecific identification.¹⁷ Functions and evolutionary trajectory of different species in the family Plantaginaceae remained controversial had different active components, leading to different pharmacological effects.^{18,19} DNA barcoding in chloroplast is increasingly popular in phylogeny analyses because of its unmatched accuracy.²⁰⁻²² Moreover, genes detected in the chloroplast can help determine phylogenetic relationships of species belonging to the genus *Plantago*.²³ Different types of chloroplast genes have been reported to have diverse effects on metabolite regulation. The gene *yefS* can modulate tetrapyrrole metabolism and biosynthesis, and overexpression of *accD* can increase fatty acid content in leaves.^{24,25} Based on taxonomic studies, Plantaginaceae was regarded as one of the clades of Scrophulariaceae, while it is now formally called Plantaginaceae, although it has

CONTACT Daiyin Peng; Shihai Xing ✉ pengdy@ahtcm.edu.cn; xshshihai@163.com College of Pharmacy, Anhui University of Chinese Medicine, Hefei 230012, China; Shihai Xing ✉ xshshihai@163.com Institute of Traditional Chinese Medicine Resources Protection and Development, Anhui Province Key Laboratory of Research & Development of Chinese Medicine, Hefei, China

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15592324.2022.2163345>

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Figure 1. (a). The whole plant of *P. asiatica*. (b). The flowers of *P. asiatica*. (c). The original plant of *P. depressa*. (d). The specimen of *P. asiatica*.

drastically expanded from what systematists considered it to be just a few years ago.²⁶ An ambiguous perspective always existed that whether Plantaginaceae was more closely relative to Plumbaginaceae in Caryophyllales or Scrophulariaceae in Lamiales.^{26–29} Regrettably, little available chloroplast data was published on *P. asiatica*, the most representative species of Plantaginaceae. Such data could help compare similar species and describe their phylogenetic relationships.

There is a need for better Cp genome data for *P. asiatica*. We use Illumina technology to sequence the whole chloroplast genome of *P. asiatica*, reporting the overall assembly, annotation, and analysis and making comparisons to other species. It will firstly provide a comprehensive analysis that is favorable for the further development and utilization of the resources.

Results and discussions

Primary characteristics of the plantago asiatica chloroplast genome

Clean data of the Cp genome of *P. asiatica* from Illumina sequencing was obtained with 30,787,584 reads representing 4,618,137,600 bp. After optimizing the raw data, the effective

ratio of Q20 and Q30 were 97.63% and 93.39%, respectively. The sequencing quality was considered good.

The complete Cp genome of *P. asiatica* has a typical circular shape, the same as most angiosperms, with a length of 164,992 bp, which consists of four distinct regions: a large single-copy (LSC) region of 82,983 bp, a small single-copy (SSC) region of 4,715 bp, and a pair of inverted repeats (IR) regions of 38, 647 bp (Table 1). The GC content can also reflect the characteristics of

Table 1. The primary characteristics of the <i>P. asiatica</i> chloroplast genome.		
Sequence Region		<i>P. asiatica</i>
Total chloroplast genome size (bp)		164,992
LSC length (bp)		82,983
SSC length (bp)		4,715
IR length (bp)		38, 647
Total number of genes		141
Protein-coding genes		95
tRNA genes		8
rRNA genes		38
Genes duplicated by IR		26
Genes with introns		18
GC content	Total (%)	37.98
	LSC (%)	36.63
	SSC (%)	30.27
	IR (%)	39.90
	All genes (%)	38.13

chloroplast genome composition. It was detected that the GC content of *P. asiatica* Cp genome was 37.98%, which was far lower than the AT content (62.02%). The characteristics of extreme AT richness always existed in chloroplast genomes,³⁰ consistent with the consequence of this study. The IR region sequence contains eight genes encoding rRNA, so the corresponding GC content (39.90%) is significantly higher than that of the LSC (36.63%) and SSC (30.27%) regions. The skewness of it was reported to be an

instruction of DNA leading and lagging chains, replication starting, and terminal points.³¹

Gene classification

The chloroplast genome encodes 141 genes belonging to three categories, including 95 protein coding genes, 8 ribosomal RNA (rRNA) genes, and 38 transfer RNA (tRNA) genes (Figure 2). When calculating duplicated

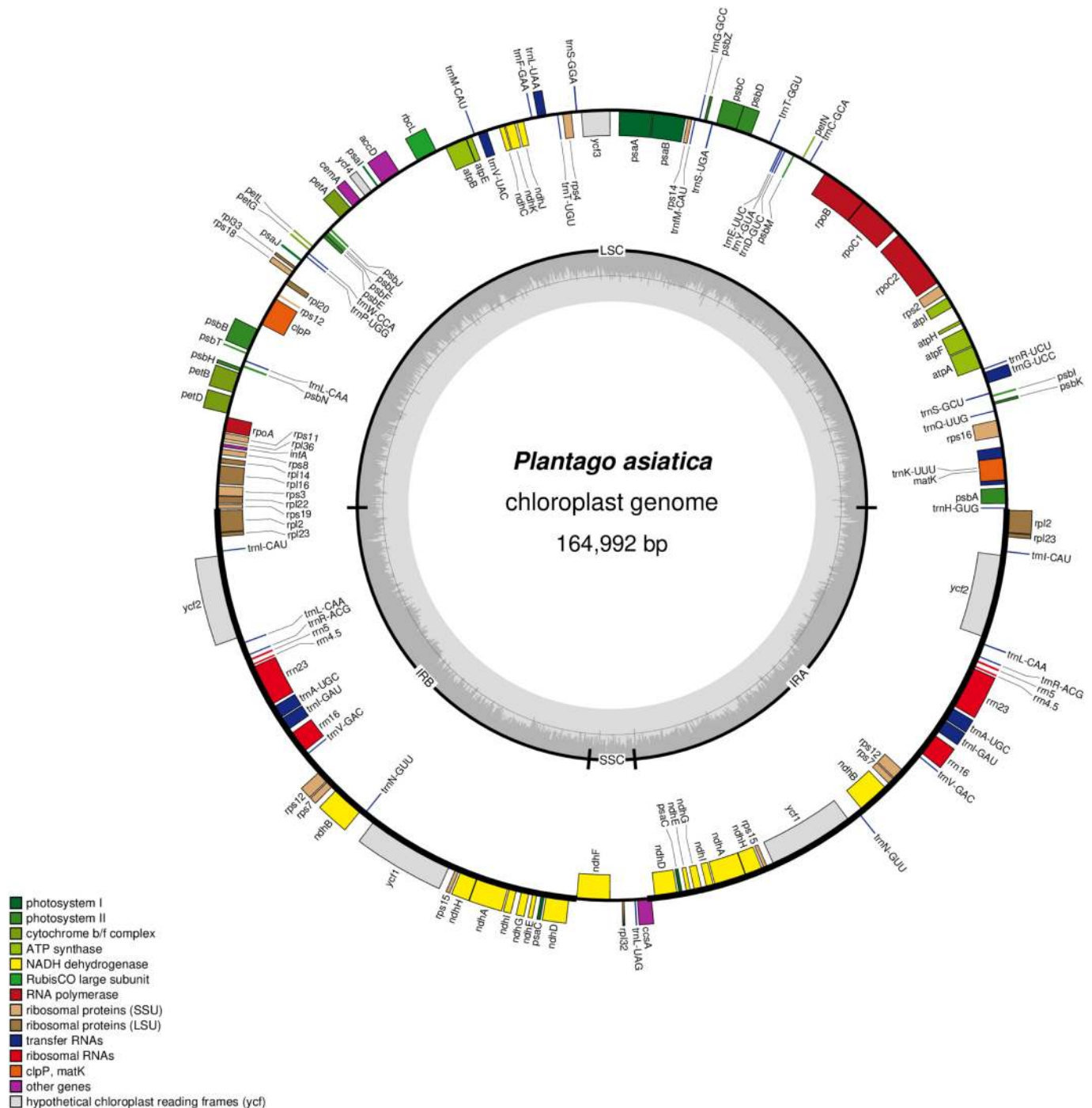


Figure 2. Circular map of the chloroplast genome of *P. asiatica*. Genes drawn inside the circle are transcribed clockwise, while those outside are transcribed counter-clockwise. Genes are color-coded to imply functional groups. The dark gray area in the inner circle corresponds to the GC content while the light gray corresponds to the AT content of the genome. The SCC, LSC, IRA and IRb regions are noted in the inner circle.

Table 2. The length of introns and exons in genes with introns in *P. asiatica*.

Gene	Region	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
<i>atpF</i>	LSC	145	733	416		
<i>clpP1</i>	LSC	69	732	291	640	237
<i>ndhA*</i>	IRA; IRB	551	1074	532		
<i>ndhB*</i>	IRA; IRB	723	677	756		
<i>ycf3</i>	LSC	124	708	228	729	155
<i>petB</i>	LSC	6	730	642		
<i>petD</i>	LSC	8	712	475		
<i>rpl16</i>	LSC	9	849	399		
<i>rpl2*</i>	IRA; IRB	391	677	434		
<i>rpoC1</i>	LSC	451	748	1613		
<i>rps12#</i>	IRA; IRB; LSC	113		228	536	25
<i>rps16</i>	LSC	227	136,163	40		
<i>trnA-UGC*</i>	IRA; IRB	38	818	35		
<i>trnG-UCC</i>	LSC	23	700	48		
<i>trnI-GAU*</i>	IRA; IRB	37	812	35		
<i>trnK-UUU</i>	LSC	37	2444	35		
<i>trnL-UAA</i>	LSC	35	513	50		
<i>trnV-UAC</i>	LSC	36	579	37		

*Identical genes are duplicated in the IR region with introns. #The *rps12* is a trans-spliced gene with the 5' end located in the LSC region and duplicated in the 3' end in the IR regions.

genes in the IR region, among them, 26 genes were both in the IRA and IRB regions, including 15 protein-coding genes (*ccsA*, *ndhA*, *ndhB*, *ndhD*, *ndhE*, *ndhG*, *ndhH*, *ndhI*, *psaC*, *rpl2*, *rpl23*, *rps12*, *rps15*, *rps7*, *ycf2*, *ycf1*), four rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, *rrn5*) and seven tRNA genes (*trnA-UGC*, *trnL-CAA*, *trnI-GAU*, *trnM-CAU*/*trnL-CAU*, *trnN-GUU*, *trnR-ACG*, *trnV-GAC*). The SSC region consisted of three protein-coding genes (*ccsA*, *rpl32*, *ndhF*) and one tRNA gene (*trnL-UAG*). The LSC region comprised 62 protein-coding genes and 22 tRNA genes (Supplementary Table S1).

