Title: 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

Kyoto University

Ryohei Terauchi

Kyoto University

Shahrokh Kazempour-Osaloo (✉ skosaloo@modares.ac.ir)

Tarbiat Modares University

Research Article

Keywords: *Onobrychis gaubae*, genomes, population genetics

Posted Date: March 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-290026/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Plastid genome sequences provide valuable markers for surveying the evolutionary relationships and population genetics of plant species. In the present study, the complete plastid genome of *Onobrychis 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 123,645 bp in length and included a large single-copy (LSC) region of 81,034 bp, a small single-copy (SSC) region of 13,788 bp and one copy of the inverted repeat (IR_b) of 28,823 bp. The genome encoded 110 genes, including 76 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosome RNA (rRNA) genes and possessed 89 simple sequence repeats (SSRs) and 28 repeated structures with the highest proportion in the LSC. Comparative analysis of the chloroplast genomes across IRLC revealed three hotspot genes (*ycf1*, *ycf2*, *clpP*) which could be used as molecular markers for resolving phylogenetic relationships and species identification. IRLC plastid genomes also showed multiple gene losses and inversions. 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.

Introduction

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 encode 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 Asteraceae⁷, Campanulaceae⁸, Onagraceae⁹, Fabaceae¹⁰ and Geraniaceae¹¹.

Fabaceae (legumes) is the third largest family of angiosperms which shows the most extensive structural variation¹². Currently accepted classification of the legumes based on plastid gene *matK* includes six subfamilies: Caesalpinioideae, Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, and Papilionoideae¹³. 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 genome⁵. In this subfamily, a loss of one of the IR¹⁴, the presence of many repetitive sequences¹⁵ 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, Millettoids, 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 Hedysareae²⁰ with nine genera, the plastomes of some *Hedysarum* species and only one species of *Onobrychis* have been reported. In the present study, the complete plastome of *O. gaubae* Bornm. belonging to Hedysareae was sequenced (GenBank accession number: ???). *Onobrychis* has more than 130 species and is the second largest genus after *Hedysarum* and mainly found throughout temperate and subtropical regions of Eurasia, N and NE Africa²¹. *Onobrychis gaubae* 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 And Discussion

General features of the *O. gaubae* plastid genome

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}. The plastid genome of *O. gaubae* with 123,645 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 *rps16* and *rp122* genes and intron 1 of *clpP* 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. A total of 16 genes contained a single intron, whereas *ycf3* exhibits two introns (Supplementary Table S1). The *rps12* gene is a trans-splicing gene which does not have introns in the 3'-end. The *trnK*-UUU has the largest intron encompassing the *matK* gene, with 2,495 bp, whereas the intron of *trnL*-UAA is the smallest (542 bp). The overall GC content of the *O. gaubae* chloroplast genome sequence was 35.1%, which is consistent with other IRLC species, whose plastomes have GC-contents ranging from 33.6–35.1% (Table 2). Different GC content occurs in the LSC (34.4%), SSC (30.8%) and IR (39.2%) regions (Supplementary Table S2). Higher GC content was usually detected in the IRs compared to the other regions of plastome, which is mainly due to the presence of rRNA genes (*rrn23*, *rrn16*, *rrn5*, *rrn4.5*) with high GC content (50%-56.4%) in IRs^{6,27}.

The GC content of the protein-coding regions of *O. gaubae* chloroplast genome comprised 36.02%. Within these regions, the GC contents for the first, second and third positions of the codons were 42.4%, 36.1% and 29.5%, respectively.

Table 1

Genes predicted in the chloroplast genome of *O. gaubae*. The number of asterisks after the gene names indicates the number of introns contained in the genes.

