Conserved gene clusters in the scrambled plastomes of IRLC legumes (Fabaceae: Trifolieae and Fabae).

Saemundur Sveinsson1*, Quentin Cronk2

1Faculty of Land and Animal Resources, Agricultural University of Iceland, Keldnaholt, 112 Reykjavik, Iceland; 2Department of Botany and Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver BC V6T 1Z4, Canada.

Author for correspondence:

Saemundur Sveinsson
Tel: 354-4335219
Email: saemundur@lbhi.is

<table>
<thead>
<tr>
<th>Section</th>
<th>Word Count</th>
<th>Figures</th>
<th>Supporting Information Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total word count</td>
<td>2884</td>
<td>3</td>
<td>3 (Figs 1 and 2 in colour)</td>
</tr>
<tr>
<td>Summary</td>
<td>179</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Introduction</td>
<td>397</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>1038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>663</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussions</td>
<td>719</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>61</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

The plastid genome retains several features from its cyanobacterial-like ancestor, one being the co-transcriptional organization of genes into operon-like structures. Some plastid operons have been identified but undoubtedly many more remain undiscovered. Here we utilize the highly variable plastome structure that exists within certain legumes of the inverted repeat lost clade (IRLC) to find conserved gene clusters. These plastomes exhibit an unusually high frequency of translocations and inversions. We analysed the plastomes of 23 legume species and identified 32 locally collinear blocks (LCBs), which are regions within the plastid genomes that occur in different orientation and/or order among the plastid genomes but are themselves free from internal rearrangements. Several represent gene clusters that have previously been recognized as plastid operons. It appears that the number of LCBs has reached saturation in our data set, suggesting that these LCBs are not random, but likely represent legume plastid operons protected from internal rearrangement by functional constraint. Some of the LCBs we identify, such as psbD/C/Z, are previously known plastid operons. Others, such as rpl32-ndhF-psbA-matK-rbcL-atpB-atpE, may represent novel polycistronic operons in legumes.

Key words (5-8): Fabeae, IRLC, massively parallel sequencing, plastid operons, plastome evolution, plastome rearrangements, Trifolieae.
**Introduction**

The plastid genome, also known as the plastome, refers to the total genetic information of a single plant organelle, the plastid, which takes many developmental forms, the most notable being the chloroplast (Bock, 2007). Plastid genomes are circular structures of double stranded DNA, usually consisting of about 100-120 genes and are around 120-160 kb long in photosynthesizing plants (Bock, 2007). Their size, structure and gene content are highly conserved across land plants (Wicke et al., 2011). However there are exceptions, such as the Geraniaceae and Campanulaceae, which are two angiosperm families known to contain species with highly rearranged plastomes (Haberle et al., 2008; Guisinger et al., 2011). A dominating feature of plastid genomes is the presence of a large inverted repeat (Wicke et al., 2011; Zhu et al., 2015) separated by a small single copy region that is variable in orientation (Walker et al. 2015). However some plant groups have lost one copy of the repeat, one being a clade within papilionoid legumes (Fabaceae), known as the inverted repeat lost clade (IRLC) (Wojciechowski et al., 2000).

Plants obtained their plastid organelles through an endosymbiosis event with a cyanobacteria-like organism, about 1.5 – 1.6 billion years ago (Margulis, 1970; Hedges et al., 2004). Its bacterial origin gives the plastid genome many prokaryotic features, such as small (70S) ribosomes and the absence of mRNA 3’ polyA tails (see Stern et al., 2010 for a review). An additional ancestral feature of the plastid genome is the organization of its coding region into multiple gene clusters, or operons (Sugita & Sugiura, 1996; Sugiura et al., 1998). These gene clusters are stretches of the plastome consisting of several genes that are transcribed into di- or polycistronic units, which are then processed before translation (Stern et al., 2010). Several such clusters have already been identified in the plastid (Adachi et al., 2012; Ghulam et al., 2013; Stoppel & Meurer, 2013).

