The Complete Chloroplast Genome of Onobrychis Gaubae (Fabaceae- Papilionoideae): Comparative Analysis with Related IR-lacking Clade Species

Mahtab Moghaddam
Tarbiat Modares University
Atsushi Ohta
Kyoto University
Motoki Shimizu
Iwate Biotechnology Research Center
Ryohei Terauchi
Kyoto University
Shahrokh Kazempour-Osaloo (skosaloo@modares.ac.ir)
Tarbiat Modares University

Research Article

Keywords: Plastome, IRLC, Onobrychis, Hypervariable region, Phylogenetic relationship

Posted Date: August 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-290026/v2

Version of Record: A version of this preprint was published at BMC Plant Biology on February 19th, 2022.
See the published version at https://doi.org/10.1186/s12870-022-03465-4.

Abstract

Background: Plastome (Plastid genome) sequences provide valuable markers for surveying the evolutionary relationships and population genetics of plant species. Papilionoideae (papilionoids) has different nucleotide and structural variations in plastomes, which makes it an ideal model for genome evolution studies. Therefore, by sequencing the complete chloroplast genome of *Onobrychis gaubae* in this study, the characteristics and evolutionary patterns of plastome variations in IR-loss clade were compared.

Results: In the present study, the complete plastid genome of *O. gaubae*, endemic to Iran, was sequenced using Illumina paired-end sequencing and was compared with previously known genomes of the IRLC species of legumes. The *O. gaubae* plastid genome was 122,873 bp in length and included a large single-copy (LSC) region of 81,486 bp, a small single-copy (SSC) region of 13,805 bp and one copy of the inverted repeat (IR_b) of 29,100 bp. The genome encoded 110 genes, including 76 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosome RNA (rRNA) genes and possessed 83 simple sequence repeats (SSRs) and 50 repeated structures with the highest proportion in the LSC. Comparative analysis of the chloroplast genomes across IRLC revealed three hotspot genes (*ycf*1, *ycf*2, *clp*P) which could be used as DNA barcode regions. Moreover, seven hypervariable regions (*trn*L(UAA)-*trn*T(UGU), *trn*T(GGU)-*trn*E(UUC), *ycf*1, *ycf*2, *ycf*4, *acc*D and *clp*P) were identified between *Onobrychis* species, which could be used to distinguish the *Onobrychis* species. Phylogenetic analyses revealed that *O. gaubae* is closely related to *Hedysarum*. The complete *O. gaubae* genome is a valuable resource for investigating evolution of *Onobrychis* species and can be used to identify related species.

Conclusions: Our results reveal that the plastomes of the IRLC are dynamic molecules and show multiple gene losses and inversions. The identified hypervariable regions could be used as molecular markers for resolving phylogenetic relationships and species identification and also provide new insights into plastome evolution across IRLC.

Background

Chloroplast is a vital organelle in plant cells that has an important role in plant carbon fixation and numerous metabolic pathways[1,2]. In angiosperms, the chloroplast genome (plastome) typically has a circular structure that ranges from 120 to 180 kb in length. Plastomes mostly exhibit a quadripartite structure in which a pair of inverted repeats (IRa and IRb; usually around 25 kb, but can vary from 7 to 88 kb each) separate the large single-copy (LSC, ca. 80 kb) and the small single-copy (SSC, ca. 20 kb) regions[1,2]. Most plastomes encodes 80 protein-coding genes primarily involved in photosynthesis and other biochemical processes along with 30 tRNA and 4 rRNA genes[3,4]. In contrast to mitochondrial and nuclear genomes, the plastomes across seed plants are highly conserved with respect to gene content, structure and organization[5,6]. However, mutations including duplication, rearrangements, and losses have been reported at the genome and gene levels among some angiosperm lineages, including Asteraceae[7], Campanulaceae[8], Onagraceae[9], Fabaceae[10] and Geraniaceae[11].

Fabaceae (legumes) is the third largest family of angiosperms which shows the most extensive structural variation[12]. Currently accepted classification of the legumes based on plastid gene *mat*K includes six subfamilies: Caesalpinioideae, Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, and Papilionoideae[13]. Gene content and gene order in plastomes of subfamilies are highly conserved and similar to the ancestral angiosperm genome organization except for Papilionoideae, which exhibits numerous rearrangements and gene/intron losses and have a smaller genome[5]. In this subfamily, a loss of one of the IR[14], the presence of many repetitive sequences[15] and the presence of a localized hypermutable region[15,16] have been documented. The Papilionoideae is further divided into six major clades: the Genistoids, Dalbergioids, Mirbelioids, Millettioids, Robinioids and the inverted-repeat lacking clade (IRLC)[14]. IRLC is the largest legume lineage which contains over 4000 species in 52 genera and eight tribes[14,17–19]. Recently, with the advent of next generation sequencing (NGS) technology, plastomes of several taxa from different tribes in this clade have been sequenced. The majority of IRLC plastomes sequenced to date were restricted to the tribes Fabeae, Trifolieae and Caraganeae. Thus it is essential to investigate the members from other lineages to better understand plastome evolution within the IRLC, and more broadly within Papilionoideae. In the tribe Hedysareae[20] with nine genera, the plastomes of some *Hedysarum* species and only one species of *Onobrychis* (*O. viciifolia* within subgenus *Onobrychis*) have been reported. In the present study, the complete plastome of *O. gaubae* Bornm. belonging to Hedysareae was sequenced (GenBank accession number: ?). *Onobrychis*, which is the second largest genus after *Hedysarum*, is composed of two subgenera (*Onobrychis* and *Sisyrosema*) and has more than 130 species. This genus mainly found throughout temperate and subtropical regions of Eurasia, N and NE Africa[21]. *Onobrychis gaubae* belongs to the subgenus *Sisyrosema*, is a polymorphic species restricted to the southern slopes of Alborz mountain range, Iran[22,23].

The main goal of this study is to assemble the chloroplast genome of *O. gaubae*, and to annotate the genome and characterize its structure to provide new genomic resource of this species. We also performed comparative analyses of the genome and phylogenetic reconstruction to evaluate the sequence divergence in the plastomes across the IR-lacking clade.

Results

Characteristics of the chloroplast genome of *O. gaubae*

The number of paired-end raw reads obtained by the Illumina HiSeq 2000 system is 43,189,861 for *O. gaubae* sample. The plastid genome of *O. gaubae* with 122,873 bp in length and having only one copy of the IR region, which is in accordance with the reports and its genome structure is similar to those of other IRLC species. In this context, the lack of *rps*16 and *rpl*22 genes and intron 1 of *clp*P in the plastome of *O. gaubae* should be noted; these genes, are absent from the chloroplast genomes of entire IRLC[24-26]. The assembled chloroplast genome of *O. gaubae* contained 110 genes, including 76 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosome RNA (rRNA) genes (Fig. 1, Table 1). The LSC (81,066 bp),

SSC (13,777 bp) and IR (28,802 bp) regions along with the locations of 110 genes in the chloroplast genome are shown in Fig. 1.

Table 1

Genes predicted in the chloroplast genome of *O. gaubae*.

Category of genes	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl*14, *rpl*16*, *rpl*2*, *rpl*20, *rpl*23, *rpl*32, *rpl*33, *rpl*36
	Small subunit of ribosomal proteins	*rps*2, *rps*3, *rps*4, *rps*7, *rps*8, *rps*11, *rps*12*, *rps*14, *rps*15, *rps*18, *rps*19
	DNA-dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1*,*rpo*C2
	Ribosomal RNA genes	*rrn*16S, *rrn*23S, *rrn* 4.5S, *rrn* 5S
	Transfer RNA genes	30 *trn* genes (5 contain an intron)
Genes for photosynthesis	Subunits of NADH-dehydrogenase	*ndh*A*, *ndh*B*, *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunits of photosystem II	*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
	Subunits of cytochrome b/f complex	*pet*A, *pet*B*, *pet*D*, *pet*G, *pet*L, *pet*N
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F*, *atp*H, *atp*I
	Subunit of rubisco	*rbc*L
Other genes	Maturase K	*mat*K
	Envelope membrane protein	*cem*A
	Subunit of Acetyl-CoA-carboxylase	*acc*D
	C-type cytochrome synthesis gene	*ccs*A
	Protease	*clp*P*
Genes of unkown function	Conserved hypothetical chloroplast open reading frames	*ycf*1, *ycf*2, *ycf*4, *ycf*3**
The number of asterisks after the gene names indicates the number of introns contained in the genes.		

A total of 16 genes contained a single intron, whereas *ycf*3 exhibits two introns (Additional File 1: Table S1). The *rps*12 gene is a trans-splicing gene which does not have introns in the 3'-end. The *trn*K-UUU has

the largest intron encompassing the *mat*K gene, with 2,495 bp, whereas the intron of *trn*L-UAA is the smallest (542 bp). The *O. viciifolia* (the only other species of *Onobrychis* which cp genome has been sequenced) plastome with 121,932 bp in length showed similar conserved gene contents, order and orientations to *O. gaubae*. The most important structural differences in the *O. viciifolia* plastid genome compared to the *O. gaubae* are the lack of the *atp*F intron and inversion of *ycf*2 gene.