Category of genes	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	<i>rp14, rp16*, rp12*, rp120, rp123, rp132, rp133, rp136</i>
	Small subunit of ribosomal proteins	<i>rps2, rps3, rps4, rps7, rps8, rps11, rps12*, rps14, rps15, rps18, rps19</i>
	DNA-dependent RNA polymerase	<i>rpoA, rpoB, rpoC1*, rpoC2</i>
	Ribosomal RNA genes	<i>rrn16S, rrn23S, rrn 4.5S, rrn 5S</i>
	Transfer RNA genes	30 <i>trn</i> genes (5 contain an intron)
Genes for photosynthesis	Subunits of NADH-dehydrogenase	<i>ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
	Subunits of photosystem I	<i>psaA, psaB, psaC, psal, psaJ</i>
	Subunits of photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbl, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
	Subunits of cytochrome b/f complex	<i>petA, petB*, petD*, petG, petL, petN</i>
	Subunits of ATP synthase	<i>atpA, atpB, atpE, atpF*, atpH, atpI</i>
	Subunit of rubisco	<i>rbcL</i>
	Other genes	
	Maturase K	<i>matK</i>
	Envelope membrane protein	<i>cemA</i>
	Subunit of Acetyl-CoA-carboxylase	<i>accD</i>
	C-type cytochrome synthesis gene	<i>ccsA</i>
	Protease	<i>clpP*</i>
Genes of unknown function	Conserved hypothetical chloroplast open reading frames	<i>ycf1, ycf2, ycf4, ycf3**</i>

Table 2
Chloroplast genome information from sampled IRLC species and the newly assembled
O. gaubae. LSC: Large Single Copy, SSC: Small Single Copy, IR: Inverted Repeat.

Species	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	GC (%)
<i>Astragalus mongholicus</i>	123,582	80,986	13,773	28,823	34.1%
<i>Caragana microphylla</i>	130,029	85,436	14,106	30,487	34.3%
<i>Carmichaelia australis</i>	122,805	80,588	14,074	28,143	34.3%
<i>Cicer arietinum</i>	125,319	82,583	13,820	28,916	33.9%
<i>Glycyrrhiza glabra</i>	127,943	84,714	14,187	29,042	34.2%
<i>Hedysarum semenovii</i>	123,407	80,288	13,679	29,440	34.9%
<i>Lens culinaris</i>	122,967	81,659	13,833	27,604	34.4%
<i>Lessertia frutescens</i>	122,700	80,698	13,750	28,252	34.2%
<i>Medicago sativa</i>	125,330	83,756	13,383	28,191	34%
<i>Melilotus albus</i>	127,205	84,279	13,806	29,120	33.6%
<i>Meristotropis xanthioides</i>	127,735	84,629	14,150	28,956	34.2%
<i>Onobrychis gaubae</i>	123,645	81,066	13,777	28,802	35.1%
<i>Oxytropis bicolor</i>	122,461	80,170	14,017	28,274	34.2%
<i>Tibetia liangshanensis</i>	123,372	79,916	13,513	29,943	34.7%
<i>Wisteria floribunda</i>	130,561	87,193	14,127	29,628	34.4%

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 arginine showed the least abundance in this species (Supplementary Table S3). 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 *psbL*, GTG for *rps8* and ACG for *ndhD*.

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^{28,29}. The total number of codons among protein-coding genes in the IRLC species varies from 20,381 codons in *Hedysarum taibeicum* (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 (Supplementary 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 (Supplementary 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 plants³⁰.

Analysis of repeats

Repeat analysis of the *O. gaubae* plastome identified 28 repeat structures with lengths ranging from 30 bp to 81 bp. These structures included 12 palindromic repeats with lengths in the range of 33–50 bp, 11 forward repeats of 30–81 bp, four reverse repeats which varied from 30 bp to 31 bp and one complementary repeats with a length of 30 bp (Supplementary Table S5). Among the 28 repeats, 75% are located in the LSC region, 14.28% in the SSC region and 10.71% in the IR region. Also, most of the repeats (60.71%) were found in the intergenic spacer regions (IGS), 25% were distributed in coding region (*psaA*, *psal*, *psbJ*, *trnR-UCU*, *trnS-UGA*) and 14.28% were located in the introns (*ndhA* and *ycf3*). 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 taibeicum*, followed by *Vicia sativa* of 517 bp and *Caragana microphylla* of 455 bp, which were much longer than other species studied. 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 genome⁶; therefore, these repeats can provide valuable information for phylogenetic and population studies²⁹.