Most of the protein-coding genes in the chloroplast genome of *P. asiatica* consisted of one exon. A total of 15 genes (*rpoC1**, *rps16**, *rpl2**, *rpl6**, *ndhA**, *ndhB**, *atpF**, *petB**, *petD**, *trnA-UGC**, *trnG-UCC**, *trnI-GAU**, *trnK-UUU**, *trnL-UAA**, *trnV-UAC**) contained one intron and three genes (*clpP1**, *ycf3**, *rps12**) embodied two introns (Table 2). In the anti-clockwise direction, the region between LSC and SSC is defined as IRA, and the region from SSC to LSC is IRb. What made the difference was the expression model of the trans-spliced gene *rps12* observed due to separate transcription. This was verified as a gene in *P. asiatica* consisting of three exons, whose first

Table 3. A list of genes found in the plastid genome of *P. asiatica*.

Category for Genes	Group of Gene	Name of Gene				
RNA genes	ribosomal RNA transfer RNA	<i>rrn5</i>	<i>rrn4.5</i>	<i>rrn16</i>	<i>rrn23</i>	
		<i>trnA-UGC^a</i>	<i>trnC-GCA</i>	<i>trnD-GUC</i>	<i>trnE-UUC</i>	<i>trnF-GAA</i>
		<i>trnfM-CAU</i>	<i>trnG-GCC</i>	<i>trnG-UCC^a</i>	<i>trnH-GUG</i>	<i>trnI-CAU</i>
		<i>trnI-GAU^a</i>	<i>trnK-UUU^a</i>	<i>trnL-CAA</i>	<i>trnL-UAA^a</i>	<i>trnL-UAG</i>
		<i>trnM-CAU</i>	<i>trnN-GUU</i>	<i>trnP-UGG</i>	<i>trnQ-UUG</i>	<i>trnR-ACG</i>
		<i>trnR-UCU</i>	<i>trnS-GCU</i>	<i>trnS-GGA</i>	<i>trnS-UGA</i>	<i>trnT-GGU</i>
		<i>trnT-UGU</i>	<i>trnV-GAC</i>	<i>trnV-UAC^a</i>	<i>trnW-CCA</i>	<i>trnY-GUA</i>
		Transcription and translation related genes	transcription	<i>rpoA</i>	<i>rpoB</i>	<i>rpoC1^a</i>
Small subunit of ribosome	<i>rps2</i>		<i>rps3</i>	<i>rps4</i>	<i>rps7</i>	<i>rps8</i>
	<i>rps11</i>		<i>rps12^a</i>	<i>rps14</i>	<i>rps15</i>	<i>rps16^a</i>
	<i>rps18</i>		<i>rps19</i>			
Large subunit of ribosome	<i>rpl2^a</i>		<i>rpl14</i>	<i>rpl20</i>	<i>rpl22</i>	<i>rpl23</i>
	<i>rpl32</i>		<i>rpl33</i>	<i>rpl36</i>	<i>rpl16^a</i>	
Photosynthesis related genes	Rubisco		<i>rbcL</i>			
	Photosystem I	<i>psaA</i>	<i>psaB</i>	<i>psaC</i>	<i>psal</i>	<i>psaJ</i>
	Photosystem II	<i>psbA</i>	<i>psbB</i>	<i>psbC</i>	<i>psbD</i>	<i>psbE</i>
		<i>psbF</i>	<i>psbH</i>	<i>psbI</i>	<i>psbJ</i>	<i>psbK</i>
		<i>psbL</i>	<i>psbM</i>	<i>psbT</i>	<i>psbZ</i>	<i>psbN</i>
	NADPH dehydrogenase	<i>ndhA^a</i>	<i>ndhB^a</i>	<i>ndhC</i>	<i>ndhD</i>	<i>ndhE</i>
		<i>ndhG</i>	<i>ndhH</i>	<i>ndhI</i>	<i>ndhJ</i>	<i>ndhK</i>
	ATP synthase	<i>atpA</i>	<i>atpB</i>	<i>atpE</i>	<i>atpF^a</i>	<i>atpH</i>
		<i>atpI</i>				
	cytochrome b/f complex	<i>petA</i>	<i>petB^a</i>	<i>petD^a</i>	<i>petG</i>	<i>petL</i>
		<i>petN</i>				
	Assembly and stability of photosystem I	<i>ycf3^a</i>	<i>ycf4</i>			
	Other genes	cytochrome c synthesis	<i>ccsA</i>			
Maturase		<i>matK</i>				
translation initiation factor		<i>infA</i>				
carbon metabolism		<i>cemA</i>				
Subunit of acetyl-CoA-carboxylase		<i>accD</i>				
ATP-dependent Clp protease		<i>clpP1^a</i>				
conserved reading frames		<i>ycf1</i> ,	<i>ycf2</i>			
Genes of unknown function						

^aIndicate the intron-containing genes.

exon was located in the LSC region, far from the other two exons distributed to the IR regions. The *rps12* gene ligation between exon 1 and 2 has been confirmed through complementary DNA sequencing of *rps12* mRNA.³² Almost all protein-coding genes started with the standard initiator codon ATG, with some exceptions. The *rpoC2* and *rps19* began with GTC and GTG, respectively) which has been reported in the chloroplast genome of other plants.³³

Most chloroplast genes may play an essential role in photosynthetic pathways and self-replication. Also, some genes possess special or undefined functions. Based on the functions these genes may perform, genes in the complete Cp genome of *P. asiatica* were sorted into the following types overall: (1) RNA genes; (2) Transcription and translation-related genes; (3) Photosynthesis related genes; (4) Other genes; and (5) Genes of unknown function (Table 3).

The gene function annotation of *P. asiatica*

The functional annotations from the non-redundant (NR) database of all protein-coding genes were demonstrated (Supplementary Table S2). Among them, 68 protein-coding genes were annotated in species of Plantaginaceae. There were 64 homologous genes to *Plantago media*, while four to *Plantago maritima*. The other group of genes were homologous to those in other families such as *Digitalis lanata*, *Cajanus cajan*, and *Mucuna pruriens*.

Functional annotation was done using Gene Ontology (GO) terms, including molecular functions (MF), cellular components (CC), and biological processes (BP) subcategories (Supplementary Table S3). The top five GO terms of these three classes are given in Figure 3a. This class was overrepresented among predicted molecular functions, binding, catalytic activity, transporter activity, structural molecule activity, and translation regulator activities. Most genes were related to metabolic and cellular processes attached to BP (57 and 56,

respectively), followed closely by protein-containing complex and cell part in CC (52 and 42, respectively). Several genes were also involved in localization and biological regulation, among other functions.

By Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation,³⁴ most genes were related to metabolism, especially energy metabolism, nucleotide metabolism, and carbohydrate metabolism (Figure 3b). This was an important aspect of the link between the most dominant term of metabolism in the KEGG pathways and metabolic process in the GO terms. The second group was genetic information processing, such as translation and transcription. Besides, these genes were also characterized and distributed to 15 specific pathways (Supplementary Table S4). Thirty-two genes were predicted to participate in the most recurrent term “Photosynthesis”, mainly in the chloroplast, where various metabolites were produced. Notably, the gene *accD* in the chloroplast of *P. asiatica* was related to fatty acid biosynthesis and metabolism, which had been verified to improve fatty acid content in tobacco, extend leaf longevity, and result in a two-fold increase in seed yield.²⁴ The *clpP1* gene encoded serine protease, which ensures normal physiological metabolic processes by degrading and removing misfolded proteins.³⁵ The *clpP* disruption also induced will lead to deformed leaf development and affect the accumulation of metabolites.³⁶ Hence, further work on exploring the functions of genes in chloroplast about metabolites is of great significance in *P. asiatica*.

SSRs and long repeat analysis

Plant genome SSRs are well-described as microsatellites and short tandem repeats (STRs). They were widely employed for molecular genetic markers and plant typing.³⁷ A total of 72 short repeats were detected, mostly single nucleotide repeats and mainly A (51) and T (56) (Supplementary Table S5). The bulk of SSRs was located in the LSC region ($n = 49$), with 1, 11,

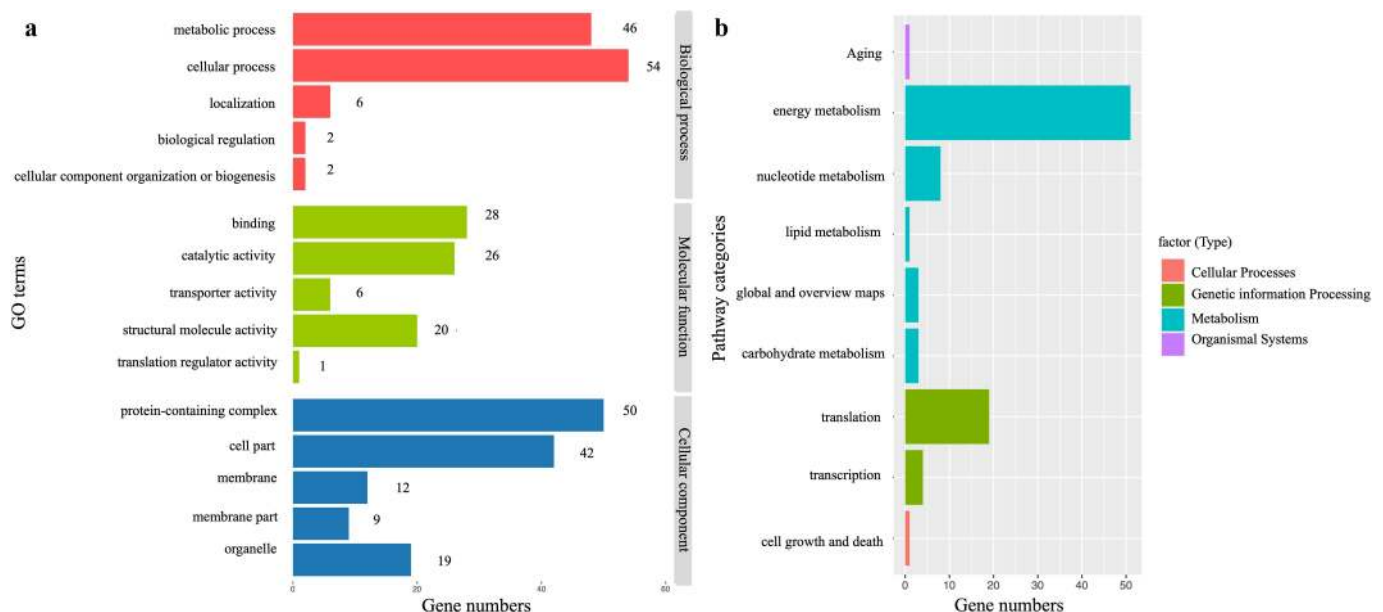


Figure 3. Genes functional annotation in the complete chloroplast genome of *P. asiatica*. (a). The GO classification of genes is based on biological processes, molecular functions, and cellular components. (b). The KEGG pathway categories corresponding to genes in *P. asiatica* chloroplast.