The length of plastome in the IRLC taxa in this study ranging from 121,020 to 130,561 bp. All plastomes exhibited the typical structure of IR-loss clade composed of LSC region (79,916 to 87,193), SSC region (13,383 to 14,187) and only one inverted repeat region (27,604 to 30,487) (Table 2).

Table 2

Chloroplast genome information from sampled IRLC species and the newly assembled *O. gaubae*.

Species	Size (bp)	LSC (bp)	GC (%) (LSC)	SSC (bp)	GC (%) (SSC)	IR (bp)	GC (%) (IR)	GC (%) Total
Astragalus mongholicus	123,582	80,986	33.4%	13,773	29.9%	28,823	38.1%	34.1%
Caragana microphylla	130,029	85,436	33.3%	14,106	30.4%	30,487	38.8%	34.3%
Carmichaelia australis	122,805	80,588	33.5%	14,074	30.2%	28,143	38.6%	34.3%
Cicer arietinum	125,319	82,583	33%	13,820	29.9%	28,916	38.3%	33.9%
Galega officinalis	125,086	82,915	33.2%	13,347	30.5%	28,824	38.7%	34.2%
Glycyrrhiza glabra	127,943	84,714	33.1%	14,187	30.1%	29,042	39.6%	34.2%
Hedysarum semenovii	123,407	80,288	34.1%	13,679	30.5%	29,440	38.9%	34.9%
Lens culinaris	122,967	81,659	33.7%	13,833	30.2%	27,604	38.7%	34.4%
Lessertia frutescens	122,700	80,698	33.4%	13,750	29.9%	28,252	38.4%	34.2%
Medicago sativa	125,330	83,756	32.9%	13,383	30.2%	28,191	38.6%	34%
Melilotus albus	127,205	84,279	32.7%	13,806	29.8%	29,120	38.1%	33.6%
Meristotropis xanthioides	127,735	84,629	33.1%	14,150	30.1%	28,956	39.6%	34.2%
Onobrychis gaubae	122,873	79,968	33.7%	13,805	30.5%	29,100	38.8%	34.5%
Onobrychis viciifolia	121,932	78,986	33.8%	13,821	30.4%	29,125	38.8%	34.6%
Oxytropis bicolor	122,461	80,170	33.5%	14,017	30%	28,274	38.3%	34.2%
Tibetia liangshanensis	123,372	79,916	33.9%	13,513	30.6%	29,943	38.6%	34.7%
Wisteria floribunda	130,561	87,193	33.2%	14,127	30%	29,628	39.4%	34.4%
LSC: Large Single Copy, SSC: Small Single Copy, IR: Inverted Repeat.								

Gene order and gene/intron content in plastomes of all IRLC taxa are highly conserved. The overall GC content of the *O. gaubae* chloroplast genome sequence was 34.5%, which is consistent with other IRLC species, whose plastomes have GC-contents ranging from 33.6% to 35.1% (Table 2). Different GC content occurs in the LSC (32. 7% - 34.1%), SSC (29.8% - 30.6%) and IR (38.1% - 39.6%) regions (Table 2).

Codon usage bias

The total coding DNA sequences (CDSs) were 81,121 bp in length and encoded 75 genes including 24,765 codons which belonged to 61 codon types. Codon usage was calculated for the protein-coding genes present in the *O. gaubae* cp genome. Phenylalanine was the most abundant amino acid, whereas Alanine showed the least abundance in this species (Additional File 1: Table S2). Most protein-coding genes employ the standard ATG as the initiator codon. Among the *O. gaubae* protein-coding genes, three genes used alternative start codons; ACG for *psb*L and *ndh*D, and GTG for *rps*8. Similar codon usage pattern was exhibited in the *O. viciifolia* species (Additional File 1: Table S3).

The chloroplast genomes of the IRLC were analyzed for their codon usage frequency according to sequences of protein-coding genes and relative synonymous codon usage (RSCU). RSCU is an important indicator to measure codon usage bias in coding regions. This value is the ratio between the actual observed values of the codon and the theoretical expectations. If RSCU=1, codon usage is unbiased; if RSCU>1, specific codon frequency is higher than other synonymous codons, otherwise, the frequency is low[27,28]. The total number of codons among protein-coding genes in the IRLC species varies from 20,381 codons in *Hedysarum taipeicum* (as the smallest number) to 24,765 codons in *O. gaubae*. The most often used synonymous codon was AUU, encoding isoleucine, and the least used was CGC/CGG, encoding arginine (Additional File 2: Table S4). In the IRLC, the standard AUG codon was usually the start codon for the majority of protein-coding genes and UAA was the most frequent stop codon among three stop codons. Methionine (AUG) and tryptophan (UGG) showed RSCU = 1, indicating no codon bias for these two amino acids. The highest RSCU value was for UUA (~2.04) in leucine and the lowest was GGC (~0.35) in glycine. Leucine preferred six codon types (UUA, UUG, CUU, CUC, CUA, and CUG) and actually showed A or T (U) bias in all synonymous codons (Additional File 2: Table S4). The result of distributions of codon usage in the IRLC species showed that RSCU>1 was recorded for most codons that ended with an A or a U, except for UUG codon, resulting in the bias for A/T bases. As well as, more codons with the RSCU value less than one, ended with base C or G. So, there is high A/U preference in the third codon of the IR-loss clade coding regions, which is a common phenomenon in cp genomes of higher plants[29].

Analysis of repeats

Repeat analysis of the *O. gaubae* plastome identified 50 repeat structures with lengths ranging from 30 bp to 179 bp. These structures included 29 forward repeats with lengths in the range of 30-179 bp, 19 palindromic repeats of 30-81 bp and two reverse repeats with a length of 31 and 37 bp (Additional File 3: Table S5). Among the 50 repeats, 66% are located in the LSC region, 18% in the IR region and 16% in the SSC region. Also, most of the repeats (42%) were found in coding region (*acc*D, *psa*A, *psa*B, *psb*C, *psb*J, *ycf*1, *ycf*2, *ycf*4, *rps*12, *trn*R-UCU, *trn*K-UUU), 40% were distributed in the intergenic spacer regions (IGS) and 18% were located in the introns (*ndh*A, *rpl*16, *rps*12, *pet*B, *ycf*3). The pattern of repeat structures (both in frequency and location) in *O. gaubae* is similar to that of *O. viciifolia* (Additional File 3: Table S6). In the majority of the studied IRLC species, the most frequently observed repeats were forward, then palindromic, and the least was the reverse (Fig. 2).

The forward type was the most abundant repeat with length in the range 30-50 bp in all the IRLC species. The longest repeats were also of the forward type, with length of 560 bp were detected in the *Hedysarum taipeicum*, followed by *Vicia sativa* of 517 bp and *Caragana microphylla* of 455 bp, which were much longer than other species studied.

Simple sequence repeats (SSRs), or microsatellites, are a type of tandem repeat sequences which contain 1-6 nucleotide repeat units and have wide distribution throughout the genome[28,30]. Accordingly, microsatellites play a crucial role in the genome recombination and rearrangement. These nucleotide motifs show a high level of polymorphism that can be widely used for phylogenetic analysis, population genetics and species authentication[28,31]. A total of 83 SSRs were detected in the *O. gaubae* plastome, which were composed by a length of at least 10 bp. Among them, 47 (56.62%) were mono-repeats, 24 (28.91%) were di-repeats, 6 (7.22%) were tri-repeats, five (6.02%) were tetra-repeats and one were penta-repeats (1.2%). No hexanucleotide SSRs existed in *O. gaubae* genome (Additional File 3: Table S7). *Onobrychis viciifolia* with 101 SSRs including 50 mono-repeats (49.5%), 30 (29.7%) di-repeats, nine (8.91%) tri-repeats, 11 (10.89%) tetra-repeats and one penta-repeat (0.99%), exhibited similar SSR distribution pattern in the plastome (Additional File 3: Table S8). The number of SSRs in the IRLC cp genomes (cpSSRs) ranged from 68 (*Vicia sativa* and *Lens culinaris*) to 151 (*Melilotus albus*) across the IRLC species (Fig. 3A). The mononucleotide repeats (P1) were identified at a much higher frequency, which varied from 45 (*Tibetia liangshanensis*, *Glycyrrhiza glabra*) to 93 (*Melilotus albus*) (Fig. 3B).

