PLOS ONE

RESEARCH ARTICLE

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

Mahtab Moghaddam[☯], **Shahrokh Kazempour-Osaloo**[*☯]

Department of Plant Biology, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran

☯ These authors contributed equally to this work.
* skosaloo@modares.ac.ir

Citation: Moghaddam M, Kazempour-Osaloo S (2020) 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 15(3): e0229846. https://doi.org/10.1371/journal.pone.0229846

Editor: Shilin Chen, Chinese Academy of Medical Sciences and Peking Union Medical College, CHINA

Received: October 5, 2019

Accepted: February 15, 2020

Published: March 5, 2020

Data Availability Statement: The data underlying the results presented in the study are available from GenBank. All taxa together with their origin, voucher information and GenBank accession numbers are listed in S1 Table.

Funding: The author(s) received no specific funding for this work.

Abstract

The genome organization and gene content of plastome (plastid genome) are highly conserved among most flowering plant species. Plastome variation (in size and gene order) is rare in photosynthetic species but size variation, rearrangements and gene/intron losses is attributed to groups of seed plants. Fabaceae (legume family), in particular the subfamily Papilionoideae and the inverted repeat lacking clade (IRLC), a largest legume lineage, display the most dramatic and structural change which providing an excellent model for understanding of mechanisms of genomic evolution. The IRLC comprises 52 genera and ca 4000 species divided into seven tribes. In present study, we have sampled several representatives from each tribe across the IRLC from various herbaria and field. The *ycf*4 gene, which plays a role in regulating and assembly of photosystem I, is more variable in the tribe Fabeae than in other tribes. In certain species of *Lathyrus*, *Pisum* and *Vavilovia*, all belonging to Fabeae, the gene is either absent or a pseudogene. Our study suggests that *ycf*4 gene has undergone positive selection. Furthermore, the rapid evolution of the gene is locus and lineage specific and is not a shared character of the IRLC in legumes.

Introduction

Previous gene mapping and genomic sequencing has demonstrated that variation in plastid genome (plastome) size, gene order and gene/intron content is relatively rare. Plastome organization in most of angiosperms is identical which comprises two large (~ 25 kb) copies of inverted repeat (IR) regions separated by two single copy regions, the large (~ 85 kb) single copy region (LSC) and the small (~ 15 kb) single copy region (SSC) [1–3]. Due to the conserved nature of plastid genome, variation at the nucleotide and structural level can provide

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

powerful phylogenetic markers and allow a comparative evaluation of genomic evolutionary history [3, 22]. There are several unrelated angiosperm lineages that have been identified with different structural rearrangements such as inversion, transposition, gene duplication, gene/intron loss, insertion/deletion and IR expansion/contraction [3–4]. Variation in plastome size is often related to IR expansion/contraction [4–6]. Expansion in IR size have been documented in *Pelargonium* x *hortorum* (217 942 bp) from the Geraniaceae which is Known as the largest seed plant plastome size [5].Cases of IR contraction can be found in Geraniaceae [7] and Pinaceae [8]. Some angiosperm families like Orobanchaceae and Fabaceae show IR loss in their plastomes [9–11].

Fabaceae (legumes) are the third largest family of angiosperms which have experienced a remarkable number of plastome rearrangements [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 order and gene/intron content in plastomes of all subfamilies except Papilionoideae are highly conserved and similar to the ancestral angiosperm genome organization [6]. Papilionoideae, the largest group of legumes due to its ecological and economical importance, exhibit numerous rearrangements and gene/intron losses and have smaller genome [6, 14]. The remarkable loss of the one of the plastid IR in the inverted repeat lacking clade (IRLC), a largest legume lineage, is an example of genome variation in papilionoids [14–15]. This clade comprises 52 genera (e.g., *Wisteria*, *Glycyrrhiza*, *Astragalus*, *Colutea*, *Trifolium*, *Lathyrus*, . . .) and ca 4000 species divided into seven tribes [14, 16–18]. The existence of a diversity of changes in the organization and gene content of IRLC plastome have made it as an excellent model for genome evolution studies. To date, different rearrangements have been reported in various tribes of the IRLC. For example the introns of *rps*12 and *clp*P have been lost in most members of the IRLC [19–22] and in tribe Trifolieae, *acc*D gene has been transferred to the nucleus in some *Trifolium* species [22–24].

Within the IRLC, the plastid genome of tribe Fabeae (including five main genera: *Lathyrus* L., *Lens* Mill., *Vicia* L., *Pisum* L. and *Vavilovia* Fed.) exhibits variation in size and expansion/contraction which have occurred during its evolution, due to the presence of repetitive DNA [25–27]. Also, *ycf*4 gene in *Pisum sativum* and *Lathyrus odoratus*, is either absent or pseudogene [21].

The *ycf*4 gene (orf184) is located in LSC region and its product is a thylakoid protein which is involved in regulating the assembly of the photosystem I complex [4, 28–30]. It is a part of a gene cluster, including *psa*I and *acc*D genes at the upstream and *cem*A gene at the downstream of it, which is considered as local mutation hotspot in plastid genome [21]. This genomic region has undergone numerous rearrangements not only in the IRLC legumes (as mentioned above) but also in different lineages. For example; within *Jasminum* and *Menodora*, both of them are belonging Oleaceae, *acc*D have been lost and the *ycf*4-*psa*I region in *Jasminum* section *Primulina* was relocated [31], the *acc*D was found as a pseudogene in some species of *Primula* (Primulaceae) [32] and chloroplast genome of *Tylosema esculentum* as one of the basal legumes in the Caesalpinioideae, has an unique inverted region which is including six genes *rbc*L, *acc*D, *psa*I, *ycf*4, *cem*A and *pet*A [33].

Unusual evolution of *ycf*4 and genomic region around of it in legumes, especially in the IRLC, as well as relatively small number of IRLC samples in previous studies [6, 21–22], prompted us to evaluate the evolutionary history of *ycf*4 gene across the IRLC with special reference to the tribe Fabeae. Furthermore, in this study two other widely sequenced plastid genes, *mat*K and *rpl*32, were used for the comparison of nonsynonymous and synonymous nucleotide substitution rates. In this paper, we examined the following points: 1) to determine the presence or absence of *ycf*4 gene across the IRLC, 2) to assess phylogenetic utility of this

gene at the tribal to generic and species levels across the IRLC, and 3) to investigate the evolutionary rate of *ycf*4 in the IRLC and Fabeae with comparing nonsynonymous and synonymous nucleotide substitution rates.

Materials and methods

Taxon sampling

Sampling was designed to include different species from across the IRLC [14]. The plastid *ycf*4 gene was obtained from representatives of IRLC genera not sampled previously and combined with a large number of sequences already in Genbank plus appropriate outgroups (*Lotus japonicus* and *Robinia pseudoacacia*) in Loteae and Robineae. In addition to the newly generated *ycf*4 (60), *mat*K (27) and *rpl*32 (24) gene sequences, 62 *ycf*4, 80 *mat*K and 72 *rpl*32 sequences were retrieved from GenBank. All taxa together with their origin, voucher information and GenBank accession numbers are listed in S1 Table.

DNA extraction, amplification and sequencing

Total genomic DNA was isolated from leaf materials using the modified CTAB method of Doyle and Doyle [34]. The *ycf*4 gene was amplified using the primer of *acc*D and *cem*A [21] and also *Psa*I and *cem*A (designed in this study). The information (location and base composition) of each of the primers used in this study and a schema of the *acc*D-*cem*A regions with positions of forward and reverse primers are shown in S2 Table and S1 Fig, respectively.