Several legume genera within the IRLC are known to harbour highly rearranged plastomes, as a result from multiple translocations and/or inversions: *Trifolium* (Cai et al., 2008; Sabir et al., 2014; Sveinsson & Cronk, 2014), *Pisum* (Palmer & Thompson, 1982), *Lathyrus* (Magee et al., 2010), *Lens* and *Vicia* (Sabir et al., 2014). The aim of this study is to analyse these rearrangements in these genera within IRLC, in order to investigate whether they can be used to study the organization of plastid genomes into operons.
Material and Methods

Source of plant material
The plant material for this study came from three sources. First, live plants were collected in the field and transplanted to a glasshouse facility. Secondly, seeds were obtained from a commercial provider, Roger Parsons Sweet Peas (Chichester, UK). Thirdly, seeds were received from the USDA germplasm collection at Pullman, Washington (W6). A full list of germplasm used is given in Table 1. All plants were grown in glasshouse facilities at UBC. In all cases where plants required critical determination they were grown until flowering, and herbarium voucher specimens were then collected (UBC).

Illumina sequencing
Total DNA was extracted from fresh leaf material following a modified version of the CTAB protocol (Doyle & Doyle, 1987). RNase treatments were performed (cat. 19101, QIAGEN, Germantown, MD) and DNA quality was assessed by visual inspections on 1% agarose gels. Illumina sequencing libraries were constructed from high quality DNA, using the NEXTflexTM DNA sequencing kit (100 bp Paired-End reads) (cat: 5140-02, BiooScientific Corp, TX). We followed the manufacturer’s protocol and c. 400 bp DNA fragments were size selected using Agencourt AMPure XpTM magnetic beads (cat. A63880, BeckmanCoulter Genomics, MA). Completed libraries were pooled and sequenced on a lane of the Illumina HiSeq-2000 platform.

Plastid genome assemblies and annotation
Trimmomatic v.0.3 (Lohse et al., 2012) was used to trim and remove low quality Illumina reads, with the following flags: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15MINLEN:36. High quality reads were used in all subsequent analysis and singlet reads, i.e. reads without a paired end, were discarded. We used the de novo method implemented in CLC Genomic Workbench v.7.0.2 to generate assemblies for each species, using the default settings. Contigs of plastid origin were identified by a BLASTN search (Altschul et al., 1997) to a plastid genome of a closely related species. These were generally the largest and most high coverage contigs in the de novo assembly and always had an E-value of 0 when blasted to the reference plastome. Regions with nucleotides
scored as Ns were manually resolved by retrieving sequence information directly from the quality-trimmed reads. For most species, the *de novo* assembly returned a single large plastid contig. When needed, multiple contigs containing plastid sequence were joined by hand, using information from the quality-trimmed reads. The quality of each plastome assembly was verified by visually by inspecting a BWA mem pileup, v. 0.7.5a (Li & Durbin, 2009), of paired end reads using Tablet v.1.13.12.17 (Milne *et al*., 2013). We made sure that the connections between manually joined contigs were supported by paired-end read mapping. Finally all plastome assemblies were annotated using DOGMA (Wyman *et al*., 2004).

**Phylogenetic analysis**

Due to the extensive rearrangements observed in the plastomes (see Sabir *et al*., 2014; Sveinsson & Cronk, 2014), we restricted our plastome phylogenetic analysis to protein coding genes. We used a custom phylogenetic pipeline, plast2phy, that extracted protein coding regions from DOGMA annotated plastomes, aligned individual gene with Mafft v. 7.0.5(-auto flag) (Katoh & Standley, 2013), trimmed alignment gaps using trimAl v.1.2 (-automated1 flag) (Capella-Gutiérrez *et al*., 2009) and finally generated a concatenated alignment of all genes. The pipeline, Plast2phy, written in Python, is available at [https://github.com/saemi/plast2phy](https://github.com/saemi/plast2phy). Model of base substitution were tested for the concatenated matrix using jModelTest v.2.1.1 (Guindon & Gascuel, 2003; Darriba *et al*., 2012). Using the Akaike information criterion (AIC), we determined the GTR+G+I model optimal for the concatenated plastome alignment. We analysed the dataset under maximum likelihood (ML; Felsenstein, 1973) using GARLI (Zwickl, 2006). We ran GARLI v. 2.0 with default settings, using ten independent searches and 100 bootstrap replicates. Bootstrap consensus was calculated using SumTrees v. 3.3.1 in the DendroPy package (Sukumaran & Holder, 2010). Trees from phylogenetic analysis were drawn using FigTree v.1.4.0 ([http://tree.bio.ed.ac.uk/software/figtree/](http://tree.bio.ed.ac.uk/software/figtree/)), rooted with *Cicer aretinum*.