In the mononucleotide repeats, A/T motifs were the most abundant but no G/C motif was detected in the cp genome. Likewise, the majority of the dinucleotides and trinucleotides were found to be particularly rich in AT sequences. Therefore, the AT richness in the SSRs of the chloroplast genome of *O. gaubae* is consistent with the results of previous studies[28,32] which have shown that in the cp genome, SSRs generally composed of polythymine (poly T) or polyadenine (poly A) repeats[33].

Sequence divergence analysis

The average nucleotide diversity (Pi) among the protein-coding genes of 23 species of the IRLC was estimated to be 0.05736. Furthermore, comparison of nucleotide diversity in the LSC, SSC and IR regions indicated that the IR region exhibits the highest nucleotide diversity (0.11549) and the SSC region shows the least (0.04132). We detected three hyper-variable regions with Pi values>0.1 among the IRLC species; *ycf*1 and *ycf*2 from IR region and *clp*P from LSC region (Fig. 4).

These genes might be undergoing rapid nucleotide substitution in IR-loss clade at the genus and species levels. Among these, *ycf*1 encoding a protein of 1800 amino acids has the highest nucleotide diversity (0.18745). The average nucleotide diversity was also investigated between two *Onobrychis* plastid genome sequences. The average value of Pi between the *Onobrychis* species was estimated to be 0.05632 (Additional File 4: Fig. S1). High nucleotide variations were observed for the protein-coding regions *ycf*1, *ycf*2, *clp*P, *acc*D and *ycf*4 and intergenic regions such as *trn*L(UAA)-*trn*T(UGU) and

*trn*T(GGU)-*trn*E(UUC). Sliding window analysis results revealed the same variable regions in the cp genome of the two *Onobrychis* species.

Moreover, we compared whole chloroplast genome sequences of different taxa of IRLC to analyze gene order and content with mVISTA. We found that, similar to other plant species, the gene coding regions were more conserved than the noncoding regions (Additional File 5: Fig. S2). High nucleotide variations were observed across IRLC for the protein-coding regions *ycf*1, *ycf*2 and *clp*P. Similar results were also obtained from calculation of nucleotide diversity (Pi).

Ka/Ks analysis

In this study, the non-synonymous (Ka) to synonymous (Ks) rate ratio (Ka/Ks) was estimated for 75 protein-coding genes across the 28 IRLC species analyzed (Additional File 6: Table S9). In general, the Ka/Ks values were lower than 0.5 for almost all genes. The *ycf*4 gene which is involved in regulating the assembly of the photosystem I complex had the highest nonsynonymous rate, 0.165691, while the *ycf*1 gene with unknown functions had the highest synonymous rate, 0.181067. The Ka/Ks ratio (denoted as ω) is widely used as an estimator of selective pressure for protein-coding genes. An $\omega > 1$ indicates that the gene is affected by positive selection, $\omega < 1$ indicates purifying (negative) selection, and ω close to 1 indicates neutral mutation[34]. In the present study, the Ka/Ks ratio was calculated to be 0 for *psb*L gene which encodes one of the subunits of photosystem II. The Ka/Ks ratio indicates purifying selection in 73 protein-coding genes. The highest Ka/Ks ratio which indicates positive selection was observed in *acc*D gene which encodes a subunit of the acetyl-CoA carboxylase (ACCase) enzyme.

Prediction of RNA editing sites

RNA editing as a post-transcriptional modification process, mainly occurs in chloroplasts and mitochondrial genomes. In higher plants, some chloroplast RNA editing sites which provide a way to create transcript and protein diversity are conserved[28].

RNA editing sites of *O. gaubae* plastid genes were predicted using Prep-CP prediction tool (Additional File 7: Table S10). In total, 58 editing sites were present in 19 chloroplast protein-coding genes and all of the editing sites were C-to-U conversions (Additional File 7: Table S10). Among them, nine editing sites, the highest number, were found in the region encoding *ndh*B gene followed by seven editing sites in *pet*B. There were six editing sites detected each in *ndh*A and *rpo*B genes. *acc*D, *ndh*G and *pet*D had three editing sites, and *ndh*D and *ndh*F had two editing sites. Two editing sites were also found in *ccs*A, *mat*K and *rpo*C1 genes. The remaining seven genes had only one editing site. The results showed that *ndh* genes exhibited the most abundant editing sites which were nearly 39.6% of the total editing sites. Furthermore, we predicted 65 RNA editing sites out of 22 plastid genes in chloroplast genomes of *O. viciifolia*. In this species, the highest number of editing sites belongs to the *pet*B, *rpo*C1 and *ndh*B genes with 9, 8 and 7

sites, respectively (Additional File 7: Table S11). In these two *Onobrychis* species, the amino acid conversion from S to L occurred most frequently, while R to W occurred least.

Phylogenetic analysis

Phylogenetic relationships within the IRLC were reconstructed using the representative taxa (28 species from different tribes) and two species as outgroup based on 75 protein-coding genes of their chloroplast genomes. The total concatenated alignment length from the 75 protein-coding genes was 87,455 bp. The maximum likelihood (ML) analysis resulted in a well-resolved tree and the Bayesian inference yielded a well-resolved topology with high support values (Fig. 5).

The ML and Bayesian trees were largely congruent. The reconstructed phylogeny is in agreement with previous studies[5,14,25,35] indicated that IRLC was monophyletic and consisted of several clades. As shown in the previous studies, *Glycyrrhiza* + *Meristotropis* were monophyletic, along with the tribe Wisterieae was sister to the rest of the IRLC[19,36]. Then, there are two major clades: clade I and II (Fig. 5). Clade I comprises tribes Caraganeae[17], Hedysareae[20] and Coluteae[18] as well as genera *Oxytropis* and *Astragalus*. Our results showed a close relationship between *O. gaubae* and *O. viciifolia* and *Hedysarum* species and confirmed *O. gaubae* phylogenetic position in the tribe Hedysareae. Furthermore, our plastid DNA analyses which are consistent with the previous study[35], show that *Oxytropis* is sister to the tribe Coluteae. In clade II, tribe Cicereae is the basal branch and formed a sister group relationship with the paraphyletic Trifolieae and the monophyletic tribe Fabeae.

Discussion

General features of the *O. gaubae* plastid genome

In our study, we determined the first complete chloroplast genome sequence of *O. gaubae* within *O.* subgenus *Sisyrosema* using the Illumina platform and deposited in the NCBI Genbank. Our assembly and annotation results showed that the length of the cp genome is 122,873 bp and its structure is similar to those of other IRLC species. Gene number, order, and type were found to be very similar between *O. gaubae* and *O. viciifolia*. In other words, the two *Onobrychis* plastid genomes were relatively conserved with the IR region more conserved than LSC and SSC regions. In this regard, one of the structural changes detected in the *O. viciifolia* cp genome is the loss of the *atp*F intron; whereas, *O. gaubae* possesses this intron. The *atp*F gene of *O. gaubae* is 1261 bp long with one intron of 702 bp, exon 1 of 144 bp and exon 2 of 415 bp. While the *atp*F gene of *O. viciifolia* is 558 bp long. The *atp*F gene has a conserved group II intron which has been found in the most previously sequenced land plant plastomes[37]. The *atp*F intron is rarely lost in flowering plants but some intronless chloroplast genomes have been reported, including *Manihot* (Euphorbiaceae)[37], *Passiflora* (Passifloraceae)[38] and several taxa across IRLC (*Colutea nepalensis*, *Lessertia frutescens*, *Oxytropis bicolor*, *O. recemosa* and *Sphaerophysa salsula*)[35]. It has been suggested that recombination between an edited mRNA and the *atp*F gene may be a possible reason for

the lack of intron[37]. Structural variations such as intron presence/absence can be useful as a molecular marker and provide informative characters at low taxonomic levels in phylogenetic studies[25]. Another structural change in plastome of *O. viciifolia* is the inversion of *ycf*2 gene. Plastome inversions due to the relative rarity and easily determined homology (no homoplasy), are highly valuable and useful in phylogenetic studies[5]. The main cause of inversions is not fully understood, but intramolecular recombination between dispersed short inverted/direct repeats and tRNA genes is an accepted explanation[39,40].

It was previously reported that the plastomes of Papilionoideae, particularly IR-loss clade, are not conserved in their genomic structure in terms of gene order and gene content and exhibit numerous rearrangements and gene/intron losses[5,24,25]. In this context, our results showed that the lengths of the IRLC plastid genomes ranged from 121,020 to 130,561 bp. This suggests that the IRLC cp genomes may have undergone different evolutionary processes such as gene/intron loss, insertion/deletion and IR/LSC/SSC expansion/contraction[41]. The plastomes among the IRLC taxa were similar in GC content but higher GC contents were usually detected in the IR region compared to the other regions of cp genome, which is mainly due to the presence of rRNA genes (*rrn*23, *rrn*16, *rrn*5, *rrn*4.5) with high GC content (50%-56.4%) in IRs [6,32]. One of the factors that shape codon usage biases in different organisms is the GC content in codon positions. Codon usage bias indicates the importance of molecular evolutionary phenomena. As mentioned above, codon usage patterns are similar between two *Onobrychis* species and also across IRLC.