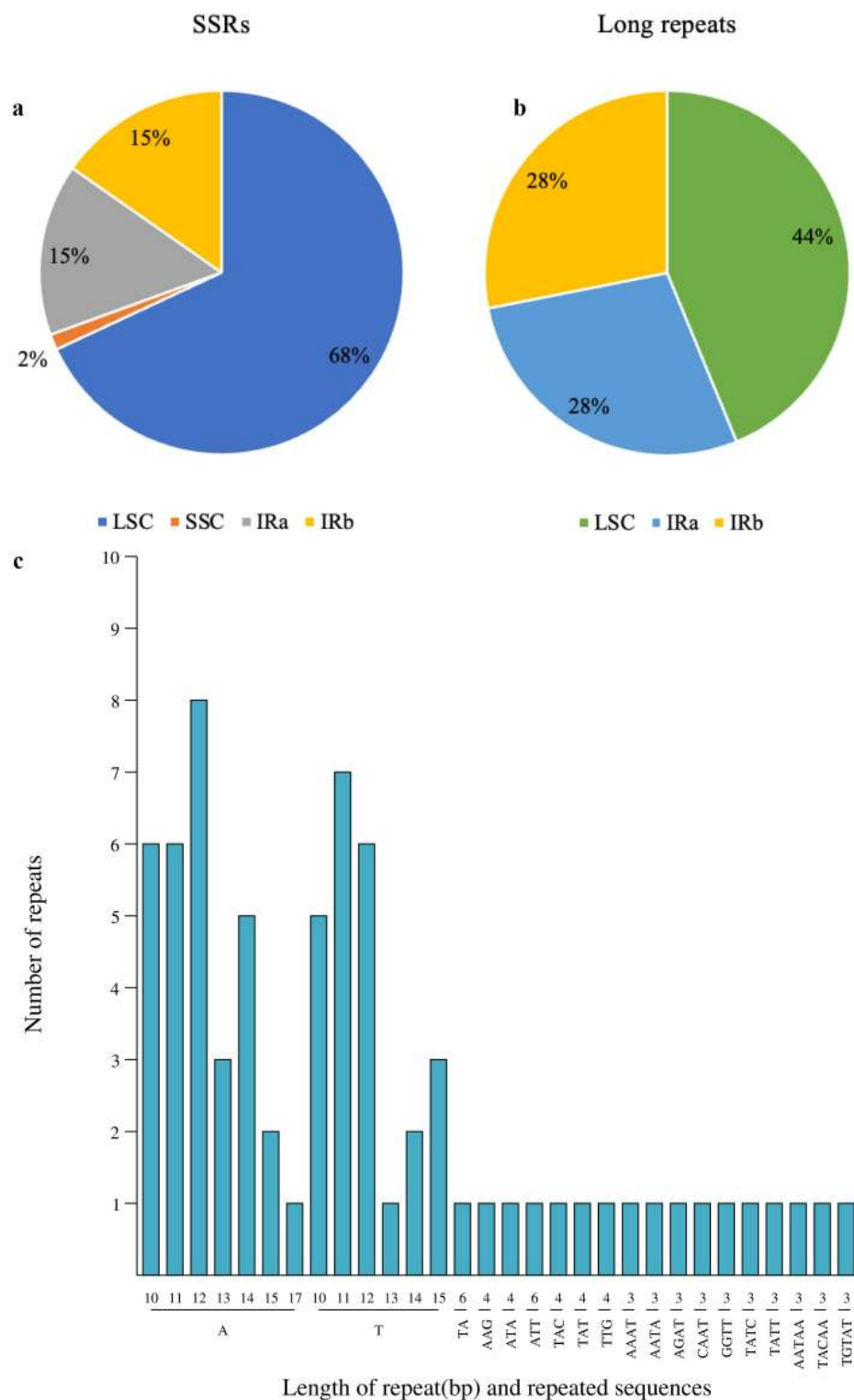


Figure 4. The type, number, and presence of SSRs and long repeats in the Cp genome of *P. asiatica*. (a) Presence of SSRs in the LSC, SSC, IRa, and IRb regions. (b) Presence of long repeats in the LSC, IRa, and IRb regions. (c) Type and number of SSRs in the Cp genome of *P. asiatica*.

and 11 SSRs in the SSC, IRa, and IRb regions, respectively (Figure 4a). Of these, 30 short repeat types were diverse from each other, which were comprised of five SSR types: mononucleotide ($n = 13$), dinucleotide ($n = 1$), trinucleotide ($n = 6$), tetranucleotide ($n = 7$), and pentanucleotide ($n = 3$) (Figure 4c). Among these SSRs, 21 SSRs were only composed of A or T bases. In the other nine simple repeats, more than

half of the sequence consisted of A or T bases but rarely covered tandem G or C bases. Therefore, SSRs in the *P. asiatica* Cp chloroplast genome were AT-rich, following the overall AT abundance of the complete sequence of the chloroplast. This finding had been reported in other chloroplast genomes, with speculation that it may be due to A-T transformation being easier than G-C transformation.³⁸

The SSRs unevenly distributed in the *P. asiatica* chloroplast genome can provide a theoretical basis for developing SSR molecular markers. Moreover, we also analyzed the large repeat sequence in the chloroplast. A total of 49 long tandem repeats were identified in the *P. asiatica* chloroplast genome (Supplementary Table S6), with the largest percentage of 44% being distributed to the LSC region. There was no long repeat located in the SSC region, while 28% repeat sequences in IRA and IRb regions (Figure 4b). Long tandem repeats, unlike SSRs, were dispersed in the whole genome, with 27 forward (F) repeats and 22 palindromic (P) repeats. The long repeats ranged from 40 to 659 bp.

Codon usage bias

Codon usage exerts a significant influence on shaping genome evolution. As one of the numerous factors that can decide codon usage, mutational bias has a particularly crucial role in shaping the plastome evolutionary phenomenon.³⁹ These factors have been demonstrated to influence codon usage at the mutational and translational levels, but mutational pressure is the dominant force at the level of chloroplast genomes. Other mechanisms that potentially help shape the codon use bias include strand asymmetry, which causes strand-specific bias in organelle genomes.⁴⁰ Because of codon degeneracy, each amino acid has one to six codons. Above all, understanding the mutational and translational bias of codon usage can help explore chloroplast evolution. The inequality of synonymous codon usage results from natural selection, mutation, and genetic drift. Relative synonymous codon usage (RSCU) and codon usage measures in *P. asiatica* such as Nc (number of codons) values suggesting the degree of the codon usage bias,

frequency of A, T, G, and C were exhibited (Supplementary Table S7). The Nc value for each protein-coding gene ranged from 23.64 (*psbI*) to 58.78 (*clpP1*), among which Nc values for *rps16*, *rpl36*, and *psbT* could not be calculated because they contained no amino acids with synonymous codons. Consequently, there were 28,949 codons detected in all genes (Figure 5, Supplementary Table S8). The leucine encoded by six types of codons was the most abundant amino acid in this chloroplast genome (3,220 codons, 11.12%). The second abundant was isoleucine (2,362 codons, 8.16%). However, cysteine only accounted for 1.22% of amino acids. The highest value of RSCU was UUA (1.82), and the lowest was AGC (0.315). The RSCU value of 32 codons was greater than 1.00, of which 28 codon bases ended in A or U, and the remaining four ended in G. Similar to *Albizia julibrissin*,⁴¹ the G or C terminal codons were used more frequently than expected (RSCU value < 1). And in *P. asiatica*, the stop codons preferred the use of TAA.

Prediction of RNA editing sites

RNA editing in the plastid can contribute to regulating chloroplast development.⁴² We predicted the 47 editing sites in 16 genes (Table 4), all of which were C to T conversions. These changes in RNA editing sites all led to the transformation of amino acids, including TCA (S) to TTA (L), GCG (A) to GTG (V), CAT (H) to TAT (Y), ACC (T) to ATC (I), TCG (S) to TTG (L), and so on. The most common change was serine (S) to leucine (L), accounting for 19 (40.42%). Amino acid conversion from hydrophilic to hydrophobic can lead to an increase in protein hydrophobicity.⁴³ RNA editing enriches the genetic information of the genome as a single gene translate to different proteins. Some genes in the chloroplast must be