**Identification of locally collinear blocks (LCBs) in plastid genomes**

The progressive alignment method, implemented in the MAUVE v.2.3.1 package (Darling *et al*., 2010), was used with the default parameters to identify locally collinear blocks (LCBs) among the plastid genomes listed in Table 1. In this study, a LCB represents a region within a plastid genome that can occur in different orientation and/or order among the studied plastid
genomes, but is free from any internal rearrangements (see Darling et al., 2010). These regions are therefore putatively orthologous in nature. I used two programs, projectAndStrip and makeBadgerMatrix (downloaded from http://gel.ahabs.wisc.edu/mauve/snapshots/, on 11 November 2014) to generate a LCB boundary file from the MAUVE alignment. The LCB boundary file contained information on where the LCBs start and end in each of the analysed plastome. I used the chickpea plastome (Cicer arietinum) [NCBI Reference Sequence NC_011163] (Jansen et al., 2008), as the reference plastome. We visualized the observed plastid rearrangements using Circos v.0.66 (Krzywinski et al., 2009), were a custom Python script was to generate Circos input files and executing the program for each of plastid genome (Figs. 1 and 2). Figure 2 was generated by manually combining all the Circos maps with a phylogenetic cladogram using inkscape (www.inkscape.org). Information regarding the size of LCBs and lengths of protein coding genes in Table 2 are based on the reference plastome of Cicer arietinum [NC_011163]. Putative unannotated tRNAs were identified using MITOS (Bernt et al., 2013).

Relationship among divergence time and plastid rearrangements

In order to investigate the relationship among species divergence and the number plastid rearrangements, we estimated two relevant parameters in a pairwise manner. Firstly we estimated the reversal distance using GRIMM v. 2.0.1 (Tesler, 2003). Reversal distance is the minimum number of reversal steps for two genomes to become completely syntenic (Tesler, 2003). GRIMM uses the LCB boundary file, described in the previous section, as its input file. Secondly we used the synonymous substitution rate (Ks) between pairs of plastid genomes as an indicator for the divergence between species. Ks values were calculated using MEGA v.6.0 (Tamura et al., 2013), from a concatenated alignment file of the plastid protein genes. Divergence times can be estimated using published estimates of plastid mutation rates, which range from 1.1 – 2.9 silent substitutions per billion years (Wolfe et al., 1987). The reversal distance was plotted against divergence time using R v. 3.0.2 (R Core Team, 2014) and ggplot2 (http://ggplot2.org/), and their relationship visualized using a smoothing curve, using the following command: geom_smooth(degree=1, shape=2/3, method='loess', level=.95).
Results

Phylogeny of species with rearranged plastomes

Mapping the plastome architecture onto the phylogenetic tree of the investigated legume species indicates that the plastomes have clearly undergone multiple multiple and rounds of inversions and translocations throughout the tree (Figs. 1a, 2a and 2b.) The phylogenetic analysis of the protein coding regions of our completely sequenced plastomes proved useful in resolving the relationships among the studied species (Fig. 2a.). The phylogeny reported here largely confirms previous studies (LPWG 2013) and is consistent with the Trifolium phylogeny using the same methods reported previously (Sveinsson & Cronk 2014) and Fabae (previously reported in Magee et al., 2010). The tree is rooted on Cicer, which along with Medicago, shows no evidence of plastid rearrangements compared to other IRLC legumes (Supporting Information 1) and, with the exception of the lost inverted repeat (Wojciechowski et al., 2000), their plastomes are collinear with Lotus japonicus (Supporting Information 1, Fig. 3).