The PCR amplifcation was carried out in the volume of 20 μl, containing 8 μl deionized water, 10 μl of the 2 × Taq DNA polymerase master mix Red (Amplicon) 0.5 μl of each primer (10 pmol/μl), and 1 μl of template DNA.PCR procedures for *ycf*4 region were 2 min at 94˚C for predenaturation followed by 38 cycles of 1 min at 94˚C for denaturation, 3 min 20 s at 58˚C (when using *acc*D and *cem*A primers) and 50 s at 55˚C (when using *Psa*I and *cem*A primers) for primer annealing and 50 s at 72˚C for primer extension, followed by a final primer extension of 5 min at 72˚C. The ensuring PCR fragments were separated by electrophoresis in 1% agarose gels in 1 × TAE (pH = 8) buffer, stained with ethidium bromide and were photographed with a UV gel documentation system (UVItec, Cambridge, UK). PCR products along with the primers used for amplifcation were sent for Sanger sequencing at Macrogen (Seoul, South Korea).

Sequence alignment

*ycf*4 gene data set was aligned using the web-based version of MUSCLE [35] under default parameters followed by manual adjustment. Indels were treated as missing data in all phylogenetic analyses. Highly variable sites were excluded to construct datasets. Removing regions of the alignments that have a gap (some species in which the *ycf*4 is pseudogene and lacks start and stop codons) before performing PAML analysis is common. It is often done with the hope that the remaining sequence has a better quality alignment and, thus, the results are more reliable. In order to determine the coding range of *ycf*4 gene in some species, open reading frame (ORF) Finder (https://www.ncbi.nlm.nih.gov/orffinder/) was used. To calculate dN/dS, we will require codon-based alignments of the DNA sequences of all genes in each ortholog group; therefore gaps should be positioned so as not to change the reading frame. Among the aligners, we have used PRANK [36, 37], which is unique in that it takes evolutionary information into consideration during alignment to infer positive selection acting on genes and codons.

Phylogenetic analysis

Bayesian phylogenetic analysis was conducted at the CIPRES Science Gateway V. 3.3 [38] using MrBayes version 3.2.6 [39] with default priors (uniform priors) and the best-fit model of sequence evolution for dataset. The Markov chain Monte Carlo algorithm was run in two separated analyses, each run with four simultaneous Markov chains (one cold and three heated with a heating parameter of 0.2) for 10 million generations, trees sampled at every 100 generations. The first 25,000 trees (25%) were discarded as a conservation burn-in, and the remaining trees were used to construct the 50% majority rule consensus tree with posterior probability values (PP). Stationary of the chains was ascertained using Tracer v.1.6 [40]. Tree visualization was carried out using Dendroscope v.2.7.4 [41].

In Bayesian analysis, the best model of molecular evolution of nucleotide substitutions was evaluated using MrModeltest 2.2 [42], based on the Akaike Information Criteria (AIC) [43]. On the basis of model test results, the GTR+G were identified as the best model for *ycf*4 gene.

Selective pressure analysis

Nonsynonymous (dN) and synonymous (dS) substitution rates and the ratio (ω = dN/dS) between them which provide information about the evolutionary forces operating on a gene, in the codon-based nucleotide alignments of 3 chloroplast genes (*ycf*4, *mat*K, *rpl*32) in the IRLC and tribe Fabeae were estimated using yn00 from the PAML package [44] and JCoDA package [45] across branches. Furthermore, the inference of selection was performed using the branch-site models of CODEML algorithm [44] implemented in EasyCodeML [46]. In general, when positive selection dominates, ω is greater than 1, natural selection is acting to promote nonsynonymous substitutions and fixation of advantageous mutations and adaptive evolution would be inferred when dN>dS and when negative selection (also called purifying selection) dominates, ω is less than 1, natural selection suppresses protein changes and acting against deleterious nonsynonymous substitutions. In genetic regions under strong negative selection, mutations are quickly removed from the gene pool, resulting in highly conserved stretches of the genome [47–49]. Since dN and dS are usually estimated using complete coding sequences and we do not expect an entire gene to evolve under positive selection, it is very rare to see dN>dS and large structural rearrangement has seldom been observed in chloroplast genes. Under selective neutrality, ω is close to 1, the positive and negative selection forces balance each other and synonymous and nonsynonymous substitution rates should be equal. In other words, coding sequences of gene evolving under no influence of selection.

Results

Sequence considerations

The aligned *ycf*4 dataset was 1128 nucleotide sites long in the IRLC (S1 File), of which 846 sites were potentially parsimony informative. The 50% majority rule consensus tree resulting from the Bayesian analysis of the *ycf*4 dataset for IR loss clade with posterior probabilities is shown in Fig 1. The monophyly of the IRLC and all its tribes is in accordance with all previous studies [14, 50]. The IRLC comprises all members of several well supported tribes/lineages including Cicereae, Hedysareae, Caraganeae, Trifolieae, Fabeae, Astragalean clade (including tribe Coluteae and genera *Astragalus*, *Oxytropis* and *Erophaca*) and *Galega* (Galegeae) as well as Wisterieae together with *Adinobotrys* and *Glycyrrhiza* [14, 16–18].

We sequenced the region flanking the *ycf*4 locus for different samples of different tribes in the IRLC and compared it to the available data for other legumes. The length of *ycf*4 varies from 564–567 bp in *Astragalus*, *Oxytropis* and tribe Hedysareae (as the smaller length of