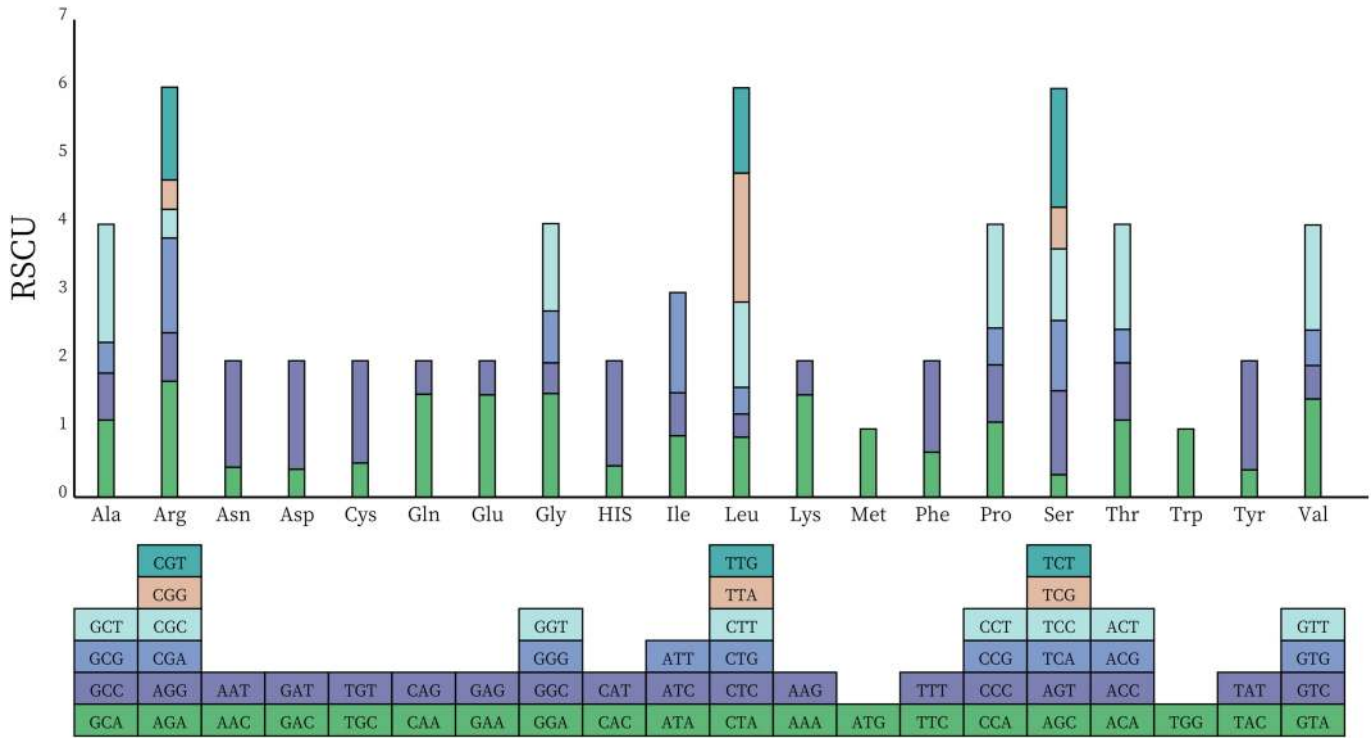


Figure 5. RSCU histogram of *P. asiatica*. The blocks underneath stand for different codon encoding amino acids. The columns on the top depict the sums of RSCU values of the 20 amino acids.

Table 4. The predicted RNA editing sites in *P. asiatica* chloroplast genome.

Gene	Codon Position	Amino Acid Position	Codon (Amino Acid) Conversion	Score
<i>accD</i>	1090	364	CCA (P) = > TCA (S)	1
<i>atpA</i>	791	264	CCC (P) = > CTC (L)	1
<i>ccsA</i>	227	76	ACC (T) = > ATC (I)	0.86
<i>matK</i>	1001	334	GCC (A) = > GTC (V)	0.86
	1154	385	ACC (T) = > ATC (I)	1
	1465	489	CAT (H) = > TAT (Y)	0.86
<i>ndhA</i>	341	114	TCA (S) = > TTA (L)	1
	566	189	TCA (S) = > TTA (L)	1
	1064	355	TCC (S) = > TTC (F)	1
<i>ndhB</i>	95	32	TCA (S) = > TTA (L)	1
	413	138	CCA (P) = > CTA (L)	1
	532	178	CAT (H) = > TAT (Y)	1
	557	186	TCG (S) = > TTG (L)	0.8
	692	231	TCT (S) = > TTT (F)	1
	776	259	TCA (S) = > TTA (L)	1
	782	261	TCA (S) = > TTA (L)	1
	857	286	ACT (T) = > ATT (I)	1
	869	290	ACA (T) = > ATA (I)	1
	1238	413	TCC (S) = > TTC (F)	1
	1427	476	CCA (P) = > CTA (L)	1
<i>ndhD</i>	29	10	ACG (T) = > ATG (M)	1
	94	32	CTT (L) = > TTT (F)	1
	905	302	TCA (S) = > TTA (L)	1
<i>ndhG</i>	320	107	ACA (T) = > ATA (I)	0.8
<i>psbE</i>	214	72	CCT (P) = > TCT (S)	1
<i>rpl2</i>	593	198	GCG (A) = > GTG (V)	0.86
<i>rpl20</i>	59	20	TCT (S) = > TTT (F)	1
	221	74	GCA (A) = > GTA (V)	0.86
	302	101	TCA (S) = > TTA (L)	0.86
	328	110	CCG (P) = > TCG (S)	0.86
<i>rpoA</i>	82	28	CAT (H) = > TAT (Y)	0.86
	329	110	GCC (A) = > GTC (V)	0.86
<i>rpoB</i>	211	71	CAT (H) = > TAT (Y)	1
	473	158	TCA (S) = > TTA (L)	0.86
	551	184	TCA (S) = > TTA (L)	1
	566	189	TCG (S) = > TTG (L)	1
	2426	809	TCA (S) = > TTA (L)	0.86
<i>rpoC1</i>	305	102	GCG (A) = > GTG (V)	1
<i>rpoC2</i>	832	278	CTT (L) = > TTT (F)	0.86
	2293	765	CGG (R) = > TGG (W)	1
	2324	775	GCT (A) = > GTT (V)	1
	2629	877	CCC (P) = > TCC (S)	0.86
	3352	1118	CAC (H) = > TAC (Y)	0.86
	3713	1238	TCG (S) = > TTG (L)	0.86
<i>rps14</i>	80	27	TCA (S) = > TTA (L)	1
	149	50	CCA (P) = > CTA (L)	1
	190	64	CTT (L) = > TTT (F)	0.86

edited to be translated normally. The gene *rpl2* in maize chloroplast can be transcribed and translated normally only after the initiation codon ACG mutating into ATG.⁴⁴ In our study, the *ndhB* gene had the largest number of editing sites, up to 16. The next gene, *rpoC2*, has six editing sites, followed by *rpoB* ($n = 5$), *rpl20* ($n = 4$), *matK* ($n = 3$), *ndhA* ($n = 3$), *ndhD* ($n = 3$), *rps14* ($n = 3$), *rpoA* ($n = 2$). Only one editing site was predicted for genes *accD*, *atpA*, *ccsA*, *ndhG*, *psbE*, *rpl2*, and *rpoC1*, respectively.

Comparative analysis of the chloroplast genomes

To explore the potential sequence divergence of *P. asiatica*, we used annotation of the model plant *Arabidopsis thaliana*, the far relative species *Platycodon grandiflorus* and three species published in the genus *Plantago*. We plotted the total Cp genome identities in mVISTA (Figure 6). Overall, sequence divergence was low across the family Plantaginaceae plastid genomes compared to other species. Among them, *P. depressa*

possessed the highest similarity with *P. asiatica*, while the distantly related species *P. grandiflorus* had marked differences when using the *A. thaliana* chloroplast genome as the external reference sequence. It is worth noting that in chloroplasts of species in the same family, Plantaginaceae and other families, the substitution rates in LSC and SSC regions were mildly higher than in the IR regions.⁴⁵ This phenomenon has been described in many plants, which may be because of copy correction by gene conversion or the presence of conserved rRNA genes in the IR region.⁴⁶

As expected, the non-coding regions showed higher sequence divergence than coding regions.⁴⁷ The coding regions of the *clpP1*, *accD*, *ndhD*, *ccsA*, *ycf1*, and *ycf2* genes were quite diverse among the species in the family Plantaginaceae. The *clpP1* and *ycf1* genes were also different between *P. asiatica* and *P. depressa*. It is suggested that these conserved coding genes in chloroplast can be used to trace the phylogenetic relationships among lots of eudicot plants,⁴⁸ *P. asiatica* included.

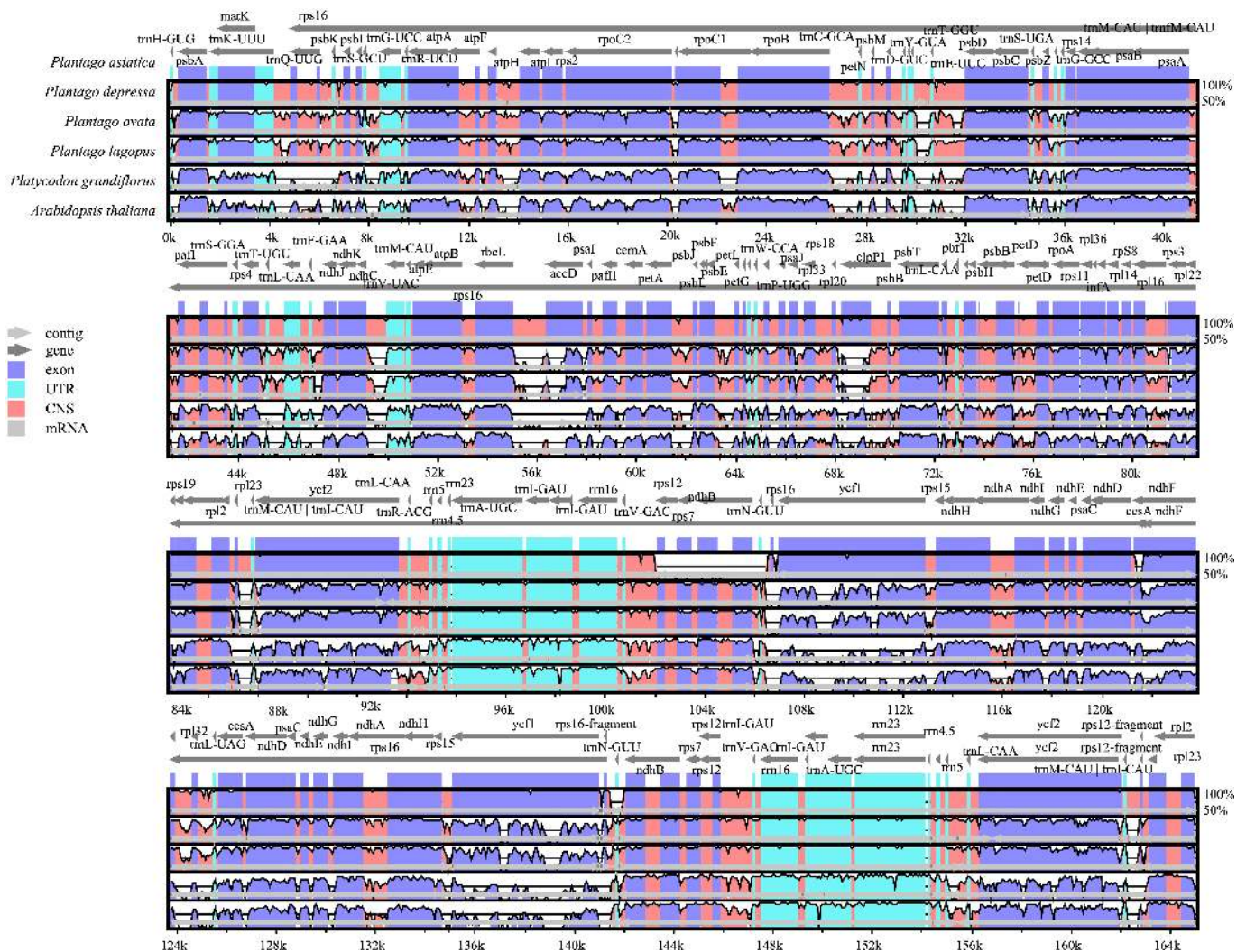


Figure 6. Comparison of the Cp genome sequences of *P. asiatica*, *P. depressa*, *P. ovata*, *P. lagopus*, *P. grandiflorus*, and *A. thaliana* generated using mVISTA. Gray arrows symbolize the position and direction of the genes. Red and blue areas indicate intergenic and genic regions, respectively. Black lines represent regions of sequence identity with *P. asiatica*, with a 50% identity cutoff. Dashed rectangles denote highly divergent regions when *P. asiatica* compared to *P. depressa*, *P. ovata*, *P. lagopus*, *P. grandiflorus*, and *A. thaliana*.