Many locally collinear blocks (LCBs) correspond to previously reported plastid operons

The entire plastid genome in these species has been broken up multiple times by rearrangement but certain genomic blocks have never been broken up. MAUVE identified a total of 32 localized collinear blocks (LCB) (Fig. 2b) in the 23 analysed plastomes. Out of these 32 LCBs, 26 contained protein-coding genes and one LCB was made up of the plastid rRNA genes (see Table 2). These 26 blocks varied in size and contained gene clusters (GCs) that varied in the number of genes that they encompass (see Table 2). Nine of the blocks (gene clusters) contained only a single gene, five blocks were composed of two genes and the remaining 12 blocks consisted of more than 2 genes. The largest gene cluster (GC) is GC-1, 13.8 kb in length, containing the following genes: rpl32, ndhF, psbA, matK, rbcL, atpE and atpB (Table 2). The smallest gene cluster detected was GC-8, about 1.2 kb in length, containing only a single gene, petN. Many of the gene clusters have previously been recognized as plastid operons (i.e. transcriptional units), such as GC-2, 6, 7, 18, 27 and 31 (see Sugita & Sugiura, 1996). Several other gene clusters share extensive similarities with previously reported plastid operons but can differ in the presence or absence of a single gene, e.g. GC-11, 12 and 21. Gene containing LCBs cover about 98% of the total length of the
This suggests that the delimitation of these clusters is not random and is under functional constraint (see discussion).

The number of plastome rearrangements increases with divergence time, but levels off

We investigated the relationship between species divergence and the number of plastome rearrangements, by plotting sequence evolution against genome rearrangement. Specifically we plotted pairwise synonymous substitution rate (Ks) against pairwise reversal distances (see Fig. 3). The rationale for this analysis was that if formation of new LCBs in the plastomes is constrained by the presence of plastid operons, blocks should increase in number until functional constraint does not allow further break-up of blocks. We find a pattern consistent with this constraint hypothesis. With evolutionary distance (approximating time) rearrangements increase until saturation is reached. When a smoothing curve is fitted through the data, we observe what seems to a strong positive correlation between evolutionary divergence and plastome rearrangements up until about Ks~0.10 where it starts to level off, which relates to about 9 in reversal distance (Fig. 3). If the plastomes were under no functional constraint there would be no obvious reason that the relationship between divergence and reversal distance would level off at that point. Our results suggest that there is functional constraint on the observed plastome rearrangements and its most likely source is the preservation of functional di- and polycistronic plastid transcriptional units (see discussion). This functional constraint appears to place a limit on the number LCBs (Fig. 3), i.e. limit the extent to which blocks of genes can be broken up. The block cannot be further divided, by inversion and translocation, without breaking co-transcriptional units.

Discussion

The rearrangements of plastomes in IRLC legumes

Plastid genomes of analysed Trifolium and the Fabaceae (Lens, Vicia, Pisum and Lathyrus) species are highly rearranged, as a result of multiple rounds of translocations and/or inversions (Fig. 2 and Fig. 2). These rearrangements have previously been reported (Palmer & Thompson, 1982; Cai et al., 2008; Magee et al., 2010; Sabir et al., 2014; Sveinsson & Cronk, 2014). Plastid genomes tend to be quite conserved structurally across land plants (see
However, besides these IRLC legumes, there are other well-known exceptions, such as Geraniaceae (Guisinger et al., 2011) and Campanulaceae (Haberle et al., 2008). The plastome rearrangements described here for the Fabae appear to be most similar to those reported in *Trachelium caeruleum* (Campanulaceae), since they do not involve proliferation of repeated elements, such as in certain *Trifolium* species (Sveinsson & Cronk, 2014) or in the Geraniaceae (Guisinger et al., 2011; Weng et al., 2013). The functional cause of these rearrangements is not known. The stability of plastid genomes is maintained through recombinational mechanisms, which are controlled by a large number of nuclear genes (see Maréchal & Brisson, 2010). The loss of the inverted repeat may be involved in the genome instability but, if so, the relationship is not simple as some IRLC legumes such as *Medicago* and *Cicer* have well conserved typical legume gene orders. Whatever the functional changes that result from this unprecedented genome instability, it is clear that it offers a unique opportunity for the study the organization of inviolable transcriptional units within the plastid genomes of flowering plants. Against a background of extensive genome scrambling, blocks with conserved gene order stand out.