In this study, we found many repeat regions including forward repeats, palindromic repeats and reverse repeats, of which forward types were the most frequent in the IR-loss clade. Furthermore, in the IRLC, repeat sequences involved in genome rearrangement, were mainly distributed in non-coding regions (IGS). Repeat structures induce indels and substitutions resulting in the mutation hotspot in the reconfiguration of genome[6]; therefore, these repeats can provide valuable information for phylogenetic and population studies[28]. Moreover, we detected SSRs in two *Onobrychis* species and representative species of the IRLC. In the IR-loss clade, mononucleotide repeats were highly abundant and were mostly composed of AT than G/C repeats. Strong AT bias in SSR loci was also observed in other legumes such as *Vigna radiate*[42], *Arachis hypogaea*[43] and *Stryphnodendron adstringens*[32] which, like other plastomes of species, may contribute to the bias in base composition[6]. The results showed that SSR loci of LSC regions appeared more frequently than in SSC or IR regions, which may be hypothesized that this phenomenon is relevant to the lack of one IR region in IR-loss clade. In general, cpSSRs show abundant variation and might provide useful information for detecting intra- and interspecific polymorphisms at the population level[30,33].

Divergent hotspots in the IRLC chloroplast genomes

In this study, the sequence divergence regions that could be used as effective molecular markers were identified based on mVISTA and sliding window analysis. In the IR-loss clade, the divergence was distributed in the LSC and IR regions. Three highly variable regions (*clp*P, *ycf*1, *ycf*2) were observed with the higher Pi values and were located at LSC and IR regions, respectively. The gene *ycf*1 with the highest nucleotide diversity is more variable than *mat*K and it can be useful for molecular systematics at low taxonomic levels[44,45]. Furthermore, variable regions including *ycf*1, *ycf*2, *clp*P, *acc*D and *ycf*4 (as the protein-coding regions) and *trn*L(UAA)-*trn*T(UGU) and *trn*T(GGU)-*trn*E(UUC) (as the intergenic regions) could be ideal as molecular markers to distinguish the *Onobrychis* species. Several studies[14,18,46] analyzed the phylogenetic reconstructions of the IRLC species at various taxonomic levels based on different fragments of plastid coding genes such as *mat*K, *ndh*F and *rbc*L, the nuclear ribosomal ITS and the combined sequences of these genes/spacers. We could use the highly variable regions acquired from this study to develop the potential phylogenetic markers which can be useful for species authentication and reconstruction of phylogeny within different tribes/genera of IR-loss clade in further studies.

Papilionoideae, in particular the IRLC, displays structural variations which provide informative characters to increase phylogenetic resolution and make the taxon an excellent model for genome evolution studies[5,25]. The plastomes of several members of the IRLC have regions with significant variation and rearrangement and accelerated mutation rates, including loss of introns from *rps*12 and *clp*P genes[24], absence of *rps*16 gene[26] and transfer/loss of *rpl*22 to the nucleus[24]. Numerous studies have also shown some other rearrangements in some IRLC taxa, such as loss of *acc*D gene in six species of *Trifolium*[10,25], loss of *rpl*23 and *rpl*33 genes in some species of *Lathyrus*, *Pisum* and *Vicia*[31] and loss of *ycf*4 gene in some species of *Lathyrus* and *Pisum*[15,16]. As revealed in other studies, there are several reasons for occurrence of rearrangements in the plastome, such as the lack of one IR region, variable IR region size and many tandemly repeated sequences[47]. For example, the loss of the *rps*16 gene was probably due to the presence of a nuclear *rps*16 copy, which contributed to pseudogenization of the plastid copy[47]. Likewise, the lack or expansion of the *acc*D gene was explained by the presence of tandemly repeated sequences[6,15].

Positive selection analysis

Moreover, we estimated the Ka/Ks for each gene in DnaSP v.6.12. Acceleration of the evolutionary rate was observed only in the *acc*D gene. Some studies have investigated whether selective pressure is acting on a particular protein-coding gene in different genera/tribes of IR-loss clade. For instance, tests for positive selection suggested that *Lathyrus*, *Pisum* and *Vavilovia*, all belonging tribe Fabeae, have undergone adaptive evolution in the *ycf*4 gene[15,16].

Legumes chloroplast genome, and in particular IRLC, have regions with high mutation rates, including *rps*16-*acc*D-*psa*I-*ycf*4-*cem*A region. *rps*16 gene was lost from cpDNA in the common ancestor of the IR-

loss clade[15]. *acc*D coding region was completely absent in the *T.* subgenus *Trifolium* and has nuclear copies in *Medicago truncatula* and *Cicer arietinum*[25]. Three consecutive genes *psa*I-*ycf*4-*cem*A is situated in a local mutation hotspot and has been lost in some species of *Lathyrus*[15,16].

Plastid RNA editing prediction

RNA editing is one of the post-transcriptional events which converts cytidine (C) to uridine (U) or U to C at specific sites of RNA molecules and modifies the genetic information from the genome in the plastids and mitochondria of land plants. RNA editing serves as a mechanism to correct missense mutations of genes by inserting, deleting and modifying nucleotides in a transcript[48]. In this study, between two species of *Onobrychis* (*O. gaubae* and O. *viciifolia*) and also in the IRLC genera, the most editing sites were observed for *ndh* genes. In this regard, the highest number of plastid editing sites was found in the *ndh* group genes in flowering plants[48]. Moreover, the *ndh* genes encoding for a thylakoid Ndh complex, have been lost or pseudogenized in different species of algae, bryophytes, pteridophytes, gymnosperms, monocots, eudicots, magnoliids, and protists[49-51]. The RNA editing is probably important for the NDH protein complex function and may also lead to improved photosynthesis and display positive selection during evolution[48].

Phylogenetic relationships

With the use of the whole cp genome coding sequence from 28 representative species of the IR-loss clade, a highly consistent topology was recovered by ML and Bayesian analyses (Fig. 5). The monophyly of the IRLC and all its tribes was consistent with all previous studies[5,14,25,35]. Wisterieae together with *Glycyrrhiza* and *Meristotropis* were the first diverging lineage as sister to the remaining taxa. Tribes Caraganeae and Hedysareae were grouped together. Many previous studies showed that *Astragalus* was sister to the genus *Oxytropis* but recent study on the chloroplast phylogenomics of *Astragalus* reported that *Astragalus* is a monophyletic clade and *Oxytropis* is sister to the Coluteoid clade[35], which is in agreement with the present study. Cicereae + Trifolieae + Fabeae formed well supported sister groups. The results of the present study suggest that there is no conflict between the phylogeny made by whole cp genome and that inferred by individual gene datasets. Therefore, a phylogenetic reconstruction for IR-loss clade species studied here showed that plastid genome database will be a helpful resource for molecular phylogeny at the higher taxonomic level (generic to tribal rank).

Conclusions

In this study, the complete plastome sequence of *O. gaubae* (122,873 bp) was determined. The gene contents and gene orientation of *O. gaubae* plastome are similar to those found in the plastid genome of other IRLC species. Comparison of plastomes across IRLC showed that the coding regions are more conserved than non-coding regions and IR is more conserved than LSC and SSC regions. The present

study also analyzed genetic information in the IRLC plastomes including the distribution and location of repeat sequences and SSRs, codon usage, RNA editing prediction, hotspot regions and phylogenomic analysis. Moreover, we identified three hotspot genes (*ycf*1, *ycf*2, *clp*P) which provided sufficient genetic information for species identification and phylogenetic reconstruction of the IRLC species. Seven hypervariable regions including *ycf*1, *ycf*2, *clp*P, *acc*D and *ycf*4 (as the protein-coding regions) and *trn*L(UAA)-*trn*T(UGU) and *trn*T(GGU)-*trn*E(UUC) (as the intergenic regions) were also identified between *Onobrychis* species, which could be used to distinguish species. Finally, the data obtained from this study could provide a useful resource for further research on tribe Hedysareae and also IR-loss clade at the genomic scale.

Methods

Chloroplast DNA extraction and sequencing

The young leaves of *O. gaubae* were collected from the southern slopes of Alborz mountain range in Tehran, Iran. It was identified by Professor S. Kazempour-Osaloo. This species were preserved in the Tarbiat Modares University Herbarium (TMUH) (voucher code: 2016-1). Permission was not necessary for collecting the samples, which has not been included in the list of national key protected plants. The fresh leaves were immediately dried with silica gel for further DNA extraction. Our experimental research, including the collection of plant materials, are complies with institutional, national or international guidelines. Genomic DNA was extracted from dried leaves using a DNeasy Plant Kit (Qiagen) according to the manufacturer's instructions. DNA quality and quantity were confirmed using 1% agarose gel electrophoresis and the resulting DNA was sequenced using the Illumina HiSeq-2000 platform at Iwate Biotechnology Research Center. The paired-end libraries were constructed according to the manufacturer's protocol (Illumina Inc., San Diego, CA). In total, 43,189,861 paired-end reads each comprising 100-bp sequence were obtained.