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^{29,31}. 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^{29,32}. A total of 89 SSRs were detected in the *O. gaubae* plastome, which were composed by a length of at least 10 bp. Among them, 53 (59.55%) were mono-repeats, 21 (23.55%) were di-repeats, 10 (11.23%) were tri-repeats and five (5.61%) were tetra-repeats. No penta- and hexanucleotide SSRs existed in *O. gaubae* genome (Supplementary Table S6). These SSR loci were located primarily in the LSC region (68.53%), followed by the SSC region (16.85%) and IR (14.6%)

(Fig. 3A). 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^{27,29} which have shown that in the cp genome, SSRs generally composed of polythymine (poly T) or polyadenine (poly A) repeats³³. The number of SSRs in the cp genomes (cpSSRs) ranged from 68 (*Vicia sativa* and *Lens culinaris*) to 151 (*Melilotus albus*) across the IRLC species (Fig. 3B). The mononucleotide repeats (P1) were identified at a much higher frequency, which varied from 45 (*Tibetia liangshanensis*, *Glycyrrhiza glabra*) to 93 (*Melilotus albus*). In all cases, the P1s were AT-rich (Fig. 3C). Strong AT bias in SSR loci was also observed in other legumes such as *Vigna radiate*³⁴, *Arachis hypogaea*³⁵ and *Stryphnodendron adstringens*²⁷ which, like other plastomes of species, may contribute to the bias in base composition⁶. 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^{31,33}.

Divergent hotspots in the IRLC chloroplast genomes

The average nucleotide diversity (π) among the protein-coding genes of 19 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 π values > 0.1 among the IRLC species; *ycf1* and *ycf2* from IR region and *c/pP* 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, *ycf1* encoding a protein of 1800 amino acids has the highest nucleotide diversity (0.18745). *ycf1* gene is more variable than *matK* gene and it can be useful for molecular systematics at low taxonomic levels^{36,37}. Numerous studies^{14,18,38} analyzed the phylogenetic reconstructions of IRLC species at various taxonomic levels based on different fragments of plastid coding genes such as *matK*, *ndhF* and *rbcl*, 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.

The non-synonymous (K_a) to synonymous (K_s) rate ratio (K_a/K_s) was estimated for 75 protein-coding genes across the 28 IRLC species analyzed (Supplementary Table S7). In general, the K_a/K_s values were lower than 0.5 for almost all genes. The *ycf4* gene which is involved in regulating the assembly of the photosystem I complex had the highest nonsynonymous rate, 0.165691, while the *ycf1* gene with unknown functions had the highest synonymous rate, 0.181067. The K_a/K_s 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 mutation³⁹. In present study, the K_a/K_s ratio was calculated to be 0 for *psbL* 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 *accD* gene which encodes a subunit of the acetyl-CoA carboxylase (ACCase) enzyme. 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 *ycf4* gene^{15,16}.

Legumes chloroplast genome, and in particular IRLC, have regions with high mutation rates, including *rps16-accD-psal-ycf4-cemA* region. *rps16* gene was lost from cpDNA in the common ancestor of the IR-loss clade¹⁵. *accD* coding region was completely absent in the *Trifolium* subgenus *Trifolium* and has nuclear copies in *Medicago truncatula* and *Cicer arietinum*²⁵. Three consecutive genes *psal-ycf4-cemA* is situated in a local mutation hotspot and has been lost in some species of *Lathyrus*^{15,16}.

Comparative analysis of genome structure

We compared whole chloroplast genome sequences of different taxa of IRLC to analyze gene order and content. We found that, similar to other plant species, the gene coding regions were more conserved than the noncoding regions (Fig. 5). High nucleotide variations were observed across IRLC for the protein-coding regions *ycf1*, *ycf2* and *clpP*. Similar results were also obtained from calculation of nucleotide diversity (Pi). 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 *rps12* and *clpP* genes²⁴, absence of *rps16* gene²⁶ and transfer/loss of *rp122* to the nucleus²⁴. Numerous studies have also shown some other rearrangements in some IRLC taxa, such as loss of *accD* gene in six species of *Trifolium*^{10,25}, loss of *rp123* and *rp133* genes in some species of *Lathyrus*, *Pisum* and *Vicia*³² and loss of *ycf4* 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 sequences⁴⁰. For example, the loss of the *rps16* gene was probably due to the presence of a nuclear *rps16* copy, which contributed to pseudogenization of the plastid copy⁴¹. Likewise, the lack or expansion of the *accD* gene was explained by the presence of tandemly repeated sequences^{6,15}.