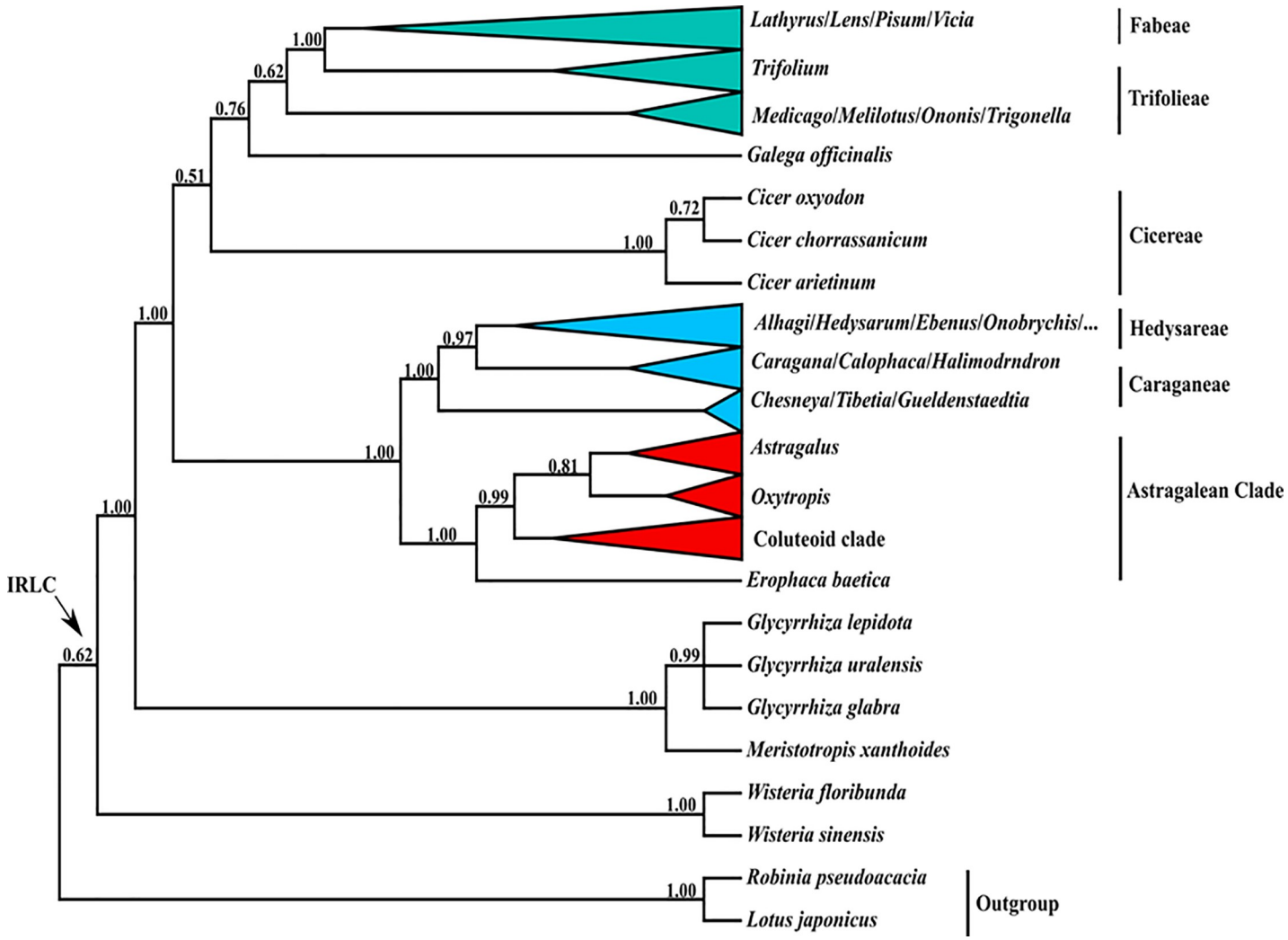

Fig 1. Fifty percent majority rule consensus tree resulting from Bayesian analysis of the *ycf*4 gene dataset. Numbers above branches are posterior probability values. Values <50% were not shown.

https://doi.org/10.1371/journal.pone.0229846.g001

*ycf*4 in the IRLC) to 630 bp in tribe Trifolieae. Gene length in various genera from different tribes of the IRLC does not show much difference except tribe Fabeae. In other words, in all tribes of the IRLC, except tribe Fabeae, *ycf*4 and its neighbors are conserved either in length or point mutations. *ycf*4 region shows extensive length variation among the Fabeae species that retain it. The aligned *ycf*4 dataset was 1085 nucleotide sites in Fabeae. As a result, the length of *ycf*4 in *Vicia* and *Lens* is 615 and 606 bp, respectively and in different species of *Lathyrus*, *Pisum* and *Vavilovia* is highly variable and in some cases is lost or pseudogene. Accordingly, they are divided into seven groups (A-G, Fig 2). In groups B, C, D, and E at least one of the two *psa*I and *ycf*4 genes was lost. In the present study, *ycf*4 gene in some species of *Lathyrus* (*L. chloranthus*, *L. hirsutus*, *L. pratensis*, *L. odoratus* and *L. annuus*), *Pisum* (*P. sativum* and *P. fulvum*) and *V. formosa* show signs of pseudogenization (groups B and D, Fig 2). Furthermore, *psa*I loss was detected in certain *Lathyrus* species (groups C, D and E, Fig 2).

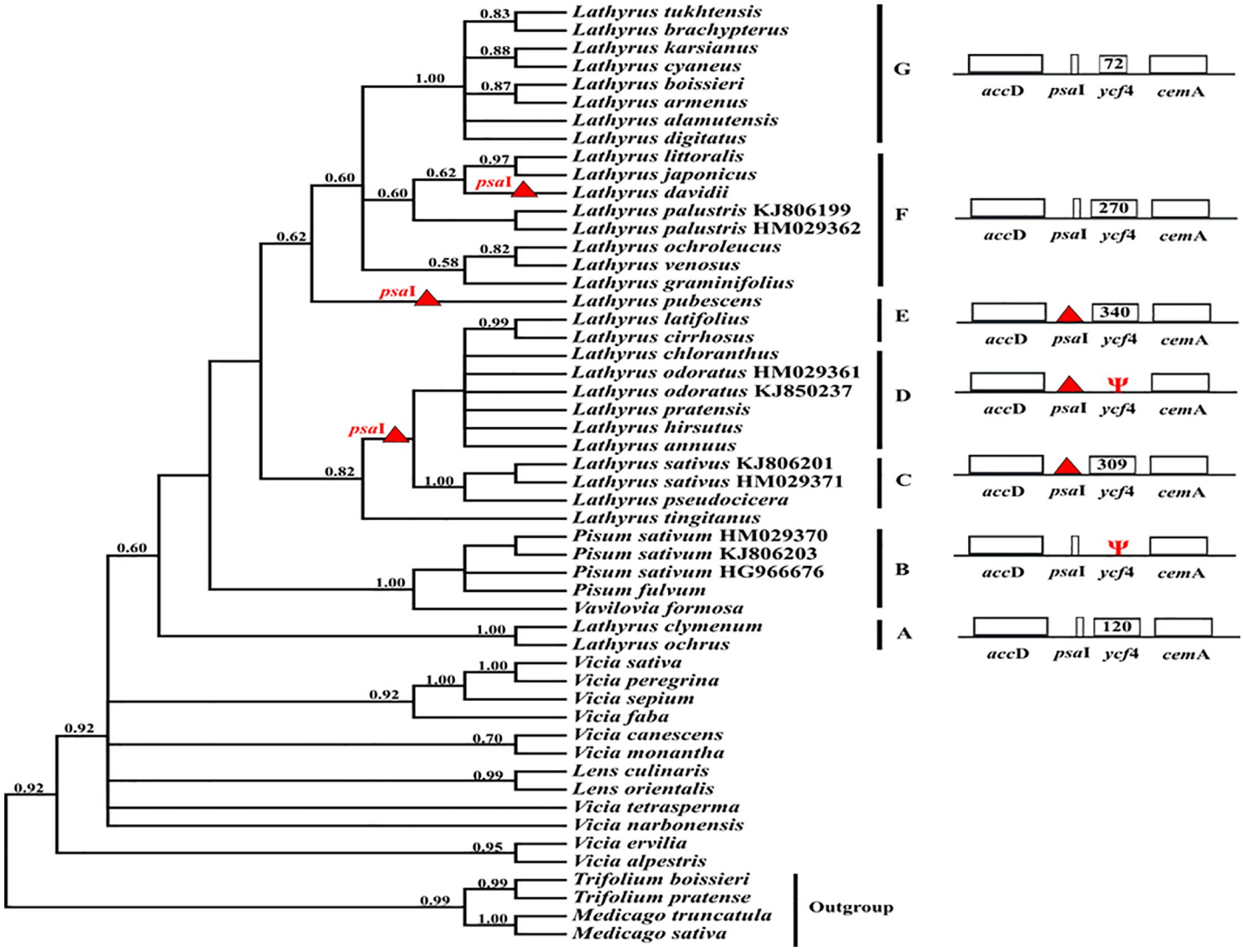

Fig 2. Presence and absence of the *ycf*4 and *psa*I genes in *Lathyrus*, *Pisum* and *Vavilovia*. Red triangles on branches indicate evolutionary losses of the *psa*I gene. Psi symbols denote pseudogenes. Numbers indicate the numbers of codons in *ycf*4 gene. The topology of the tree is the result of Bayesian analysis of tribe Fabeae based on *ycf*4 gene. Numbers above branches are posterior probability values. Values <50% were not shown.