Contraction and expansion in the IR area boundaries are recognized as recurrent evolutionary processes that have led to the observed variance in chloroplast genome size.⁴⁹ The variation of genes situated at plastome termini and boundary shifts (IR-SC) in four junctions (JLB -LSC/IRb, JSB -IRb/SSC, JSA -SSC/IRa, and JLA -IRa/LSC) were investigated. Detailed comparisons of JLB, JSB, JSA, and JLA in *P. asiatica*, *P. lagopus*, *P. depressa*, *P. media*, *P. ovata*, far relative species *Plumbago auriculata* and *A. thaliana* were depicted in Figure 7. Compared with *A. thaliana*, the *ndhF* gene overlapped at the JSB boundary of *P. asiatica*, *P. lagopus*, *P. depressa*, *P. media*, *P. ovata*, *P. auriculata*, while this gene was not in the same position of the *A. thaliana*. The junction line between SSC and IRa intersected the *ycf1* or *ccsA* gene in the family Plantaginaceae, and the gene *rps15* was observed in the JSA of *P. auriculata*. In the family Plantaginaceae, *P. asiatica* showed the largest-scale inversion of expanded IR regions among these 5 species, which may be the main cause of its

largest size of Cp genome. The boundaries in four junctions of *P. asiatica* were similar with *P. depressa* and *P. media* but differed from *P. ovata* and *P. lagopus*. There was some degree of extension into the SSC observed in *P. asiatica*, which to some extent occurred in other angiosperms.⁵⁰ The gene *rpl2* extended into the IRb regions ranging from 50 bp to 73 bp except *P. lagopus*, whose *rpl2* gene was located in LSC region. The same extension can be uncovered in the *ycf1* gene. It traversed the SSC and IRa regions (LR line) of *P. ovata* and *P. lagopus*. Significantly, *P. ovata* and *P. lagopus* had smaller Cp genome size than other species because the *ycf1* gene was partly relocated to the SSC region by IR contraction. Contrastingly, in *P. asiatica* as well as the two similar species, *ycf1* gene did not extend up to the SSC region and was completely duplicated in IRs. The extensions of these genes in *P. asiatica* may be why this was the largest genome among these five species in the family Plantaginaceae, leading to a smaller SSC region as reported in other species.⁵⁰

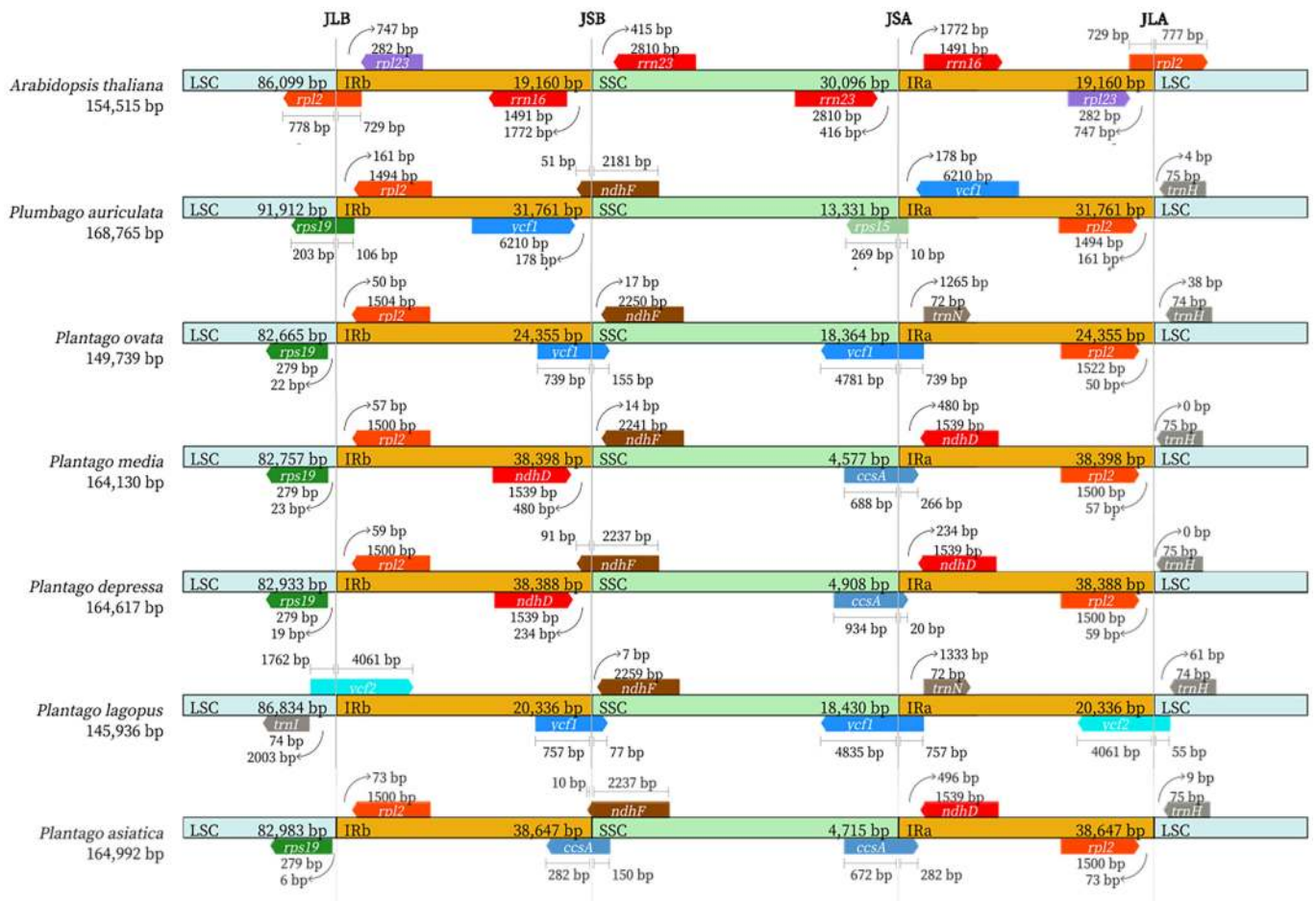


Figure 7. Comparison of the borders of the LSC, SSC, and IR regions of *P. asiatica*, *P. lagopus*, *P. depressa*, *P. media*, *P. ovata*, *P. auriculata* and *Arabidopsis thaliana*.

Phylogenetic relationship analysis

Phylogenetic relationship analysis in the light of chloroplast genome is more convenient than nuclear genome in tracing plant species lineages and species identification.⁴⁸ To resolve the phylogenetic position of *P. asiatica* in the genus *Plantago* and determine whether it is closer to Plumbaginaceae in Caryophyllales or Scrophulariaceae in Lamiales, four typical species in Gesneriaceae, Scrophulariaceae, and Plumbaginaceae were chosen, as well as five species in Plantaginaceae. The topology of the trees constructed by two different methods [maximum likelihood (ML) and Bayesian inference (BI)] was identical, confirming the robustness of our data. The bootstraps (BP) in ML analysis were presented (Figure 8). BI phylogram showing branch lengths was included in Supplementary Figure S1. In the phylogenetic tree, the species can be regarded as three clades. Clade I stood for Plumbaginaceae in Caryophyllales and occupied the basal position. It was phylogenetically distant from other families. Clade II was composed of Scrophulariaceae and Gesneriaceae, supported with 98 bootstrap BP values. Plantaginaceae was divided into Clade III, consisting of two subclades with a BP value of 100. Consequently, with strong support, it uncovered that *P. asiatica* is closely related to *P. depressa* (NC041161). *P. asiatica*, *P. depressa*, and *P. media* (NC028520) were members of one subclade, while *P. ovata* (MH205737) and

P. lagopus (NC041420) were distributed to the other branch in Plantaginaceae. What followed the conclusion was that *P. asiatica*, *P. depressa*, and *P. media* are in the subgenera of *Plantago*, but *P. ovata* and *P. lagopus* are in the *Psyllium* subgenera.⁵¹ Moreover, compared with species in Gesneriaceae and Scrophulariaceae, *P. asiatica* had the farthest relationship with species in Plumbaginaceae.

Furthermore, a DNA polymorphism analysis (Pi) was used to detect low variable sites in the chloroplast genome among 18 species, showing a phylogenetic tree of only highly conserved gene sequences (Figure 9a, Supplementary Table S9). The average Pi was 0.08059. Genes with high Pi values can be used as molecular markers for plant identification and phylogenetic analysis.⁵² In particular, the *clpP1* (Pi = 0.23005) and *matK* (Pi = 0.1536) appeared highly diverse among these species. Although the morphological and chloroplast sequence were similar between *P. asiatica* and *P. depressa*, the *clpP1* and *ycf1* genes were hyper-variable regions between *P. asiatica* and other species. On the basis of the results and the sequence divergence analysis, we proposed that the *clpP1* and *ycf1* genes with comparatively high sequence deviation, are good for interspecies phylogenetic analysis. Moreover, the four genes with the lowest Pi values were chosen to reveal phylogenetic roles: *ndhB* (0.02263), *psbL* (0.02637), *rps12* (0.02901), and *rps7* (0.0294). Using BI and ML methods, phylogenetic analysis was also performed with these four highly