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 transcript⁴². RNA editing sites of *O. gaubae* plastid genes were predicted using Prep-CP prediction tool (Supplementary Table S8). In total, 58 editing sites were present in 19 chloroplast protein-coding genes and all of the editing sites were C-to-U conversions (Supplementary Table S8). Among them, nine editing sites, the highest number, were found

in the region encoding *ndhB* gene followed by seven editing sites in *petB* (Fig. 6). There were six editing sites detected each in *ndhA* and *rpoB* genes. *accD*, *ndhG* and *petD* had three editing sites, and *ndhD* and *ndhF* had two editing sites. Two editing sites were also found in *ccsA*, *matK* and *rpoC1* 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. In flowering plants, the highest number of plastid editing sites was found in the *ndh* group genes⁴². 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^{43–45}. The RNA editing is probably important for the NDH protein complex function and may also lead to improved photosynthesis and display positive selection during evolution⁴².

Phylogenetic relationships

Phylogenetic relationships within the IRLC were reconstructed using the representative taxa (27 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 reconstructed phylogeny is in agreement with previous studies^{5,14,25,46} 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^{19,47}. Then, there are two major clades: clade I and II (Fig. 7). Clade I comprises tribes Caraganeae¹⁷, Hedysareae²⁰ and Coluteae¹⁸ 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 study⁴⁶, 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. 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).

Materials And Methods

Chloroplast DNA extraction and sequencing

The young leaves of *O. gaubae* were collected in the natural habitat in northwestern Tehran, Iran. 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 *Astragalus nakaianus* (KR296789) 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>)⁴⁸. Gaps in the cpDNA sequences were filled by PCR amplification and Sanger sequencing. The de novo assembled chloroplast genomes were annotated by GeSeq⁴⁹. We used the online tRNAscan-SE service⁵⁰ 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 Bowtie2⁵¹ 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 X⁵², 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 (π) among the plastid genomes of the representative species of the IRLC, the whole chloroplast genome sequences were aligned using MAFFT⁵³ on XSEDE v.7.402 in CIPRES Science Gateway⁵⁴. A sliding window analysis was conducted to determine the nucleotide diversity of the chloroplast genome using DnaSP v.6.12 software⁵⁵. 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 (K_s) and non-synonymous (K_a) substitution rates, as well as their ratio (K_a/K_s) 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 mode⁵⁶ among the 17 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>)⁵⁷ with a cutoff value of 0.8.

Phylogenetic reconstruction

Seventy-five protein-coding genes were recorded from 27 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 (Supplementary Table S9). 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 MrModeltest2.2⁵⁸ under the Akaike Information Criteria (AIC)⁵⁹. Maximum likelihood analyses were performed using the online phylogenetic software W-IQ-TREE⁶⁰ 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 CIPRES⁵⁴ 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.

Conclusions

In this study, the complete plastome sequence of *O. gaubae* (123,645 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 (*ycf1*, *ycf2*, *clpP*) which provided sufficient genetic information for species identification and phylogenetic reconstruction of the IRLC 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.

Declarations

Competing Interests:

The authors declare no competing interests.

Author 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.

Data Availability

The complete chloroplast sequence generated and analyzed during the current study are available in GenBank (accession number are described in the text).

References

1. Jansen, R. K., Ruhlman, T. A. Plastid Genomes of Seed Plants. *In Genomics of Chloroplasts and Mitochondria* 103–126, https://doi.org/10.1007/978-94-007-2920-9_5 (Springer, Dordrecht, 2012).
2. Ruhlman, T. A., and Jansen, R. K. The plastid genomes of flowering plants in *Chloroplast Biotechnology: Methods and Protocols* (ed. Maliga P.) 3–38 (Springer, Humana Press, 2014).
3. Jansen, R. K. et al. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395,348–384 (2005).
4. Bock, R. Structure, function, and inheritance of plastid genomes in *Cell and Molecular Biology of Plastids* (ed. Bock, R.) 29–63, https://doi.org/10.1007/4735_2007_0223 (Springer-Verlag Berlin Heidelberg, 2007).
5. Schwarz, E.N. *et al.* Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids. *J Syst Evol.* 53, 458–468 (2015).
6. Asaf, S. *et al.* Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. *PLoS ONE* 12(8): e0182281 (2017).
7. Kim, K.J., Choi, K.S., Jansen, R. K. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Mol Biol Evol.* 22(9): 1783±92 (2005).
8. Haberle, R.C., Fourcade, H.M., Boore, J.L., Jansen, R.K. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J. Mol. Evol.* 66(4): 350±61 (2008).
9. Greiner, S. *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.* 36(7): 2366±78 (2008).
10. Cai, Z. *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.* 67(6): 696±704 (2008).
11. Guisinger, M.M., Kuehl, J.V., Boore, J.L., Jansen, R.K. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol. Biol. Evol.* 28(1): 583±600 (2011).