https://doi.org/10.1371/journal.pone.0229846.g002

dN/dS analysis

For each gene, we used the method of Yang [44] to investigate the number of non-synonymous (dN) and synonymous (dS) nucleotide substitutions and the ratio of them (dN/dS, ω) across branches and found that this ratio varies significantly among lineages, under the tree topology resulting from Bayesian analysis. In the dN/dS analyses (Figs 3 and 4), acceleration of the evolutionary rate is seen in *ycf*4 in Fabeae, particularly *Lathyrus*, relative to other IRLC genera (Fig 3). The same high acceleration rate is, however, not seen in two other plastid genes (*mat*K and *rpl*32) across the IRLC (Fig 4). The level of constraint on *ycf*4 gene in *Lathyrus* is lower than in the other genera of IRLC. dN/dS ratio among the genera belonging to different tribes is less than that ratio among the related species of *Lathyrus*, for instance, dN/dS = 0.716 between *Medicago sativa* and *Astragalus membranaceous* compared with dN/dS = 1.527 within *L*. *davidii* and *L*. *littoralis*. In other words, *ycf*4 is highly conserved in all genera of the IRLC except *Lathyrus*.

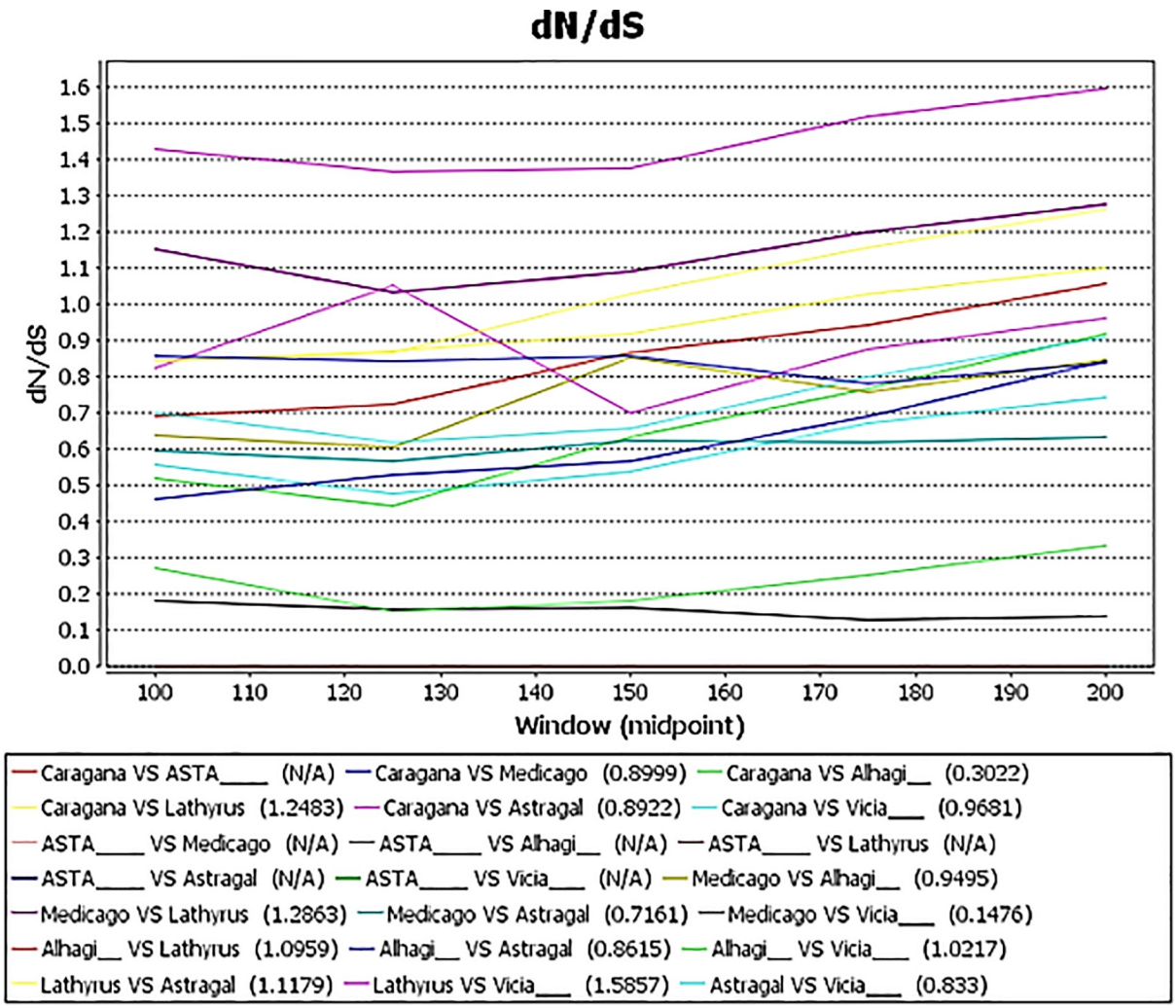

Fig 3. JCoDA output using sliding window analysis of pairwise taxa dN/dS. All parwise comparisons were performed using a window 100. All pairwise comparisons suggested positive selection of *ycf*4 in *Lathyrus*.

https://doi.org/10.1371/journal.pone.0229846.g003

We then applied the branch-site model, which accommodates heterogeneity among sites and can reflect divergent selective pressures. Parameter estimates under this model suggested that, when the *Lathyrus* branch was the foreground branch, some set of sites in *ycf*4 gene evolved under positive selective pressures (S3 Table). Using the Bayes empirical Bayes method, we identified seven codon sites in *ycf*4 gene with posterior probabilities $\geq$ 95% that evolved under positive selective pressure on the *Lathyrus* branch (1 L, 2 S, 3 V, 4 V, 5 L, 6 L, 7 T). When other branches from other genera were the foreground branches, no codon sites with posterior probabilities $\geq$ 95% were identified under positive selective pressures (S3 Table).

The *ycf*4 gene is highly variable in terms of length and point mutations compared to the other two genes among *Lathyrus* species (Table 1). The length of *ycf*4, among *Lathyrus* species which have the intact gene, varies from 219 bp in group G (including *L. tukhtensis*, *L. brachypterus*, *L. karsianus*, *L. cyaneus*, *L. boissieri*, *L. armenus*, *L. alamutensis* and *L. digitatus*) to 1023 bp in group E (including *L. latifolius* and *L. cirrhosus*). While the gene length in the other two genes (1512–1527 bp in *mat*K and 153–183 bp in *rpl*32) does not show much variation. Furthermore, the nucleotide substitution rates in *Lathyrus* species are clearly elevated in *ycf*4