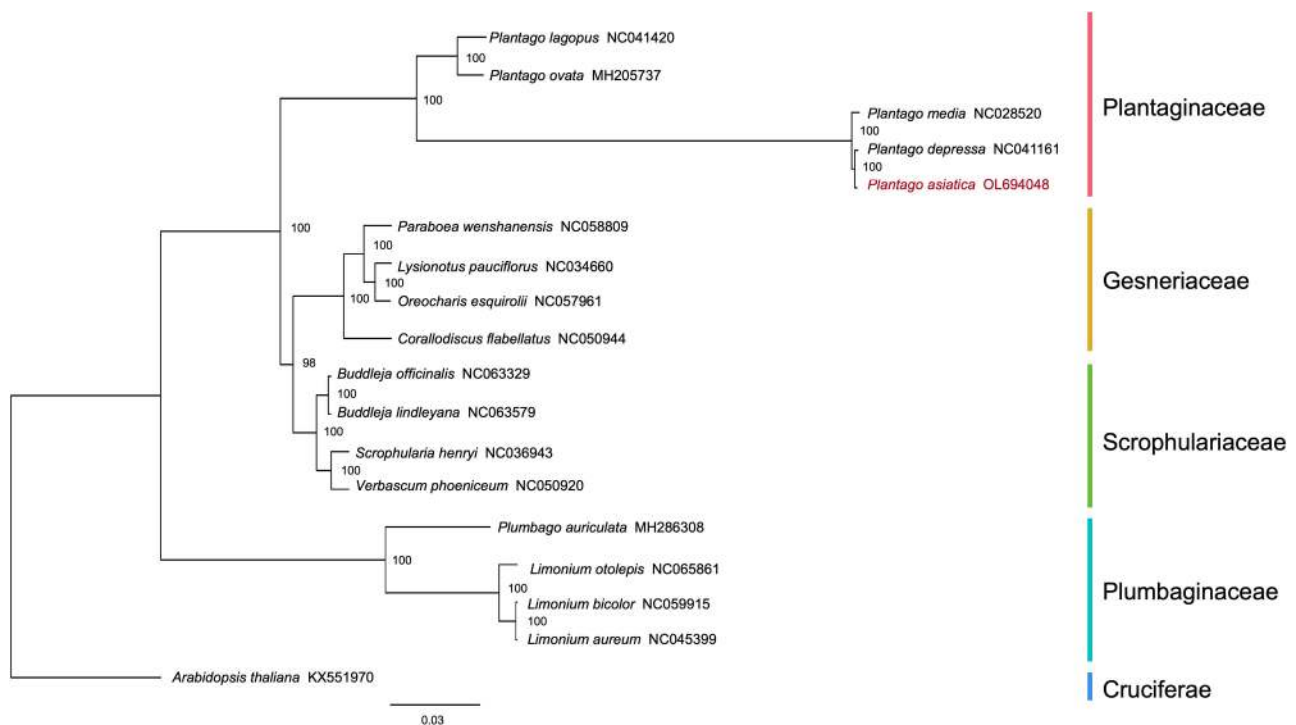


Figure 8. Phylogenetic tree plotting using ML method, based on aligning the completed chloroplast genome sequences of *Plantago asiatica* L. and 17 other representative species. The numbers above the nodes are support values with ML bootstrap values. The distance bar was implied by the ML method.

conserved genes. The ML and BI phylogenetic trees of the four genes are shown in Supplementary Figures S2–S5. The most highly conserved *ndhB* gene was created with the same topology consistent with the complete genomes, which enabled more inference of the relationship based on the phylogenetic studies (Figure 9b).⁵³ The trees generated using BI or ML methods suggested *P. asiatica* and *P. depressa* formed a single clade with high bootstrap (100%) and BI support. In addition, the position of *P. asiatica* and the family Plantaginaceae followed the topologies from previous phylogenetic studies within Lamiales.^{53,54} However, the comprehensive analysis of *P. asiatica* was not operated. And Cp genome from Scrophulariaceae was not included in the previous study. Thus, the present phylogeny verifies that Plantaginaceae is closer to Gesneriaceae and Scrophulariaceae in Lamiales than Plumbaginaceae in Caryophyllales.

Materials and methods

Plant materials

Fresh and well-grown leaves of *P. asiatica* were obtained from a single individual from the medicinal botanical garden of Anhui University of Chinese Medicine Anhui, China (N31° 56'58"; E117°23'38"), followed by the identification of associate Professor Qingshan Yang from the Anhui University of Chinese Medicine. A voucher specimen of this plant was stored in the Center of Herbarium, Anhui University of Chinese Medicine, Hefei, China (AhtcmH, xy.ahtcm.edu.cn/info/1006/6713.htm, Shi-hai Xing, xshshihai@163.com, under the voucher number 20211104) (Figure 1d). No ethical approval/

permission is required in this study. Because the material used in the study is a so normal plant that can be used as vegetable. It's neither an endangered or protected plant nor collected in any protected land. Furthermore, there's no relative local and national regulations or guidelines in China about collecting this plant. The sample was legally collected in accordance with guidelines provided by the authors' institution and national or international regulations. Field studies was complied with local legislation.

DNA extraction, filtering of raw reads and genome sequencing

Leaves were snap-frozen with liquid nitrogen prior to being stored at -80°C . Total mixed genomic DNA was extracted by the CTAB method, with minor changes. DNA quality and integrity were confirmed by gel electrophoresis and Nanodrop methods. DNA libraries with different indices were multiplexed and loaded on an Illumina HiSeq instrument according to the manufacturer's instructions (Illumina, San Diego, CA, USA). Sequencing was carried out using a 2×150 paired-end (PE) configuration, and image analysis and base calling were conducted by the HiSeq control software (HCS), OLB, and GAPipeline-1.6 (Illumina) on the HiSeq instrument.

Chloroplast genome assembly and analysis

The length of the whole chloroplast genome was predicted using KmerGenie.⁵⁵ Short reads were quality controlled and de novo, then assembled using velvet (version 1.2.10), followed by gaps being filled using SSPACE (version 3.0)⁵⁶ and GapFiller (version 1–10).^{57,58} Based on the clean data, the Cp

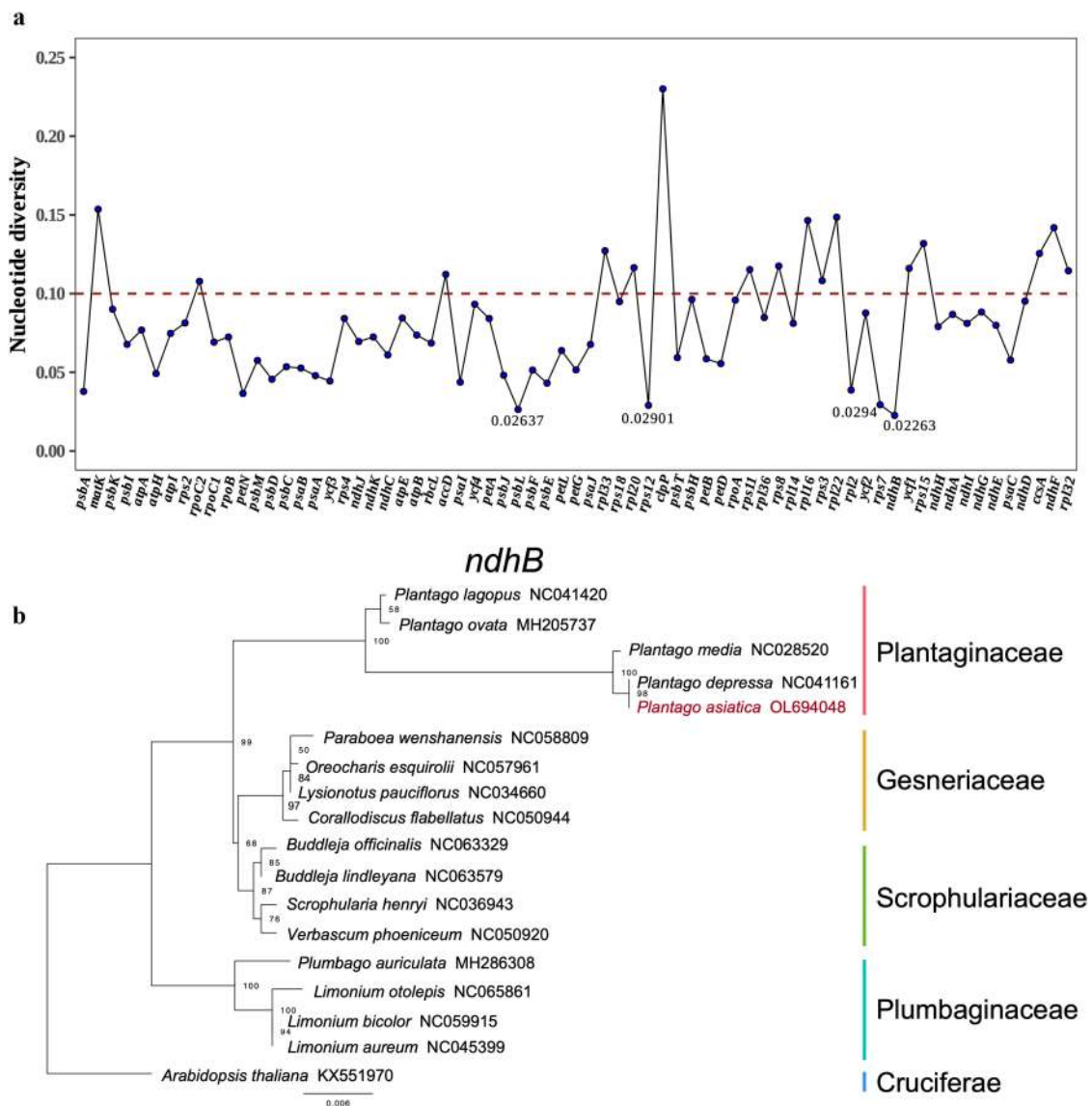


Figure 9. (a). Pi values of the protein coding genes. (b). Phylogenetic tree by maximum likelihood method of *ndhB* gene.

genome of *P. asiatica* was assembled using NOVOPlasty 2.7.2⁵⁹ and auxiliary software Spades⁶⁰ on all the contigs, which utilized the complete Cp sequences of *P. depressa* (GenBank: NC041161) as the reference genome. The assembled Cp genome sequence of *P. asiatica* was submitted to NCBI with GenBank OL694048.

Chloroplast gene annotation

For elucidation of these genes, an online tool of the Dual Organellar GenoMe Annotator (DOGMA) program⁶¹ was initially applied to annotate the *P. asiatica* chloroplast genome, whose protein-coding genes identity was set to 50 and hit number to 10. It is worth noting that start and stop sites of these annotated protein-coding genes needed manual correction. Ribosome RNA was detected by aligning rRNA sequences from other chloroplast genomes to ginseng chloroplast genome sequence using BLAST with global coverage and identified $\geq 90\%$, and tRNA was identified by tRNAscan-SE⁶² with default parameters. A circular Cp genome map was drawn using the

OGDRAW program.⁶³ Then, each divided sequence was searched against the NR database using BLASTN, GO annotation by Blast2GO, and the KEGG database for identification.