12. Palmer, J.D., Osorio, B., Thompson, W.F. Evolutionary significance of inversions in legume chloroplast DNAs. *Curr. Genet.* 14, 65–74 (1988).
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.* 66,44–77 (2017).
14. Wojciechowski, M.F., Lavin, M., Sanderson, M.J. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *Am. J. Bot.* 91, 1846–1862 (2004).
15. Magee, A.M. *et al.* Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research.* 20, 1700–1710 (2010).
16. Moghaddam M., Kazempour-Osaloo S. Extensive survey of the *ycf4* 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 Fabaeae. *PLoS ONE* 15(3), e0229846 (2020).
17. Duan, L. *et al.* A molecular phylogeny of Caraganeae (Leguminosae, Papilionoideae) reveals insights in to new generic and infrageneric delimitations. *PhytoKeys.* 70, 111–137 (2016).
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, *matK* and *rp/32-trnL(UAG)* sequences data. *Plant Biosyst.* 6, 1082–1093 (2017).
19. Compton, J.A. *et al.* The *Callerya* Group redefined and Tribe Wisterieae (Fabaceae) emended based on morphology and data from nuclear and chloroplast DNA sequences. *PhytoKeys.* 125, 1-112 (2019).
20. Amirahmadi, A., Kazempour Osaloo, S., Moein, F., Kaveh, A., Maassoumi, A.A. Molecular systematic of the tribe Hedysareae (Fabaceae) based on nrDNA ITS and plastid *trnL-F* and *matK* sequences. *Plant Syst. Evol.* 300, 729–747 (2014).
21. Lock, J.M. Tribe Hedysarae in *Legumes of the world*. (eds. Lewis G, Schrire B, Mackinder B, Lock M) 489–495 (Royal Botanical Gardens, Kew 2005).
22. Rechinger KH. Tribus Hedysareae *Papilionaceae II, Flora Iranica*. (ed. Rechinger KH) 387–464 (Akademische Druckund Verlagsanstalt, Graz 1984).
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.* 305, 33–48 (2019).
24. Jansen, R.K., Wojciechowski, M.F., Sanniyasi, E., Lee, S.B., Daniell, H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* 48, 1204–1217 (2008).
25. Sabir, J. *et al.* Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol. J.* 12, 743–754 (2014).