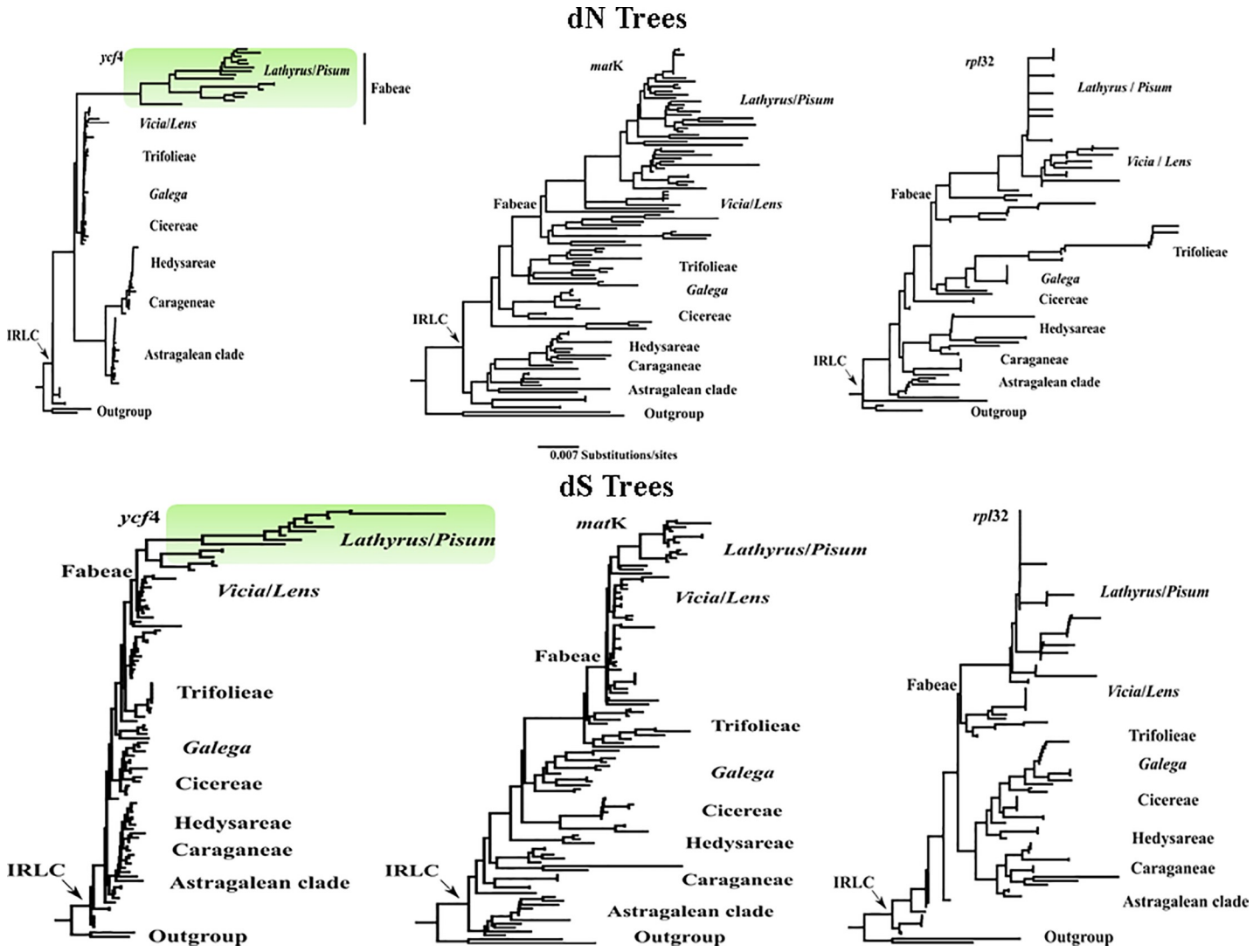

Fig 4. Synonymous and nonsynonymous divergence in IRLC *ycf4* sequences. All trees are drawn to the same scale. Green branches in the dN and dS trees indicate taxa in which *ycf*4 gene represents positive selection and acceleration in evolutionary rate. Trees for *mat*K and *rpl*32 genes do not show comparable rate at either synonymous or nonsynonymous divergence.

https://doi.org/10.1371/journal.pone.0229846.g004

when compared with other two cpDNA genes, for example, there are four nucleotide substitutions between *L. littoralis* and *L. japonicus* in *mat*K sequences and there are no differences between their *rpl*32 gene, however, there are 67 nucleotide substitutions between these two

Table 1. Sequence divergence in cpDNA regions compared among *Lathyrus* species.

Genes	*L.sativus* vs. *L.ochroleucus*			*L.davidii* vs. *L.littoralis*			*L.odoratus* vs. *L. palustris*		
	d_N/d_S	$d_N \pm SE$	$d_S \pm SE$	d_N/d_S	$d_N \pm SE$	$d_S \pm SE$	d_N/d_S	$d_N \pm SE$	$d_S \pm SE$
*ycf*4	3.589	2.006±1.74	0.559±0.43	1.527	0.218±0.136	0.143±0.150	NA	NA	NA
*mat*K	0.390	0.020±0.00	0.052±0.01	0.218	0.008±0.00	0.003±0.00	0.349	0.021±0.00	0.060±0.01
*rpl*32	0.223	0.009±0.00	0.044±0.04	0.0	0.0	0.0	0.193	0.019±0.01	0.097±0.07

For protein-coding genes the synonymous divergence (dS), the nonsynonymous divergence (dN), its standard error (SE), and the nonsynonymous-to-synonymous ratio (dN/dS, also called ω) are shown. NA, Not applicable (gene not present).

https://doi.org/10.1371/journal.pone.0229846.t001

species in *ycf*4 gene. All of the investigated species of *Lathyrus* showed signs of elevated branch lengths in the *ycf*4 gene (Fig 4). Our analysis for estimation of positive selection revealed that *Lathyrus* branch within *ycf*4 data has undergone adaptive evolution and showed ω value greater than 1 (Table 1, Fig 4).

Discussion

Phylogenetic estimations

Within the papilionoid legumes, many phylogenetic studies have been performed on the IRLC that confirm its monophyletic origin [14, 51]. The IRLC consists of several tribes (as mentioned above) [14, 18]. Tribe Fabeae comprises five genera (*Lathyrus*, *Pisum*, *Vicia*, *Lens*, *Vavilovia*) and c. 380 species and contains some important crop species. In present work in agreement with previous studies [52–54], Fabeae has monophyletic origin and is sister group to *Trifolium*. Like previous studies [52–54], our results confirm the monophyly of *Lens* and *Pisum* and the paraphyly of *Vicia* and *Lathyrus*. *Pisum* and *Vavilovia* are sister groups and nested in *Lathyrus*, which together with *Lens* is nested in *Vicia*.

Extensive phylogenetic analyses have been undertaken in the IRLC and tribe Fabeae, using mainly organellar or rDNA markers [6, 14, 53] which have resolved relationships between genera and even species at lower levels. In the present work, phylogenetic relationships have been investigated only based on *ycf*4 gene. This gene lacked adequate phylogenetic signal at the lower taxonomic level but showed better phylogenetic resolution at higher taxonomic level (generic to tribal rank). For example, in some systematic and phylogenetic studies [52–54] based on *mat*K and *rbc*L, the relationships between sections and species of *Lathyrus* have been carefully resolved, for instance, *L*. *odoratus* and *L*. *hirsutus* are closest relatives but in the present study the relationship between them is unclear due to low support value. In our study, the *ycf*4 gene did not resolve the relationships at low levels and it is suggested that *ycf*4 along with other plastid genes such as *mat*K, *rbc*L and *ndh*F be used for molecular phylogenetic studies of genera and species.

Sequence divergence value across the IRLC is less than 0.751 (0.000–0.750) and among *Lathyrus* species is less than 0.534 (0.000–0.533). Therefore, the low phylogenetic resolution at low taxonomic levels is due to the low number of synapomorphic characters and showed strong evidence of homoplasy. Homoplasy can be caused by different factors such as high selection pressures and mutation rates [55]. In *ycf*4 sequences, length and point mutations may cause homoplasy and lower phylogenetic resolution. The gene is not useful for resolving lower level relationships (e.g. at the level of species and lower).