Do conserved blocks in otherwise rearranged plastomes represent operons?

The genespace of plastid genomes is organized into transcriptional units, similar to operons in the genome of their cyanobacterial ancestors (Stern et al., 2010). However it is important to note that despite plastids being of bacterial origin, most aspects of the regulation of plastid gene expression are radically different from bacteria, mainly due to interactions with the nuclear genome (Stern et al., 2010). Nevertheless, it is well established from functional studies that many plastid genes are organized into dicistronic or polycistronic operon-like units, i.e. co-regulated gene blocks, also known as transcriptional units (Sugita & Sugiura, 1996). It is therefore reasonable to assume that any structural rearrangements that would break up these transcriptional units would be detrimental to the plastid and be selected against.

Our results are in agreement with that assumption, as many of the gene clusters that we observe are known plastid polycistronic operons (Table 2 here; and Table 2 in Sugita & Sugiura, 1996). Examples of this are: (i) Gene Cluster 21 (GC-21, Table 2) that seems to correspond to the *psbB* operon, which has been extensively studied (Stoppel & Meurer, 2013); (ii) GC-7 which contains the same genes as the *psbD/C/Z* operon, which has been
characterized in tobacco (Adachi et al., 2012); (iii) GC-28 which contains all the rRNA genes, which are necessary to construct the plastid 70S ribosome. The numerous genes that are not associated with any other and freely translocate independently, are likely to represent single gene transcriptional units, i.e. monocistronic operons. Gene Cluster 1 (rpl32-ndhF-psbA-matK-rbcL-atpB-atpE) is of particular interest, as it contains seven cistrons that previously were thought to be transcribed independently (i.e. as monocistronic units) or belong to different operons (see Table 2 in Sugita & Sugiura, 1996). Our results are highly suggestive that that GC-1 is a conserved plastid operon, at least in the legume species analysed here. Six LCBs without any annotated protein-coding or RNA genes were also identified (varying in size between 177 and 689 nt (see Table 2; Supporting Information 2). However, we identified putative unannotated tRNAs in these blocks (Table 2) and so it is possible that they too are under functional constraint.

These results demonstrate that identification of conserved gene clusters in this clade of rapid structural evolution is a powerful way of provide evidence for previously described plastid operons and potentially to find new ones. Such is the extent of the genic reorganization in the sampled species that it may be argued that the persistence of multiple intact gene blocks is implausible unless these units (Table 2) are inviolable as they represent the fundamental regulatory architecture of the legume plastid.

Acknowledgements

We thank C. Hefer with computer help, D. Kaplan for greenhouse assistance, the Western Canada Research Grid (Westgrid) for computer facilities, USDA and the Desert Legume project for seeds, A. Kuzmin for assistance with sequencing library prep. The work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants Program (grant no. RGPIN-2014-05820 grant to QC).
References


Zwickl DJ. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, University of Texas, Austin, TX, USA.
Table 1. Information regarding the plastid genomes used in this study. Details regarding the Illumina sequencing and voucher details are presented where applicable. An asterisk (*) indicates species sequenced in this study.