Genome Assembly and Annotation

Using the complete plastid genome of *Onobrychis viciifolia* (MW007721) as the reference, the paired-end reads of *O. gaubae* were filtered and assembled in to a complete plastome using Fast-Plast (https://github.com/mrmckain/Fast-Plast)[52]. Furthermore, we compared the chloroplast genome of *O. gaubae* with the complete chloroplast sequence of other Hedysareae species (*Hedysarum* and Alhagi species). Gaps in the cpDNA sequences were filled by PCR amplification and Sanger sequencing. The de novo assembled chloroplast genomes were annotated by GeSeq[53]. We used the online tRNAscan-SE service[54] to improve the identification of tRNA genes. To detect the number of matched reads and the depth of coverage, raw reads were remapped to the assembled plastomes with Bowtie2[55] as implemented in Geneious v.9.0.2. The entire chloroplast genome sequences of *O. gaubae* was deposited in GenBank (Accession Number:?).

Codon usage

Codon usage was determined for all protein-coding genes. The codon usage analysis was performed in the web server Bioinformatics (https://www.bioinformatics.org/sms2/codon_usage.html). Furthermore, the relative synonymous codon usage (RSCU) values were determined with MEGA X[56], which was used to reveal the characteristics of the variation in synonymous codon usage.

Characterization of repeat sequences

REPuter was used to identify forward repeats, reverse sequences, complementary and palindromic sequences, with a minimal size of 30 bp, hamming distance of 3 and over 90% identity. Simple sequence repeats (SSRs) were detected using the microsatellite identification tool MISA (available online: http://pgrc.ipk-gatersleben.de/misa/misa.html). The minimum numbers of the SSR motifs were 10, 5, 4, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively.

Divergent hotspots identification and synonymous (Ks) and non-synonymous (Ka) substitution rate analysis

To assess the nucleotide diversity (Pi) among the plastid genomes of the representative species of the IRLC, the whole chloroplast genome sequences were aligned using MAFFT[57] on XSEDE v.7.402 in CIPRES Science Gateway[58]. A sliding window analysis was conducted to determine the nucleotide diversity of the chloroplast genome using DnaSP v.6.12 software[59]. The window length was set to 800 bp and the step size was 200 bp. Furthermore, the protein-coding regions of the 28 chloroplast genomes were used to evaluate evolutionary rate variation within the IRLC. Thus, we aligned the 75 protein-coding genes separately using MAFFT and then estimated the synonymous (Ks) and non-synonymous (Ka) substitution rates, as well as their ratio (Ka/Ks) using DnaSP v.6.12 software.

Genome comparison

To investigate divergence in chloroplast genomes, identity across the whole cp genomes was visualized using the mVISTA viewer in the Shuffle-LAGAN mode[60] among the 19 IRLC accessions using *Glycyrrhiza glabra* as the reference.

Prediction of potential RNA editing sites

Thirty-five protein-coding genes of *O. gaubae* were used to predict potential RNA editing sites using the Predictive RNA Editor for Plants (PERP)-Cp web server (http://prep.unl.edu)[61] with a cutoff value of 0.8.

Phylogenetic reconstruction

Seventy-five protein-coding genes were recorded from 28 species within IRLC, as well as from two outgroups (*Robinia pseudoacacia* L. and *Lotus japonicus* (Regel) K.Larsen). All genes sequences were obtained from GenBank (Additional File 8: Table S12). The concatenated data were analyzed using maximum likelihood and Bayesian inference methodologies. Prior to maximum likelihood and Bayesian analyses, a general time reversible and gamma distribution (GTR + G) model was selected using the MrModeltest2.2[62] under the Akaike Information Criteria (AIC)[63]. Maximum likelihood analyses were

performed using the online phylogenetic software W-IQ-TREE[64] available at http://iqtree.cibiv.univie.ac.at. Nodes supports were calculated via rapid bootstrap analyses with 5000 replicates. Bayesian inference was performed using MrBayes v.3.2 in the CIPRES[58] with the following settings: Markov chain Monte Carlo simulations for 5,000,000 generations with four incrementally heated chains, starting from random trees and sampling one out of every 1,000 generations. The first 25% of the trees were regarded as burn-ins. The remaining trees were used to construct a 50% majority-rule consensus tree and to estimate posterior probabilities. Posterior probabilities (PP) > 0.95 were considered as significant support for a clade.

Abbreviations

SSR: Simple sequence repeat; cp: Chloroplast; IRs: Inverted repeats; LSC: Large single-copy; SSC: Small single-copy; IRLC: Inverted repeat lacking clade; ML: Maximum-likelihood; Ks: Synonymous substitution rates; Ka: Nonsynonymous substitution rates; RSCU: Relative synonymous codon usage; DnaSP: DNA sequence polymorphism; NCBI: National Center for Biotechnology; Pi: Nucleotide diversity/polymorphism; GTR: General time reversible; ITS: Internal transcribed spacer of ribosomal DNA; rRNA: Ribosomal RNA; tRNA: Transfer RNA.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of Data and materials

Sequences used in this study are available from the National Center for Biotechnology Information (NCBI) (see Additional file 8: Table S12). Annotated sequence of plastome of *O. gaubae* were submitted to GenBank under ? accession number. Sample of *O. gaubae* is saved at the Tarbiat Modares University Herbarium, Tehran, Iran.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by Kyoto University, Iwate Biotechnology Research Center and Tarbiat

Modares University. The funders had no role in the design of the study, analysis of data, decision to publish and in manuscript preparation.

Authors' contributions

M. M. and S. K. O. conceived the idea, designed the study and carried out the plant sampling; M. M., A. O. and M. S. extracted chloroplast DNA for next generation sequencing, A. O. and M. S. assembled the genome, M. M. and A. O. performed the manual genome annotation, M. M. performed the phylogenetic and computational analyses, M. M. wrote the paper. R. T. and S. K. O. edited and reviewed the paper. All authors have read and approved the final manuscript.

Acknowledgements

Not applicable.

Author details

[1]Department of Plant Biology, Faculty of Biological Sciences, Tarbiat Modares University, Tehran 14115-154, Iran. [2]Graduate School of Agriculture, Kyoto University, Kyoto 617-0001, Japan. [3]Iwate Biotechnology Research Center, Kitakami, Iwate 024-0003, Japan.

References

1. Jansen RK, Ruhlman TA. Plastid Genomes of Seed Plants. In Genomics of Chloroplasts and Mitochondria. Springer, Dordrecht; 2012. p. 103–126.
2. Ruhlman, TA, and Jansen, RK. The plastid genomes of flowering plants in Chloroplast Biotechnology: Methods and Protocols. Maliga P, editor. Springer, Humana Press; 2014. p. 3–38.
3. Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, et al. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 2005;395:348–384.
4. Bock R. Structure, function, and inheritance of plastid genomes in Cell and Molecular Biology of Plastids. Bock R, editor. Springer-Verlag Berlin Heidelberg; 2007. p. 29–63.
5. Schwarz EN, Ruhlman TA, Sabir JSM, Hajrah NH, Alharbi NS, Al-Malki AL, et al. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps*16 in papilionoids. *J Syst Evol.* 2015;53:458–468.

6. Asaf S, Khan AL, Aaqil Khan M, Muhammad Imran Q, Kang S-M, Al-Hosni K, et al. Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. PLoS ONE. 2017; doi:12(8): e0182281.
7. Kim KJ, Choi KS, Jansen RK. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol. 2005;22(9): 1783 ± 92.
8. Haberle RC, Fourcade HM, Boore JL, Jansen RK. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. J. Mol. Evol. 2008;66(4): 350 ± 61.
9. Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, et al. The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution. Nucleic Acids Res. 2008;36(7): 2366 ± 78.
10. Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, et al. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J. Mol. Evol. 2008;67(6): 696 ± 704.
11. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol. Biol. Evol. 2011;28(1): 583 ± 600.
12. Palmer JD, Osorio B, Thompson WF. Evolutionary significance of inversions in legume chloroplast DNAs. Curr. Genet. 1988;14:65–74.
13. Legume Phylogeny Working Group. Legume phylogeny and classification in the 21st century: A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. Taxon. 2017;66:44–77.
14. Wojciechowski MF, Lavin M, Sanderson MJ. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *mat*K gene resolves many well-supported subclades within the family. Am. J. Bot. 2004;91:1846–1862.
15. Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, Perry AS, et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Research. 2010;20:1700–1710.
16. Moghaddam M, Kazempour-Osaloo S. Extensive survey of the *ycf*4 plastid gene throughout the IRLC legumes: Robust evidence of its locus and lineage specific accelerated rate of evolution, pseudogenization and gene loss in the tribe Fabeae. PLoS ONE. 2020; doi:15(3), e0229846.
17. Duan L, Yang X, Liu P, Johnson G, Wen J, Chang Z. A molecular phylogeny of Caraganeae (Leguminosae, Papilionoideae) reveals insights in to new generic and infrageneric delimitations. PhytoKeys. 2016;70:111–137.
18. Moghaddam M, Kazempour Osaloo S, Hosseiny H, Azimi F. Phylogeny and divergence times of the Coluteoid clade with special reference to *Colutea* (Fabaceae) inferred from nrDNA ITS and two cpDNAs, *mat*K and *rpl*32-trnL(UAG) sequences data. Plant Biosyst. 2017;6:1082–1093.
19. Compton JA, Schrire BD, Konyves K, Forest F, Malakasi P, Mattapha S, et al. The *Callerya* Group redefined and Tribe Wisterieae (Fabaceae) emended based on morphology and data from nuclear

and chloroplast DNA sequences. PhytoKeys. 2019;125;1–112.