The lack of *ycf*4 has evolved twice within the IRLC; once in the clade containing *Pisum* and *Vavilovia* specie (group B) and again in some *Lathyrus* species including *L*. *chloranthus*, *L*. *odoratus*, *L*. *pratensis*, *L*. *hirsutus* and *L*. *annuus* (group D). *psa*I gene also shows independent evolution three times in different species of *Lathyrus* (including *L*. *davidii*, *L*. *pubescens* and groups C/D/E).

Magee et al. [21] detected some previous false reports about *ycf*4 gene losses in four legume species (*Glycine max*, *Trifolium subterraneum*, *Cicer arietinum*, *Medicago truncatula*). We also noticed that the *ycf*4 gene which was not identified in cpDNA of *Medicago sativa* [56] and 13 *Lathyrus* species [57] is, in fact, present in the plastid genome of these species. These *Lathyrus* species are still referred without *ycf*4 gene [58]. We identified the gene in the plastome of the species using ORF finder and comparison with other related species. Due to the high evolutionary rate of the *ycf*4 gene and consequently the high divergence of gene, different softwares which are used to annotate like DOGMA [59] cannot recognize it.

Positive selection and rapid evolution of *ycf4*

In the present work, we have investigated the evolutionary rate of the open reading frame *ycf*4 and its genomic region in the IRLC, and particularly in Fabeae. Our results showed that *ycf*4 gene and its upstream gene (*psa*I) are more variable in the tribe Fabeae than in other tribes of the IRLC. The *ycf*4 gene was found in all tribes of the IRLC except Fabeae completely intact and well conserved. In certain species of *Lathyrus*, *Pisum* and *Vavilovia*, *ycf*4 is either absent or pseudogenized. In legumes, in addition to *Lathyrus*, *ycf*4 shows signs of pseudogenization in *Desmodium heterocarpon* [60]. Moreover, the genomic region around *ycf*4 including *psa*I gene at the upstream of it, in some species of *Lathyrus* has been lost (Fig 2).

*ycf*4 gene encoded a thylakoid protein which is involved in the assembly of the photosystem I complex (PSI) as a part of an energy harvesting process. PSI embedded in the thylakoid membranes of phototrophs (cyanobacteria, algae and plants) and mediates the light-induced electron transfer from plastocyanin or cytochrome *c* to ferredoxin. PSI contains at least 11 subunits, 5 of which are encoded by the plastid genome (PsaA, PsaB, PsaC, PsaI and PsaJ) and other subunits (PsaD, PsaE, PsaF, PsaG, PsaH and PsaK) are nuclear-encoded subunits [4, 28, 61]. All of the plastid encoded subunits are transmembrane proteins with specific and important functional roles except subunits I and J. The function of these two subunits has not yet been determined and it seems that the presence of subunits I and J are not necessary for the PSI function [4, 28, 62]. *psa*I gene encodes a small protein which is conserved among land plants and *psa*I mutants in tobacco plants have standard growth conditions [63]. Our data demonstrated that *psa*I gene has been lost in some *Lathyrus* species (some of them also lack *ycf*4 gene such as group D in Fig 2). Therefore, according to the result of tobacco mutants, it is conceivable that *Lathyrus* species without *psa*I gene do not show abnormal phenotype and have standard growth. Some parasitic species like *Cuscuta gronovii*, *C*. *obtusiflora* (photosynthetic) and *Epifagus virginiana* (non-photosynthetic) and also some genera from green algae (*Chromera*, *Vitrella*, *Aureococcus*, *Bigelowiella natans* and *Euglena gracilis*) are other known examples in which *psa*I gene is absent [64–67].

In order to evaluate the function of *ycf*4 gene, different studies have generated stable knockout mutants for *ycf*4 [28, 63]. It is expected that the lack of *ycf*4 gene will be resulted in loss of PSI activity and reduce autotrophic growth. But the analysis of *ycf*4-deficient mutant appears that *ycf*4 is required for the assembly and stability of the PSI not for the synthesis of that. Ycf4 is one of the most important chaperons in PSI assembly process but the functional role of that has not yet been clearly identified [4, 68]. Accordingly, some studies [4, 30] have suggested renaming this factor to *paf*II (PSI assembly factor II). Krech et al. [63] have shown that Ycf4 protein is an important but non-essential factor for PSI assembly process. Given that all species which have lost these two genes are photosynthetic and certainly have a functional PSI, it seems that other alternative factors appear to perform *ycf*4 and *psa*I functions [63]. Our study suggests that the photosystem I complex of *Lathyrus*, *Pisum* and *Vavilovia* underwent unique structural changes. In the present study, by comparing the evolutionary rates of the three chloroplast genes, found that only *ycf*4 gene had dN/dS value > 1, indicating that this gene had undergone positive selection. This positive selection across the IRLC and Fabeae is only seen in *Lathyrus*, *Pisum* and *Vavilovia*. Thus, rapid evolution of *ycf*4 is locus and lineage specific and is not a shared character of the IRLC in legumes. The presence of fast-evolving protein gene in *Lathyrus* is probably due to the high point and length mutations rates that may result from repeated DNA breakage and repair [21, 69].

When a gene is lost from organellar genome, the probability given is that the organelle gene has been transferred from organelle to nuclear. In such cases, based on the ratio between the point mutation rates in the organelle and nuclear, it can be determined whether the gene has

been transmitted [21, 70]. In this context, evolutionary transfer of *acc*D gene from the plastome of *Trifolium subterraneum* to the nucleus can be mentioned [24]. Magee et al. [21] showed that there are no nuclear copies for *ycf*4 and *psa*I in the *L*. *odoratus* and *P*. *sativum*, therefore, it can be concluded that these genes have not been transferred to the nucleus. Our study demonstrated that each three tandem genes *psa*I-*ycf*4-*cem*A is situated in a local mutation hotspot in particular within *Lathyrus*, resulting in dramatic acceleration of sequence evolution in some species and evolutionary gene losses in others [6, 21, 71]. Given that in addition to *ycf*4 and *psa*I genes, other genes such as *acc*D and *rps*16 in legumes have been lost or pseudogenized (a phenomenon that is very rare in other angiosperms), it is supposed that a hotspot might have been existed across the legume evolution and caused the acceleration of the *ycf*4 but the exact location of hotspot has varied.

Conclusion

The present study investigated the presence/absence and evolutionary process of *ycf*4/*psa*I genes in the IRLC and particularly Fabeae. The *ycf*4 gene was found in all tribes of the IRLC except Fabeae completely intact and conserved. In certain species of *Lathyrus*, *Pisum* and *Vavilovia*, *ycf*4/*psa*I is either absent or pseudogenized. Tribe Fabeae comprises five genera (*Lathyrus*, *Pisum*, *Vicia*, *Lens*, *Vavilovia*) and c. 380 species and contains some important crop species. The *ycf*4 gene is highly variable in terms of length and point mutations compared to the other two genes (*mat*K and *rpl*32) among *Lathyrus* species. In the present study, by comparing the evolutionary rates of the three chloroplast genes, we found that only *ycf*4 gene had dN/dS value > 1, indicating that this gene had undergone positive selection. This positive selection across the IRLC and Fabeae is only seen in *Lathyrus*, *Pisum* and *Vavilovia*. Thus, rapid evolution of *ycf*4 is locus and lineage specific and is not a shared character of the IRLC in legumes.