26. Doyle, J. J., Doyle, J. L., Palmer, J. D. Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Syst. Bot.* 20, 272–294 (1995).
27. Souza, U.J.B.d. *et al.* The complete chloroplast genome of *Stryphnodendron adstringens* (Leguminosae - Caesalpinioideae): comparative analysis with related Mimosoid species. *Sci. Rep.* 9, 14206 (2019).
28. Sharp, P.M., Li, W.H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3), 1281–1295 (1987).
29. Li, X. *et al.* Comparison of Four Complete Chloroplast Genomes of Medicinal and Ornamental *Meconopsis* Species: Genome Organization and Species Discrimination. *Sci. Rep.* 9, 10567 (2019).
30. Li, C.J., Wang, R.N., Li, D.Z. Comparative analysis of plastid genomes within the Campanulaceae and phylogenetic implications. *PLoS ONE* 15(5), e0233167 (2020).
31. Powell, W., Morgante, M., Mcdevitt, R., Vendramin, G.G., Rafalski, J.A. Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes-Applications to the Population-Genetics of Pines. *PNAS.* 92(17), 7759±63 (1995).
32. Lei, W. *et al.* Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Sci. Rep.* 22, 6:21669 (2016).
33. Zong, D. *et al.* Comparative analysis of the complete chloroplast genomes of seven *Populus* species: Insights into alternative female parents of *Populus tomentosa*. *PLoS ONE* 14(6), e0218455 (2019).
34. Tangphatsornruang, S. *et al.* The Chloroplast Genome Sequence of Mungbean (*Vigna radiata*) Determined by High-throughput Pyrosequencing: Structural Organization and Phylogenetic Relationships. *DNA Res.* 17(1), 11±22 (2010).
35. Yin, D. *et al.* Development of chloroplast genome resources for peanut (*Arachis hypogaea* L.) and other species of *Arachis*. *Sci. Rep.* 7, 11649 (2017).
36. Neubig, K. M. *et al.* Phylogenetic utility of *ycf1* in orchids: a plastid gene more variable than *matK*. *Plant Syst. Evol.* 277, 75–84 (2009).
37. Dong, W. *et al.* *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5, 8348 (2015).
38. Schaefer, H. *et al.* Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-atlantic island lineages. *BMC Evol. Biol.* 12, 250 (2012).
39. Yang, Z., Wong, W.S.W., Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22, 1107–1118 (2005).
40. Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Muller, K.F., Quandt, D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297 (2011).
41. Keller, J. *et al.* The evolutionary fate of the chloroplast and nuclear *rps16* genes as revealed through the sequencing and comparative analyses of our novel legume chloroplast genomes from *Lupinus*. *DNA Res.* 24(4), 343-358 (2017).

42. He, P. *et al.* Abundant RNA editing sites of chloroplast protein-coding genes in *Ginkgo biloba* and an evolutionary pattern analysis. *BMC Plant Biol.* 16, 257 (2016).
43. Blazier, J., Guisinger, M. M., Jansen, R. K. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Mol. Biol.* 76, 263 – 272 (2011).
44. Ruhlman, T. *et al.* NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biol.* 15, 100 (2015).
45. Sanderson, M.J. *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.* 102(7), 1115-27 (2015).
46. Su, C. *et al.* 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.* 156, 107025. (2021).
47. Duan, L. *et al.* Chloroplast Phylogenomics Reveals the Intercontinental Biogeographic History of the Liquorice Genus (Leguminosae: *Glycyrrhiza*). *Front. Plant Sci.* 11, 793 (2020).
48. McKain, M.R., Wilson, M. mrmckain/Fast-Plast: Fast-Plast v.1.2.8. Version v.1.2.8. (2018).
<https://github.com/mrmckain/Fast-Plast>.
49. Tillich, M. *et al.* GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6-W11 (2017).
50. Schattner, P., Brooks, A.N., Lowe, T.M. The tRNA scan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686±W9 (2005).
51. Langmead, B. Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359 (2012).
52. Kumar, S. *et al.* MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35(6), 1547–1549 (2018).
53. Katoh, K., Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780 (2013).
54. Miller, M.A., 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).
55. Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34, 3299–3302 (2017).
56. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279 (2004).
57. Mower, J.P. The PREP suite: Predictive rna editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 37, W253–W259 (2009).
58. Nylander, J.A.A. MrModeltest v2. Program distributed by the author. Uppsala: Evolutionary Biology Centre, Uppsala University. (2004).

- ## Figures

Figure 1

Gene map of the *O. gaubae* 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 frequency of identified SSRs in large single-copy (LSC), inverted repeat (IR,) and small single-copy (SSC) regions; (b) The number of SSRs detected in the IRLC chloroplast genomes; (c) The number of SSR types detected in the IRLC chloroplast genomes.

Figure 4

Nucleotide variability (%) values among the IRLC species (using for coding regions).

Figure 5

Sequence identity plot comparing the IRLC chloroplast genomes with *Glycyrrhiza glabra* as a reference by using mVista. Pink bars represent noncoding sequences (CNS), and white peaks represent genomic differences. The y-axis represents the percentage identity (shown: 50–100%).

Figure 6

Distribution of number of RNA editing sites in the protein coding genes of the *O. gaubae*.

Figure 7

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.

- [TableS4.xlsx](#)
- [Supplementaryinformation.pdf](#)