20. Amirahmadi A, Kazempour Osaloo S, Moein F, Kaveh A, Maassoumi AA. Molecular systematic of the tribe Hedysareae (Fabaceae) based on nrDNA ITS and plastid *trn*L-F and *mat*K sequences. Plant Syst. Evol. 2014;300:729–747.
21. Lock JM. Tribe Hedysarae in *Legumes of the world*. Lewis G, Schrire B, Mackinder B, Lock M, editors. Royal Botanical Gardens, Kew; 2005. p.489–495.
22. Rechinger KH. Tribus Hedysareae *Papilionaceae II, Flora Iranica*. Rechinger KH, editor. Akademische Druckund Verlagsanstalt, Graz; 1984. p.387–464.
23. Kaveh A, Kazempour-Osaloo S, Amirahmadi A, Maassoumi A, Schneeweiss G. Systematics of *Onobrychis* sect. *Heliobrychis* (Fabaceae): morphology and molecular phylogeny revisited. Plant Syst. Evol. 2019;305:33–48.
24. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps*12 and *clp*P intron losses among legumes (Leguminosae). Mol. Phylogenet. Evol. 2008;48:1204–1217.
25. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NZ, Mutwakil M, et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol. J. 2014;12:743–754.
26. Doyle JJ, Doyle JL, Palmer JD. Multiple independent losses of two genes and one intron from legume chloroplast genomes. Syst. Bot. 1995;20:272–294.
27. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–1295.
28. Li X, Tan W, Sun J, Du J, Zheng C, Tian X, et al. Comparison of Four Complete Chloroplast Genomes of Medicinal and Ornamental *Meconopsis* Species: Genome Organization and Species Discrimination. Sci. Rep. 2019; doi:10.1038/s41598-019-47008-8.
29. Li CJ, Wang RN, Li DZ. Comparative analysis of plastid genomes within the Campanulaceae and phylogenetic implications. PLoS ONE. 2020; doi:15(5), e0233167.
30. Powell W, Morgante M, Mcdevitt R, Vendramin GG, Rafalski JA. Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes-Applications to the Population-Genetics of Pines. PNAS. 1995;92(17):7759 ± 63.
31. Lei W, Ni D, Wang Y, Shao J, Wang X, Yang D, et al. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. Sci. Rep. 2016; doi: 10.1038/srep21669.
32. Souza UJBD, Nunes R, Targueta CP, Diniz-Filho JAF, Telles MPD. The complete chloroplast genome of *Stryphnodendron adstringens* (Leguminosae - Caesalpinioideae): comparative analysis with related Mimosoid species. Sci. Rep. 2019; doi:10.1038/s41598-019-50620-3.
33. Zong D, Gan P, Zhou A, Li J, Xie Z, Duan A, et al. Comparative analysis of the complete chloroplast genomes of seven *Populus* species: Insights into alternative female parents of *Populus tomentosa*. PLoS ONE; 2019; doi:14(6), e0218455.

34. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of aminoacid sites under positive selection. Mol Biol Evol. 2005;22:1107–1118.
35. Su C, Duan L, Liu P, Liu J, Chang Z, Wen J. Chloroplast phylogenomics and character evolution of eastern Asian *Astragalus* (Leguminosae): Tackling the phylogenetic structure of the largest genus of flowering plants in Asia. Mol Phylogenet Evol. 2021; doi:10.1016/j.ympev.2020.107025.
36. Duan L, Harris AJ, Su C, Zhang Z-R, Arslan E, Ertugrul K, et al. Chloroplast Phylogenomics Reveals the Intercontinental Biogeographic History of the Liquorice Genus (Leguminosae: *Glycyrrhiza*). Front. Plant Sci. 2020; 11:793.
37. Daniell H, Wurdack KJ, Kanagaraj A, Lee S-B, Saski C, Jansen RK. The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atp*F in Malpighiales: RNA editing and multiple losses of a group II intron. Theor Appl Genet. 2008;116:723–737.
38. Jansen RK, Cai Z, Raubeson LA, Daniell H, de Pamphilis CW, Leebens-Mack J, et al. Analysis of 81 genes from 64 plastid genomes resolves relation-ships in angiosperms and identifies genome-scale evolutionary pat-terns. Proc Natl Acad Sci USA. 2007;104:19369–19374.
39. Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR. Arecurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol. Phylogenet. Evol. 2014;72:82–89.
40. Wang HX, Liu H, Moore MJ, Landrein S, Liu B, Zhu ZX, et al. Plastid phylogenomic insights into the evolution of the Caprifoliaceae s.l. (Dipsacales). Mol Phylogenet Evol. 2020; doi: 10.1016/j.ympev.2019.106641.
41. Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol. Biol. 2011;76:273–297.
42. Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, et al. The Chloroplast Genome Sequence of Mungbean (*Vigna radiata*) Determined by High-throughput Pyrosequencing: Structural Organization and Phylogenetic Relationships. DNA Res. 2010; doi: 10.1093/dnares/dsp025.
43. Yin D, Wang Y, Zhang X, Ma X, He X, Zhang J. Development of chloroplast genome resources for peanut (*Arachis hypogaea* L.) and other species of *Arachis*. Sci. Rep. 2017; doi: 10.1038/s41598-017-12026-x.
44. Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Norris H, et al. Phylogenetic utility of *ycf*1 in orchids: a plastid gene more variable than *mat*K. Plant Syst. Evol. 2009;277:75–84.
45. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. *ycf*1, the most promising plastid DNA barcode of land plants. Sci. Rep. 2015; doi.:10.1038/srep08348.
46. Schaefer H, Hechenleitner P, Santos-Guerra A, Sequeira MMD, Pennington RT, Kenicer G, et al. Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-atlantic island lineages. BMC Evol. Biol. 2012;12:250.

47. Keller J, Rousseau-Gueutin M, Martin GE, Morice J, Boutte J, Coissac E, et al. The evolutionary fate of the chloroplast and nuclear *rps*16 genes as revealed through the sequencing and comparative analyses off our novel legume chloroplast genomes from *Lupinus*. DNA Res. 2017;24:343–358.
48. He P, Huang S, Xiao G, Zhang Y, Yu J. Abundant RNA editing sites of chloroplast protein-coding genes in *Ginkgo biloba* and an evolutionary pattern analysis. BMC Plant Biol. 2016;16:257.
49. Blazier J, Guisinger MM, Jansen RK. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol. Biol. 2011;76:263–272.
50. Ruhlman TA, Chang W-J, Chen JJW, Huang Y-T, Chan M-T, Zhang J, et al. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biol. 2015;15:100.
51. Sanderson MJ, Copetti D, Burquez A, Bustamante E, Charboneau JLM, Eguiarte LE, et al. Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. Am. J. Bot. 2015; doi: 10.3732/ajb.1500184.
52. McKain MR, Wilson M. mrmckain/Fast-Plast: Fast-Plast v.1.2.8. Version v.1.2.8. 2018. https://github.com/mrmckain/Fast-Plast.
53. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6-W11.
54. Schattner P, Brooks AN, Lowe TM. The tRNA scan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:W686 ± W9.
55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012;9: 357–359.
56. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 2018;35:1547–1549.
57. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Sofware Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 2013;30:772–780.
58. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES science gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE); New Orleans, Louisiana. 2010.
59. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol. Biol. Evol. 2017;34:3299–3302.
60. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–W279.
61. Mower JP. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 2009;37:W253–W259.
62. Nylander JAA. MrModeltest v2. Program distributed by the author. Uppsala: Evolutionary Biology Centre, Uppsala University. 2004.

63. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 2004;53:793–808.
64. Trifinopoulos J, Nguyen LT, Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44(W1):W232–W235.