Supporting information

S1 Fig. Relative position of the PCR amplification and sequencing primers used in this study. Arrows indicate the direction of strand synthesis. Boxed areas represent coding region.
(TIF)

S1 File. Alignment of ycf4 gene sequences in IRLC legumes.
(PDF)

S1 Table. Taxa included in the *ycf*4, *mat*K and *rpl*32 analyses. (-) not available in GenBank. Abbreviations used in plant accession information: FMUH, Ferdowsi University of Mashhad Herbarium, Mashhad, Iran; GAZI, Gazi Universitesi Herbarium, Ankara, Turkey; IRAN, Iranian Research Institute of Plant Protection,Tehran, Iran; MO, Missouri Botanical Garden Herbarium, St Louis, USA; MSB Herbarium of Ludwig-Maximilians-Universitat, Munchen, Germany; TARI Herbarium of the Research Institute of Forests and Rangelands, Tehran, Iran; TMUH, Tarbiat Modares University Herbarium, Tehran, Iran; TUH, Tehran University Herbarium, Tehran, Iran; HWANRC Herbarium of West Azarbayjan Natural Resources Research Center, Urmia, Iran. [a]Sequences from GenBank. [b]Whole plastid genome.
(PDF)

S2 Table. Location and base composition of amplification and sequencing primers used in this study. * Location indicates the Start and end nucleotide positions.
(PDF)

S3 Table. Parameter estimation and likelihood ratio tests for the branch-site model.
(PDF)

Acknowledgments

This paper is part of PhD dissertation of the M.M. granted by research council of Tarbiat Modares University.

Author Contributions

Conceptualization: Shahrokh Kazempour-Osaloo.

Data curation: Mahtab Moghaddam.

Formal analysis: Mahtab Moghaddam.

Supervision: Shahrokh Kazempour-Osaloo.

Visualization: Mahtab Moghaddam.

Writing – original draft: Mahtab Moghaddam.

Writing – review & editing: Shahrokh Kazempour-Osaloo.

References

1. Palmer JD. Plastid chromosomes: structure and evolution. In: Bogorad L.; Vasil I., editors. Cell Culture and Somatic Cell Genetics of Plants. Academic Press; San Diego, California, USA: 1991. p. 5–53.
2. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 2007; 104:19369–19374. https://doi.org/10.1073/pnas.0709121104 PMID: 18048330
3. Jansen RK, Ruhlman TA. Plastid genomes of seed plants. In: Genomics of Chloroplasts and Mitochondria. Advances in Photosynthesis and Respiration (Bock R. and Knoop V. eds); 2012. pp. 103–126. The Netherlands: Springer.
4. Wicke S, Schneeweiss GM, dePamphilis CW, M€ uller KF, Quandt D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Molecular Biology. 2011; 76: 273–297. https://doi.org/10.1007/s11103-011-9762-4 PMID: 21424877
5. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, et al. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution. 2006; 23: 2175–2190. https://doi.org/10.1093/molbev/msl089 PMID: 16916942
6. Schwarz EN, Ruhlman TA, Sabir JSM, Hajarah NH, Alharbi NS, Al- Malki AL, et al. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. J Syst Evol. 2015; 53:458–468. https://doi.org/10.1111/jse.12179
7. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. Molecular Biology and Evolution. 2011; 28: 583–600. https://doi.org/10.1093/molbev/msq229 PMID: 20805190
8. Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, et al. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trn*Q, *trn*K, *psb*A, *trn*I and *trn*H and the absence of *rps*16. Molecular and General Genetics. 1992; 232: 206–214. https://doi.org/10.1007/bf00279998 PMID: 1557027
9. Palmer JD, Thompson WF. Rearrangements in the chloroplast genomes of mung bean and pea. Proceedings of the National Academy of Sciences USA. 1981; 78: 5533–5537.
10. Palmer J.D., Osorio B., Aldrich J. and Thompson W.F. Chloroplast DNA evolution among legumes: loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr. Genet. 1987; 11: 275–286.
11. Bock R, Knoop V. Genomics of chloroplasts and mitochondria. Dordrecht : Springer Netherlands. 2012.

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. https://doi.org/10.3732/ajb.91.11.1846 PMID: 21652332

15. Lavin M, Doyle JJ, Palmer JD. Evolutionary significance of the loss of the chloroplast DNA inverted repeat in the Leguminosae subfamily Papilionoideae. Evolution. 1990; 44, 390–402. https://doi.org/10.1111/j.1558-5646.1990.tb05207.x PMID: 28564377

16. Duan L, Yang X, Liu P, Johnson G, Wen J, Chang Z. A molecular phylogeny of Caraganeae (Leguminosae, Papilionoideae) reveals insights into new generic and infrageneric delimitations. PhytoKeys. 2016; 70: 111–137.

17. 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-*trn*L(UAG) sequences data. Plant Biosystems. 2017; 6: 1082–1093.

18. 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. https://doi.org/10.3897/phytokeys.125.34877 PMID: 31303810

19. Guo X, Castillo-Ramırez S, Gonzalez V, Bustos P, Fernandez-Vazquez JL, Santamaria RI, et al. Rapid evolutionary change of common bean (*Phaseolus vulgaris* L.) plastome and the genomic diversification of legume chloroplasts. BMC Genomics. 2007; 8: 228–244. https://doi.org/10.1186/1471-2164-8-228 PMID: 17623083

20. 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. https://doi.org/10.1016/j.ympev.2008.06.013 PMID: 18638561

21. 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. https://doi.org/10.1101/gr.111955.110 PMID: 20978141

22. Sabir J, Schwarz EN, Ellison N, Zhang J, Baeshen NA, Mutwakil M, et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnology Journal. 2014; 12: 743–754. https://doi.org/10.1111/pbi.12179 PMID: 24618204

23. Sveinsson S, Cronk Q. Evolutionary origin of highly repetitive plastid genomes within the clover genus (Trifolium). BMC Evolutionary Biology. 2014; 14: 228. https://doi.org/10.1186/s12862-014-0228-6 PMID: 25403617

24. Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, Mc Murtry V, et al. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. Journal of Molecular Evolution. 2008; 67: 696–704. https://doi.org/10.1007/s00239-008-9180-7 PMID: 19018585

25. Neumann P, Koblížková A, Navrátilová A, Macas J. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics. 2006; 173: 1047–56. https://doi.org/10.1534/genetics.106.056259 PMID: 16585134

26. Macas J, Neumann P, Navrátilová A. Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics. 2007; 8: 427. https://doi.org/10.1186/1471-2164-8-427 PMID: 18031571

27. Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumann P, et al. In Depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS ONE. 2015; 10(11): e0143424. https://doi.org/10.1371/journal.pone.0143424 PMID: 26606051

28. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix J. The chloroplast *ycf*3 and *ycf*4 open reading frames of *Chlamydomonas reinhardtiiare* required for the accumulation of the photosystem I complex. EMBO J. 1997; 16: 6095–6104. https://doi.org/10.1093/emboj/16.20.6095 PMID: 9321389

29. Ruf S, Kossel H, Bock R. Targeted inactivation of a tobacco intron–containing open reading frame reveals a novel chloroplast-encoded photosystem I–related gene. J Cell Biol. 1997; 139: 95–102. https://doi.org/10.1083/jcb.139.1.95 PMID: 9314531

30. Ozawa S, Nield J, Terao A, Stauber EJ, Hippler M, Koike H., et al. Biochemical and structural studies of the large Ycf4-photosystem I assembly complex of the green alga *Chlamydomonas reinhardtii*. Plant Cell. 2009; 21: 2424–2442. https://doi.org/10.1105/tpc.108.063313 PMID: 19700633

31. Lee H, Jansen RK, Chumley TW, Kim K. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol. 2007; 24: 1161–1180. https://doi.org/10.1093/molbev/msm036 PMID: 17329229

32. Ren T, Yang Y, Zhou T, Liu ZL. Comparative plastid genomes of *Primula* species: sequence divergence and phylogenetic relationships. Int. J. Mol. Sci. 2018; 19, 1050.