Analysis of long repeats and SSRs

The web-based REPuter program⁶⁴ (<https://bibiserv.cebitec.uni-bielefeld.de/reputer>) was employed to visualize the four repetitive sequences of the Cp genome, the forward, reverse, palindrome, and complement sequences included. As for all the repeat types, the constraints set in REPuter contributed to identifying all the 90% identical repeat sequences with a minimum repeat size of 30 bp with a hamming distance equal to 3 (i.e., the gap size between repeats had a maximum length of 3 bp). All overlapping repeats were detached from the final consequences, whose repeating sequence regions were between 1 and 5 bp and repeated no less than three times were considered as SSRs. Besides, the length of repeats ≥ 20 bp were considered as large repeat sequences. The Cp SSRs were detected by MISA⁶⁵ with search parameters of 10 repeat

units for mononucleotide SSRs, five repeat units for dinucleotide SSRs, four repeat units for trinucleotide SSRs, and three repeat units for tetra-, penta-, and hexanucleotide SSRs, respectively.

Prediction of RNA editing sites and codon usage analysis

The Prep-Cp was developed to computationally identify RNA editing sites.⁶⁶ For the analysis, the threshold value was set to 0.8 to ensure the accuracy of the prediction. To claim the consequences of codon usage bias, we applied the program CodonW1.4.2⁶⁷ to analyze the synonymous codon preference of protein-coding genes shorter than 300 bp in length. Each CDS contains initiation codon and termination codon. RSCU is a qualified frequency that synonymous with every codon was specific to encoding an amino acid.

Comparison and analysis among genomes

The complete genomes of six species were compared by mVISTA,⁶⁸ including *A. thaliana*, *P. grandifloras*, and four species in Plantaginaceae (*P. asiatica*, *P. depressa*, *P. lagopus*, and *P. ovata*). IRscope⁶⁹ was also employed to compare the borders of four main regions in the chloroplast genomes of the six species.

Phylogenetic position analysis

To detect divergence among the genus *Plantago* Cp genomes and its associated species, 17 species (including *P. asiatica*) were downloaded from the NCBI database to study the coding region evolution. Entire chloroplast genomes were used to examine these species' average pairwise sequence divergence. Phylogenetic analysis was done after all related sequences were first aligned using MAFFT under default parameters.⁷⁰ Alignments were trimmed using trimAI (v.1.2).⁷¹ The BIC value was used to predict the best fitting model GTRGAMMAX to perform the ML analysis by ModelTest-NG (v.0.1.4).⁷² The final maximum likelihood tree was constructed by RAxML (v. 7.7.8)⁷³ with 1000 rapid bootstrap, which was rooted by an outgroup of *A. thaliana*. In addition to this, BI was employed in MrBayes (v.3.1.2). The Markov Chain Monte Carlo (MCMC) method was run using four incrementally heated chains across 1,000,000 generations, starting from random trees and sampling 1 out of every 100 generations. Pi and sequence polymorphism of the specific 18 species were analyzed using DnaSP (v 5.10.1).⁷⁴

Conclusion

In this study, we generated the chloroplast genome of *P. asiatica* with a classical tetrameric structure of 164,992 bp in length, including 8 rRNA, 38 tRNA, and 95 protein-coding genes. Among all genes, 28,949 codons had G or C terminals. There were 47 editing sites in 16 genes, all of which were C to T conversions. Serine (S) transformation into leucine (L) can increase protein hydrophobicity. Significantly, we detected 72 SSRs in *P. asiatica* that can be utilized in DNA barcoding to

distinguish similar species. The *P. asiatica* chloroplast genome shared similar overall organization and gene contents with most chloroplast genomes, including the closest species. The phylogeny analysis showed that *P. asiatica* formed a sister to *P. depressa*. Besides, Plantaginaceae was closer to Gesneriaceae and Scrophulariaceae in Lamiales than Plumbaginaceae in Caryophyllales. Our genome comparison analysis revealed highly conserved gene sequences and several differences among Plantaginaceae species, which can assist us in further studying these evolutionary relationships. Species in the family Plantaginaceae are usually difficult to distinguish due to similarities in phenotype, while different metabolites in different plants have various effects. So, these conserved gene sequences in the chloroplast genome can be used to identify species of Plantaginaceae. Our findings not only provide new insights into chloroplast genomic and phylogenetic relationships of *P. asiatica* but also will assist with maximizing further development and application of this plant.

Acknowledgments

We would like to thank Genewiz Biotechnology (Suzhou) Co. Ltd in China for chloroplast genome sequencing and bioinformatics analysis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Author Contributions

Conceived, designed the study: SX, DP and JW; collected specimens and prepared samples for sequencing: JW and NY; Analysis and interpretation of the data: SX, JW, JZ, and XG; Drafted the manuscript: JW and NY; Revised and criticized the manuscript: SX and DP; All authors approved the final version and agreed to accountable for all aspects of the work.

Ethics declarations

No ethical approval/permission is required in this study.

Data availability

The data that support the findings of this study are openly available in GenBank of NCBI (<https://www.ncbi.nlm.nih.gov>) under the access number OL694048. The associated BioProject, SRA and Bio-Sample numbers are PRJNA797537, SRR17629686, and S SAMN25008951 respectively.

References

1. Wu P. Shen Nong Ben Cao Jing. Shanghai: Scientific and Technological Literature Publishing House; 1996.
2. Liu X, Wu X, Huang H, Zhong S, Lai X, Cao L. Herbalogical study on *Plantago asiatica* L. *Zhong Yao Cai*. 2002;25:46–48.
3. Committee CP. Pharmacopoeia of the people's republic of China. Beijing: China Chemical Industry Press; 2020.
4. Chung MJ, Woo Park K, Heon Kim K, Kim CT, Pill Baek J, Bang KH, Choi YM, Lee SJ. Asian plantain (*Plantago asiatica*) essential oils suppress 3-hydroxy-3-methyl-glutaryl-co-enzyme A reductase expression in vitro and in vivo and show hypocholesterolaemic properties in mice. *Br J Nutr*. 2008;99(1):67–75. doi:10.1017/S0007114507798926.

5. Oto G, Ekin S, Ozdemir H, Demir H, Yasar S, Levent A, Berber I, Kaki B. Plantago major protective effects on antioxidant status after administration of 7,12-Dimethylbenz(a)anthracene in rats. *Asian Pac J Cancer Prev*. 2011;12:531–535.
6. Torigoe Y. Studies on the constituent of plantago asiatica LINNÉ. (1). Yakugaku Zasshi. 1965;85(2):176–178. doi:10.1248/yakushi1947.85.2_176.
7. Türel I, Ozbek H, Erten R, Oner AC, Cengiz N, Yilmaz O. Hepatoprotective and anti-inflammatory activities of Plantago major L. *Indian J Pharmacol*. 2009;41(3):120–124. doi:10.4103/0253-7613.55211.
8. Fan W, Zhang B, Wu C, Wu H, Wu J, Wu S, Zhang J, Yang X, Yang L, Hu Z. Plantago asiatica L. seeds extract protects against cardiomyocyte injury in isoproterenol-induced cardiac hypertrophy by inhibiting excessive autophagy and apoptosis in mice. *Phytomedicine*. 2021;91:153681. doi:10.1016/j.phymed.2021.153681.
9. Niu Y, Li N, Alaxi S, Huang G, Chen L, Feng Z. A new heteropolysaccharide from the seed husks of Plantago asiatica L. with its thermal and antioxidant properties. *Food Funct*. 2017;8(12):4611–4618. doi:10.1039/C7FO01171G.
10. Gong L, Zhang H, Niu Y, Chen L, Liu J, Alaxi S, Shang P, Yu W, Yu L. A novel alkali extractable polysaccharide from Plantago asiatica L. Seeds and its radical-scavenging and bile acid-binding activities. *J Agric Food Chem*. 2015;63(2):569–577. doi:10.1021/jf505909k.
11. Yoon MY, Kim HJ, Lee SJ, Han J. The effect of antioxidant and whitening action on Plantago asiatica L. leaf ethanol extract for health care. *Technol Health Care*. 2019;27(5):567–577. doi:10.3233/THC-191744.
12. Zhao S, Su Y, Liang H. Efficiency and mechanism of formaldehyde removal from air by two wild plants: plantago asiatica L. and Taraxacum mongolicum Hand.-Mazz. *J Environ Health Sci Eng*. 2019;17(1):141–150. doi:10.1007/s40201-018-00335-w.
13. Li F, Du P, Yang W, Huang D, Nie S, Xie M. Polysaccharide from the seeds of Plantago asiatica L. alleviates nonylphenol induced intestinal barrier injury by regulating tight junctions in human Caco-2 cell line. *Int J Biol Macromol*. 2020;164:2134–2140. doi:10.1016/j.ijbiomac.2020.07.259.
14. Song Y, Chen Y, Lv J, Xu J, Zhu S, Li M, Chen N. Development of chloroplast genomic resources for Oryza species discrimination. *Front Plant Sci*. 2017;8:1854. doi:10.3389/fpls.2017.01854.
15. Greiner S, Golczyk H, Malinova I, Pellizzer T, Bock R, Börner T, Herrmann RG. Chloroplast nucleoids are highly dynamic in ploidy, number, and structure during angiosperm leaf development. *Plant J*. 2020;102(4):730–746. doi:10.1111/tpj.14658.
16. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J*. 1986;5(9):2043–2049. doi:10.1002/j.1460-2075.1986.tb04464.x.
17. Henriquez CL, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, McKain MR. Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). *Genomics*. 2020;112(3):2349–2360. doi:10.1016/j.ygeno.2020.01.006.
18. Lan JP, Tong RC, Sun XM, Zhang HY, Sun S, Xiong AZ, Wang ZT, Yang L. Comparison of main chemical composition of Plantago asiatica L. and P. depressa Willd. seed extracts and their anti-obesity effects in high-fat diet-induced obese mice. *Phytomedicine*. 2021;81:153362. doi:10.1016/j.phymed.2020.153362.
19. Zhong CL, Cao SQ, Sun DM, Cheng XR, Pan LY, Ye RL, Zhou XY, Li GW. Study on the different components of Plantaginis herba from different origins based on UPLC feature atlas. *J Guangdong Pharm Univer (In Chinese)*. 2021;37:40–46.
20. Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y, Yamamoto T, Katayama H. Complete sequence of the chloroplast genome from pear (Pyrus pyrifolia): genome structure and comparative analysis. *Tree Genet Genomes*. 2012;8(4):841–854. doi:10.1007/s11295-012-0469-8.
21. Xue S, Shi T, Luo W, Ni X, Iqbal S, Ni Z, Huang X, Yao D, Shen Z, Gao Z. Comparative analysis of the complete chloroplast genome among Prunus mume. P. Armeniaca, and P. Salicina. *Hortic Res-England*. 2019;6:89.
22. Batnini MA, Bourguiba H, Trifi-Farah N, Krichen L. Molecular diversity and phylogeny of Tunisian Prunus armeniaca L. by evaluating three candidate barcodes of the chloroplast genome. *Sci Horti*. 2019;245:99–106. doi:10.1016/j.scienta.2018.09.071.
23. Dhar MK, Friebe B, Kaul S, Gill BS. Characterization and physical mapping of ribosomal RNA gene families in Plantago. *Ann Bot*. 2006;97(4):541–548. doi:10.1093/aob/mcl017.
24. Madoka Y, Tomizawa KI, Mizoi J, Nishida I, Nagano Y, Sasaki Y. Chloroplast transformation with modified accD operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. *Plant Cell Physiol*. 2002;43(12):1518–1525. doi:10.1093/pcp/pcf172.
25. Peter E, Wallner T, Wilde A, Grimm B. Comparative functional analysis of two hypothetical chloroplast open reading frames (ycf) involved in chlorophyll biosynthesis from Synechocystis sp. PCC6803 and plants. *J Plant Phys*. 2011;168(12):1380–1386. doi:10.1016/j.jplph.2011.01.014.
26. Albach DC, Meudt HM, Oxelman B. Piecing together the “new” Plantaginaceae. *Am J Bot*. 2005;92(2):297–315. doi:10.3732/ajb.92.2.297.
27. Soltis DE, Soltis P, Endress PK, Chase MW. Phylogeny and evolution of angiosperms. Sunderland: Sinauer Associates; 2005.
28. APG II. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot J Linn Soc*. 2003;141(4):399–436. doi:10.1046/j.1095-8339.2003.t01-1-00158.x.
29. APG III. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc*. 2009;161(2):105–121. doi:10.1111/j.1095-8339.2009.00996.x.
30. Roy SW, Irimia M. Origins of human malaria: rare genomic changes and full mitochondrial genomes confirm the relationship of Plasmodium falciparum to other mammalian parasites but complicate the origins of Plasmodium vivax. *Mol Biol Evol*. 2008;25(6):1192–1198. doi:10.1093/molbev/msn069.
31. Saina JK, Gichira AW, Li ZZ, Hu GW, Wang QF, Liao K. The complete chloroplast genome sequence of Dodonaea viscosa: comparative and phylogenetic analyses. *Genetica*. 2018;146(1):101–113. doi:10.1007/s10709-017-0003-x.
32. Sharp PA. On the origin of RNA splicing and introns. *Cell*. 1985;42(2):397–400. doi:10.1016/0092-8674(85)90092-3.
33. Fonseca LHM, Lohmann LG. Plastome rearrangements in the “Adenocalymma-Neojobertia” Clade (Bignoniaceae, Bignoniaceae) and its phylogenetic implications. *Front Plant Sci*. 2017;8:1875. doi:10.3389/fpls.2017.01875.
34. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30. doi:10.1093/nar/28.1.27.
35. Gottesman S. PROTEASES AND THEIR TARGETS IN Escherichia coli. *Annu. Rev. Genet*. 1996;30(1):465–506. doi:10.1146/annurev.genet.30.1.465.
36. Shikanai T, Shimizu K, Ueda K, Nishimura Y, Kuroiwa T, Hashimoto T. The chloroplast clpP gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco. *Plant Cell Physiol*. 2001;42(3):264–273. doi:10.1093/pcp/pce031.
37. Doorduyn L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. The complete chloroplast genome of 17 individuals of pest species Jacobaea vulgaris: sNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Res*. 2011;18(2):93–105. doi:10.1093/dnares/dsr002.
38. Xie DF, Yu Y, Deng YHQ, Li J, Liu HY, Zhou SD, He XJ. Comparative analysis of the chloroplast genomes of the Chinese endemic genus urophyssa and their contribution to chloroplast