Figures

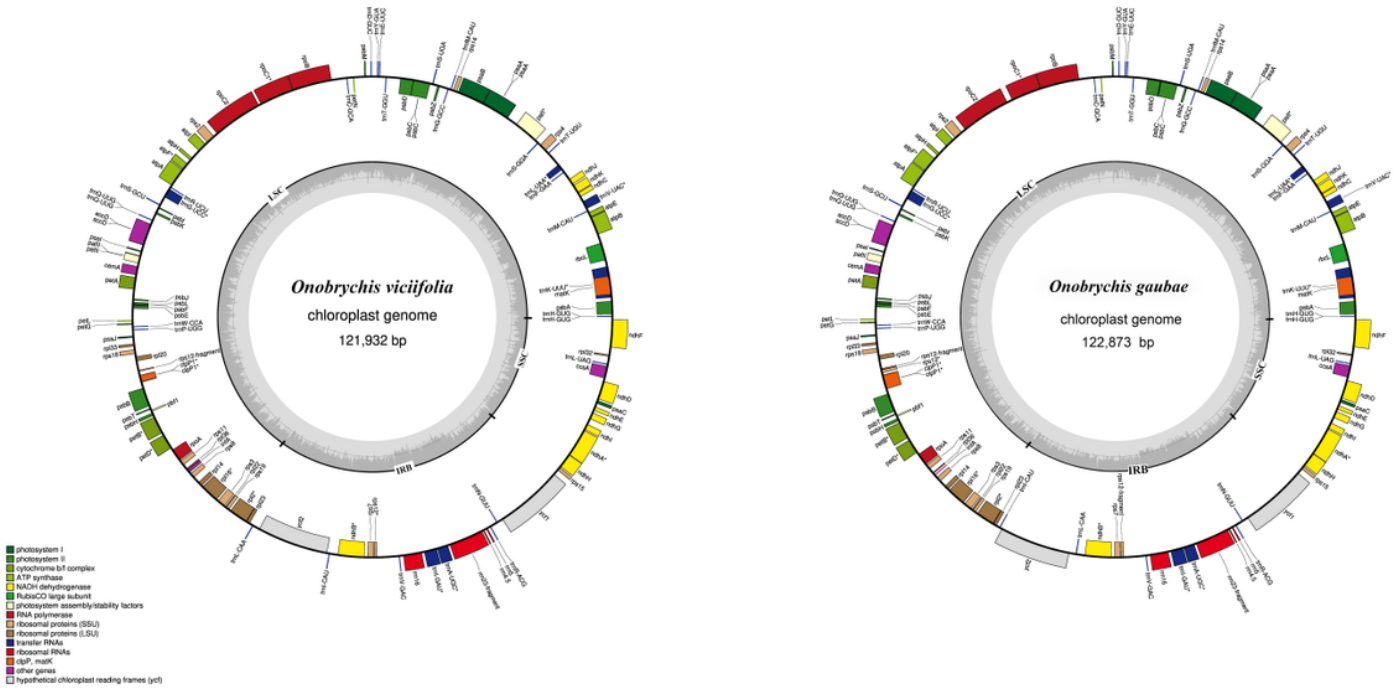

Figure 1

Gene map of two Onobrychis species chloroplast genome. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the structure of the chloroplast: The large single copy (LSC), small single copy (SSC) and inverted repeat (IR) regions. The gray ring marks the GC content with the inner circle marking a 50% threshold. Asterisks mark genes that have introns.

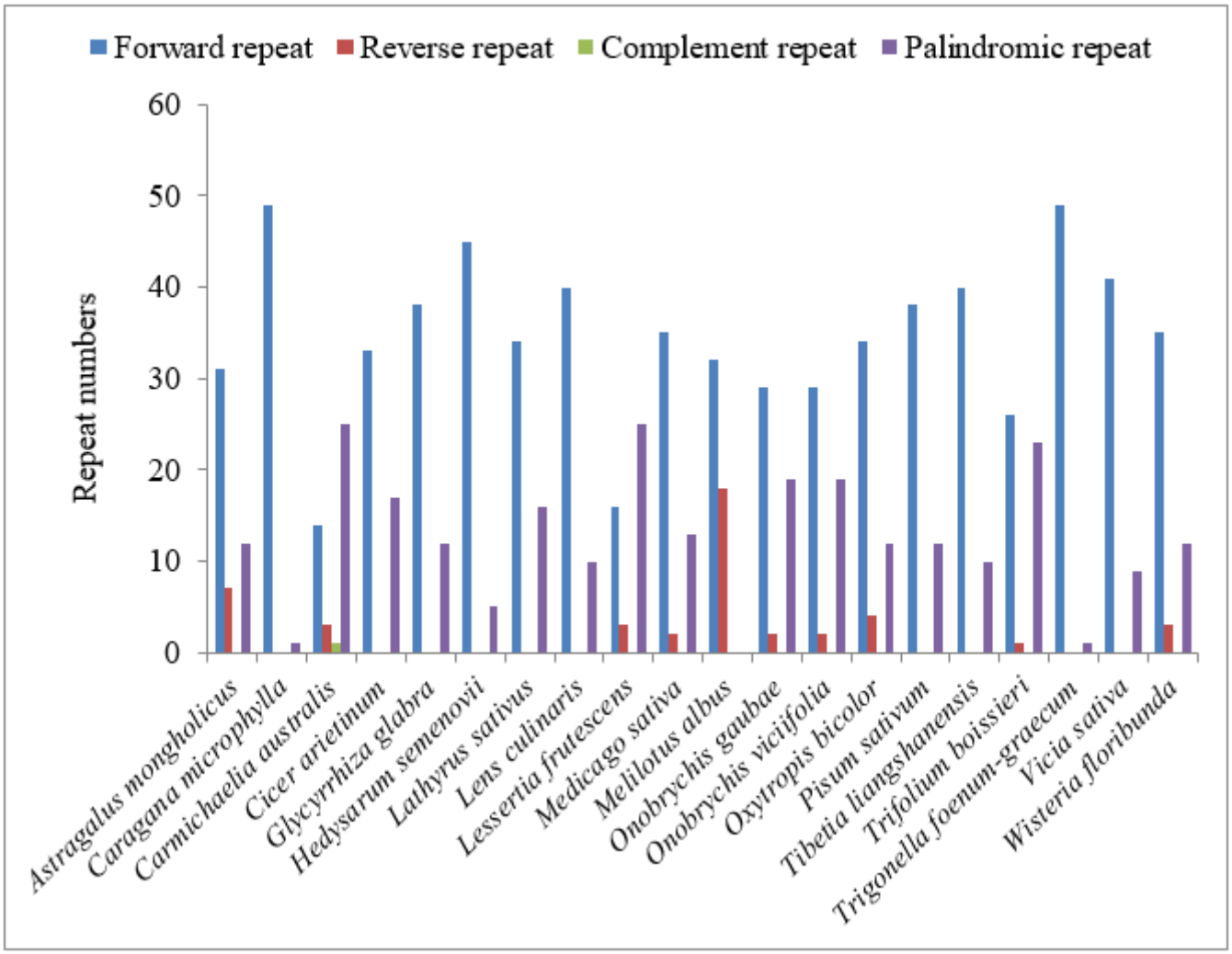

Figure 2

Analysis of repeated sequences in the IRLC species chloroplast genomes.

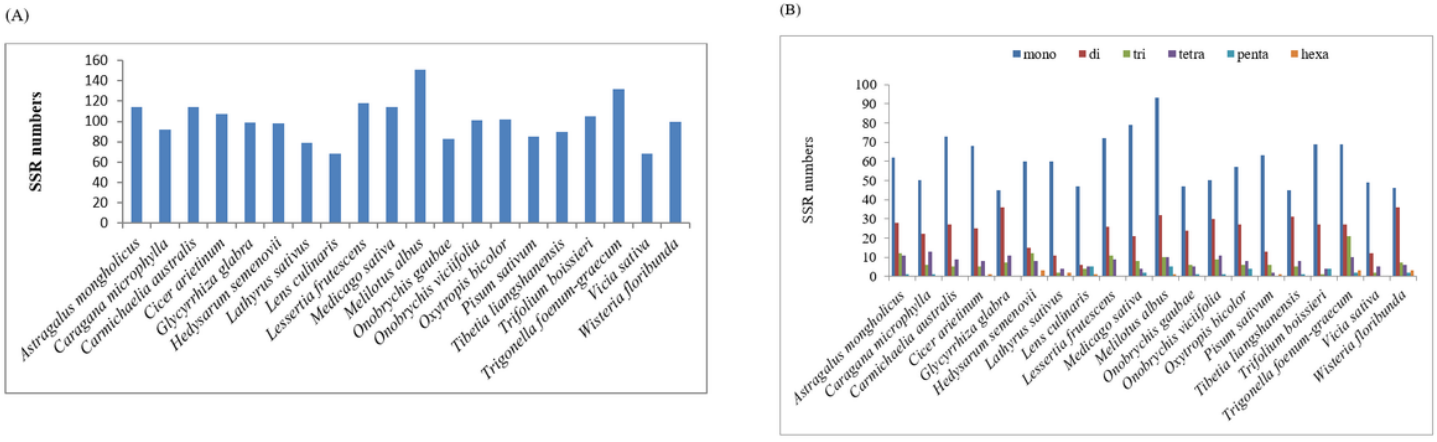

Figure 3

Analysis of perfect simple sequence repeats (SSRs) in the IRLC chloroplast genomes. (A) The number of SSRs detected in the IRLC chloroplast genomes; (B) The number of SSR types detected in the IRLC chloroplast genomes. In the mononucleotide repeats, A/T motifs were the most abundant but no G/C motif was detected in the cp genome. Likewise, the majority of the dinucleotides and trinucleotides were found to be particularly rich in AT sequences. Therefore, the AT richness in the SSRs of the chloroplast genome of O. gaubae is consistent with the results of previous studies28,32 which have shown that in the cp genome, SSRs generally composed of polythymine (poly T) or polyadenine (poly A) repeats33.