33. Kim Y., Cullis C. A novel inversion in the chloroplast genome of marama (*Tylosema esculentum*). Journal of Exoerimental Botany. 2017; 8: 2065–2072.

34. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987; 19: 11–15.

35. Edgar RC. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 2004; 32: 1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

36. Loytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci. 2005; 102: 10557–10562. https://doi.org/10.1073/pnas.0409137102 PMID: 16000407

37. Loytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008; 320: 1632–1635. https://doi.org/10.1126/science.1158395 PMID: 18566285

38. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans, LA. 2010; 8. pp. https://doi.org/10.1109/GCE.2010.5676129

39. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012; 61:539–542. https://doi.org/10.1093/sysbio/sys029 PMID: 22357727

40. Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer, version 1.6. Available from: http://beast.bio.ed.ac.uk/Tracer.

41. Huson DH, Richter DC, Rausch C, Rupp R. User Manual for Dendroscope V2.7.4. 2010. Available from: http://www-ab.Informatik. uni-tuebin.de/software/dendroscope.

42. Nylander JAA. MrModeltest v2. 2004. Program distributed by the author. Uppsala: Evolutionary Biology Centre, Uppsala University.

43. 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. https://doi.org/10.1080/10635150490522304 PMID: 15545256

44. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

45. Steinway SN, Dannenfelser R, Laucius CD, Hayes JE, Nayak S. JCoDA: a tool for detecting evolutionary selection. BMC Bioinformatics. 2010; 11:284. https://doi.org/10.1186/1471-2105-11-284 PMID: 20507581

46. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYWEasycodeml: a visual tool for analysis of selection using codeML. Ecol. Evol. 2019; 9:3891–3898. https://doi.org/10.1002/ece3.5015 PMID: 31015974

47. Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol. 1999; 16:1315–1328. https://doi.org/10.1093/oxfordjournals.molbev.a026042 PMID: 10563013

48. Koonin EV, Rogozin IB. Getting positive about selection. Genome Biol. 2003; 4:331. https://doi.org/10.1186/gb-2003-4-8-331 PMID: 12914654

49. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005; 22:1107–1118. https://doi.org/10.1093/molbev/msi097 PMID: 15689528

50. Sanderson MJ, Wojciechowski MF. Diversification rates in a temperate legume clade: are there "so many species" of *Astragalus* (Fabaceae)? Am. J. Bot. 1996; 83: 1488–1502.

51. Wojciechowski M.F. Towards a new classification of Leguminosae: naming clades using non-linnaean phylogenetic nomenclature. South African Journal of Botany. 2013; 89: 85–93.

52. Kenicer GJ, Kajita T, Pennington RT, Murata J. Systematics and biogeography of *Lathyrus* (Leguminosae) based on internal transcribed spacer and cpDNA sequence data. Am J Bot. 2005; 92(7): 1199–209. https://doi.org/10.3732/ajb.92.7.1199 PMID: 21646142

53. Schaefer H, Hechenleitner P, Santos-Guerra A, Menezes de sequeira M, 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 Evolutionary Biology. 2012; 12:250. https://doi.org/10.1186/1471-2148-12-250 PMID: 23267563

54. Oskoueiyan R, Kazempour Osaloo S, Amirahmadi A. Molecular phylogeny of the genus *Lathyrus* (Fabaceae-Fabeae) based on cpDNA *mat*K sequence in Iran. Iran J Biotech. 2014; 12(2): 41–48.

55. Crispell J, Balaz D, Gordon S.V. Homoplasy finder: a simple tool to identify homoplasies on a phylogeny. Microbial Genomics. 2019. https://doi.org/10.1099/mgen.0.000245 PMID: 30663960

56. Tao X, Ma L, Zhang Z, Liu W, Liu Z. Characterization of the complete chloroplast genome of alfalfa (*Medicago sativa*) (Leguminosae). Gene Reports. 2017; 6: 67–73.

57. Sveinsson S, Cronk Q. Conserved gene clusters in the scrambled plastomes of IRLC legumes (Fabaceae: Trifolieae and Fabeae) Saemundur Sveinsson, Quentin Cronk bioRxiv 040188; 2016. https://doi.org/10.1101/040188

58. Jiang M, Chen H, He S, Wang L, Chen AJ, Liu C. Sequencing, characterization, and comparative analyses of the plastome of *Caragana rosea* var. *rosea*. Int. J. Mol. Sci 2018; 19:1419.

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

60. Jin D-P, Choi I-S, Choi B-H. Plastid genome evolution in tribe Desmodieae (Fabaceae: Papilionoideae). PLoSONE. 2019; 14(6): e0218743.https://doi.org/10.1371/journal.pone.0218743.

61. Nelson N, Yocum CF. Structure and function of photosystems I and II. Annu Rev Plant Biol. 2006; 57: 521–565. https://doi.org/10.1146/annurev.arplant.57.032905.105350 PMID: 16669773

62. Bock R. Structure, function, and inheritance of plastid genomes. In: Bock R (ed) Cell and Molecular Biology of Plastids. Springer, Berlin Heidelberg. 2007; pp. 29–63.

63. Krech K, Ruf S, Masduki FF., Thiele W, Bednarczyk D, Albus CA, et al. The plastid genome-encoded YCF4 protein functions as a nonessential assembly factor for photosystem I in higher plants. Plant Physiology. 2012; 159: 579–591. https://doi.org/10.1104/pp.112.196642 PMID: 22517411

64. Wolfe KH, Morden CW, Palmer JD. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci. 1992; 89:10648–10652. https://doi.org/10.1073/pnas.89.22.10648 PMID: 1332054

65. Funk HT, Berg S, Krupinska K, Maier UG, Krause K. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Biol. 2007; 7:45. https://doi.org/10.1186/1471-2229-7-45 PMID: 17714582

66. McNeal JR, Kuehl JV, Boore JL, Depamphilis CW. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biol. 2007; 7:57. https://doi.org/10.1186/1471-2229-7-57 PMID: 17956636

67. Sobotka R, Esson HJ, Konik P, Trskova E, Moravcova L, Horak A, et al. Extensive gain and loss of photosystem I subunits in chromerid algae, photosynthetic relatives of apicomplexans. Scientific Reports. 2017; 7:13214. https://doi.org/10.1038/s41598-017-13575-x PMID: 29038514

68. Nellaepalli S, Ozawa S-I, Kuroda H, Takahashi Y. The photosystem I assembly apparatus consisting of Ycf3-Y3IP1 and Ycf4 modules. Nature Communications. 2018; 9:2439. https://doi.org/10.1038/s41467-018-04823-3 PMID: 29934511

69. Yang Y, Sterling J, Storici F, Resnick MA, Gordenin DA. Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. PLoS Genet. 2008; 4:e1000264. https://doi.org/10.1371/journal.pgen.1000264 PMID: 19023402

70. Brandvain Y, Wade MJ. The functional transfer of genes from the mitochondria to the nucleus: The effects of selection, mutation, population size and rate of self-fertilization. Genetics. 2009; 182:1129–1139. https://doi.org/10.1534/genetics.108.100024 PMID: 19448273

71. Keller J, Rousseau-Gueutin M, Martin G E., 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 of four novel legume chloroplast genomes from *Lupinus*. DNA Research. 2017; 24(4): 343–358. https://doi.org/10.1093/dnares/dsx006 PMID: 28338826