- phylogeny and adaptive evolution. *Int J Mol Sci.* 2018;19(7):1847. doi:10.3390/ijms19071847.
39. Liu Q, Xue Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J Genet.* 2005;84(1):55–62. doi:10.1007/BF02715890.
 40. Jia W, Higgs PG. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol Biol Evol.* 2008;25(2):339–351. doi:10.1093/molbev/msm259.
 41. Zhang J, Huang H, Qu C, Meng X, Meng F, Yao X, Wu J, Guo X, Han B, Xing S, *et al.* Comprehensive analysis of chloroplast genome of *Albizia julibrissin* Durazz. (Leguminosae sp.). *Planta.* 2021;255(1):26. doi:10.1007/s00425-021-03812-z.
 42. Wang Y, Wang Y, Ren Y, Duan E, Zhu X, Hao Y, Zhu J, Chen R, Lei J, Teng X, *et al.* white panicle 2 encoding thioredoxin z, regulates plastid RNA editing by interacting with multiple organellar RNA editing factors in rice. *New Phytol.* 2021;229(5):2693–2706. doi:10.1111/nph.17047.
 43. Drescher A, Hupfer H, Nickel C, Albertazzi F, Hohmann U, Herrmann R, Maier R. C-to-U conversion in the intergenic *ndhI/ndhG* RNA of plastids from monocot plants: conventional editing in an unconventional small reading frame. *Mol Genet Genomics.* 2002;267(2):262–269. doi:10.1007/s00438-002-0662-9.
 44. Hoch B, Maier RM, Appel K, Igloi GL, Kossel H. Editing of a chloroplast mRNA by creation of an initiation codon. *Nature.* 1991;353(6340):178–180. doi:10.1038/353178a0.
 45. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 2016;209(4):1747–1756. doi:10.1111/nph.13743.
 46. Khakhlova O, Bock R. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J.* 2006;46(1):85–94. doi:10.1111/j.1365-3113X.2006.02673.x.
 47. Yao X, Tang P, Li Z, Li D, Liu Y, Huang H. The first complete chloroplast genome sequences in actinidiaceae: genome structure and comparative analysis. *PLoS One.* 2015;10(6):e0129347. doi:10.1371/journal.pone.0129347.
 48. Provan J, Powell W, Hollingsworth PM. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol Evol.* 2001;16(3):142–147. doi:10.1016/S0169-5347(00)02097-8.
 49. Goulding SE, Olmstead RG, Morden CW, Wolfe KH. Ebb and flow of the chloroplast inverted repeat. *Mol Gen Genet.* 1996;252(1–2):195–206. doi:10.1007/BF02173220.
 50. Kaila T, Chaduvla PK, Saxena S, Bahadur K, Gahukar SJ, Chaudhury A, Sharma TR, Singh NK, Gaikwad K. *Chloroplast Genome Sequence of Pigeonpea (Cajanus cajan (L.) Millspaugh) and Cajanus scarabaeoides (L.) Thouars: genome organization and comparison with other legumes.* *Front Plant Sci.* 2016;7:1847. doi:10.3389/fpls.2016.01847.
 51. Mower JP, Guo W, Partha R, Fan W, Levens N, Wolff K, Nugent JM, Pabón-Mora N, González F. *Plastomes from tribe Plantagineae (Plantaginaceae) reveal infrageneric structural synapomorphies and localized hypermutation for Plantago and functional loss of ndh genes from Littorella.* *Mol Phylogenet Evol.* 2021;162:107217. doi:10.1016/j.ympev.2021.107217.
 52. Biju VC, P. S, Vijayan S, Rajan VS, Sasi A, Janardhanan A, Nair AS. *The complete chloroplast genome of trichopus zeylanicus, and phylogenetic analysis with dioscoreales.* *Plant Genome.* 2019;12(3):1–11. doi:10.3835/plantgenome2019.04.0032.
 53. Asaf S, Khan AL, Khan A, Khan G, Lee I-J, Al-Harrasi A. Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of *Plantago ovata*. *Sci Rep.* 2020;10(1):3881. doi:10.1038/s41598-020-60803-y.
 54. Si H, Li R, Zhang Q, Liu L. Complete chloroplast genome of *Plantago asiatica* and its phylogenetic position in Plantaginaceae. *Mitoch DNA B Res.* 2022;7:819–821.
 55. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics.* 2014;30(1):31–37. doi:10.1093/bioinformatics/btt310.
 56. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics.* 2011;27(4):578–579. doi:10.1093/bioinformatics/btq683.
 57. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol.* 2012;13(6):R56. doi:10.1186/gb-2012-13-6-r56.
 58. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–829. doi:10.1101/gr.074492.107.
 59. Dierckx N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 2017;45(4):e18. doi:10.1093/nar/gkw955.
 60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455–477. doi:10.1089/cmb.2012.0021.
 61. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics.* 2004;20(17):3252–3255. doi:10.1093/bioinformatics/bth352.
 62. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 2005;33(Web Server):W686–689. doi:10.1093/nar/gki366.
 63. Lohse M, Drechsel O, Kahlau S, Bock R. Organellar Genome DRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 2013;41(W1):W575–581. doi:10.1093/nar/gkt289.
 64. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 2001;29(22):4633–4642. doi:10.1093/nar/29.22.4633.
 65. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet.* 2003;106(3):411–422. doi:10.1007/s00122-002-1031-0.
 66. Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 2009;37(Web Server):W253–259. doi:10.1093/nar/gkp337.
 67. Du X, Zeng T, Feng Q, Hu L, Luo X, Weng Q, He J, Zhu B. *The complete chloroplast genome sequence of yellow mustard (Sinapis alba L.) and its phylogenetic relationship to other Brassicaceae species.* *Gene.* 2020;731:144340. doi:10.1016/j.gene.2020.144340.
 68. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. Vista: computational tools for comparative genomics. *Nucleic Acids Res.* 2004;32(Web Server):W273–279. doi:10.1093/nar/gkh458.
 69. Amirouyefi A, Hyvonen J, Pocza P, Hancock J. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics.* 2018;34(17):3030–3031. doi:10.1093/bioinformatics/bty220.
 70. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772–780. doi:10.1093/molbev/mst010.
 71. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972–1973. doi:10.1093/bioinformatics/btp348.
 72. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B and Flouri T. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol Biol Evol.* 2020;37(1):291–294. doi:10.1093/molbev/msz189.
 73. Stamatakis A. RAXML-VI-HP: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006;22(21):2688–2690. doi:10.1093/bioinformatics/btl446.
 74. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009;25(11):1451–1452. doi:10.1093/bioinformatics/btp187.