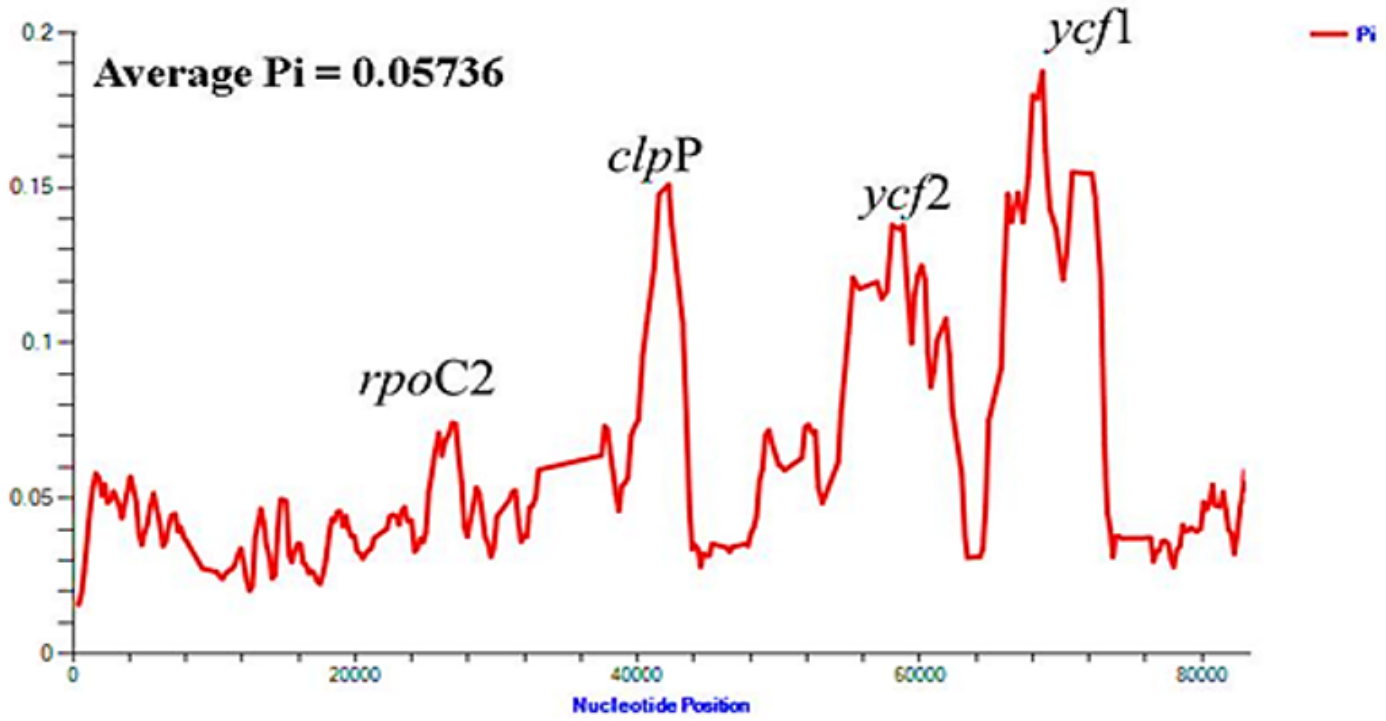

Figure 4

Nucleotide variability (%) values among the IRLC species (using for coding regions). Window length: 800 bp; step size: 200 bp. X-axis: Position of the midpoint of a window. Y-axis: Nucleotide diversity of each window.

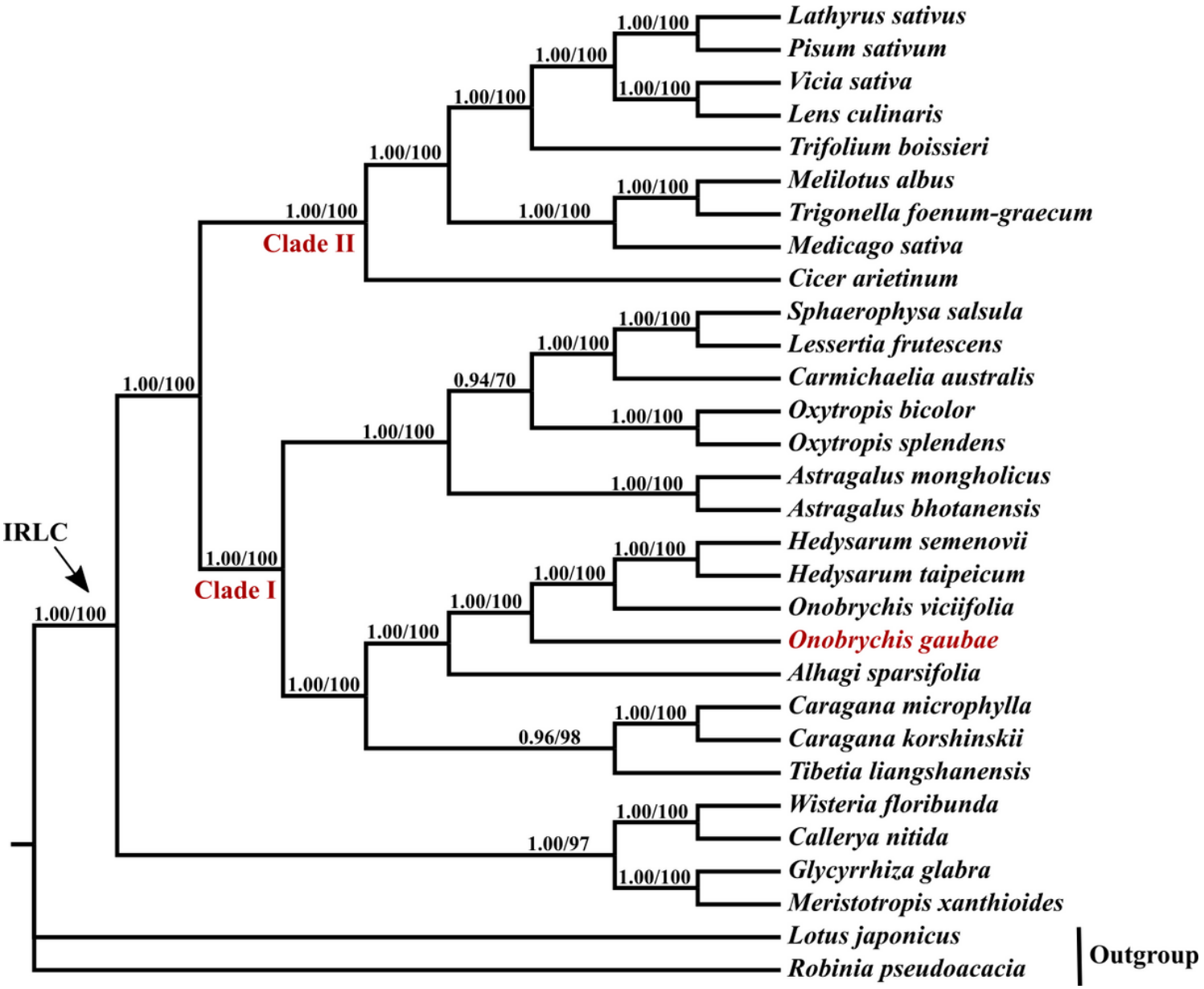

Figure 5

Fifty percent majority rule consensus tree resulting from Bayesian analysis of the 75 plastid genes of IRLC. The position of Onobrychis gaubae is shown in red. Numbers above branches are posterior probability and likelihood values, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- AdditionalFile2.xlsx
- AdditionalFile3.docx

- AdditionalFile4.tiff
- AdditionalFile5.pdf
- AdditionalFile6.docx
- AdditionalFile7.docx
- AdditionalFile8.docx