<table>
<thead>
<tr>
<th>Species (USDA seed accession)</th>
<th>GenBank accession</th>
<th>Herbarium voucher (UBC²)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cicer arietinum</em> L. (NA)</td>
<td>NC_011163</td>
<td>NA</td>
</tr>
<tr>
<td><em>Medicago truncatula</em> Gaertn. (NA)</td>
<td>NC_003119</td>
<td>NA</td>
</tr>
<tr>
<td><em>Trifolium strictum</em> L. (PI 369147)</td>
<td>KJ788292</td>
<td>V241491</td>
</tr>
<tr>
<td><em>T. grandiflorum</em> Schreb. (NA)</td>
<td>NC_024034</td>
<td>NA</td>
</tr>
<tr>
<td><em>T. aureum</em> Pollich (NA)</td>
<td>NC_024035</td>
<td>NA</td>
</tr>
<tr>
<td><em>T. boisseri</em> Guss. (PI 369022)</td>
<td>KJ788284</td>
<td>V241490</td>
</tr>
<tr>
<td><em>T. glanduliferum</em> Boiss. (PI 296666)</td>
<td>KJ788285</td>
<td>V241492</td>
</tr>
<tr>
<td><em>Vicia sativa</em> L. (PI 293436)</td>
<td>KJ850242</td>
<td>NA</td>
</tr>
<tr>
<td><em>Lens culinaris</em> Medik. (PI 592998)</td>
<td>KJ850239</td>
<td>NA</td>
</tr>
<tr>
<td><em>Pisum sativum</em> L. (W6 32866)</td>
<td>KJ806203</td>
<td>V241498</td>
</tr>
<tr>
<td><em>Lathyrus clymenum</em> L. (RP³)</td>
<td>KJ850235</td>
<td>V241501</td>
</tr>
<tr>
<td><em>L. tingitanus</em> L. (RP³)</td>
<td>KJ850238</td>
<td>V241502</td>
</tr>
<tr>
<td><em>L. sativus</em> L. (NA)</td>
<td>NC_014063</td>
<td>NA</td>
</tr>
<tr>
<td><em>L. odoratus</em> L. (RP³)</td>
<td>KJ850237</td>
<td>V241503</td>
</tr>
<tr>
<td>Species</td>
<td>Accession</td>
<td>Location</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------</td>
<td>---------------------</td>
</tr>
<tr>
<td>*L. pubescens Hook. &amp; Arn.</td>
<td>KJ806200</td>
<td>V241505</td>
</tr>
<tr>
<td>(RP^3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*L. inconspicuus L. (W6 2817)</td>
<td>KJ850236</td>
<td>V241504</td>
</tr>
<tr>
<td>*L. davidii Hance (RP^3)</td>
<td>KJ806192</td>
<td>V241506</td>
</tr>
<tr>
<td>*L. palustris L. (NA)</td>
<td>KJ806198</td>
<td>V241511</td>
</tr>
<tr>
<td>*L. japonicus Willd.(NA)</td>
<td>KJ806194</td>
<td>NA</td>
</tr>
<tr>
<td>*L. littoralis Endl.(NA)</td>
<td>KJ806197</td>
<td>NA</td>
</tr>
<tr>
<td>*L. graminifolius (S.Watson)</td>
<td>KJ806193</td>
<td>V241507</td>
</tr>
<tr>
<td>T.G.White (DLP^4 accession:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>920239)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*L. ochroleucus Hook. (NA)</td>
<td>KJ806199</td>
<td>V241489</td>
</tr>
<tr>
<td>*L. venosus Muhl. ex Willd.</td>
<td>KJ806202</td>
<td>V241509</td>
</tr>
<tr>
<td>(NA)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 SD: Standard Deviation
2 UBC Herbarium, Vancouver BC Canada
3 Roger Parsons Sweet Peas, Chichester UK
4 Desert Legume Project Germplasm, AZ USA
Table 2. Details regarding the gene content, positional boundaries, length of locally collinear blocks identified within the analysed plastome. An asterisk (*) marks putative tRNAs unannotated in the *Cicer* plastome.

<table>
<thead>
<tr>
<th>LCB ID</th>
<th>Genes residing in the LCB</th>
<th>LCB boundary(^2)</th>
<th>LCB length (coding %(^1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC-01</td>
<td>rpl32-ndhF-psbA-matK-rbcL-atpB-atpE</td>
<td>121898 - 125304 and 1 - 10457</td>
<td>13,862 (60.0)</td>
</tr>
<tr>
<td>GC-02</td>
<td>ndhC-ndhK-ndhJ</td>
<td>10458 - 13156</td>
<td>1,524 (56.5)</td>
</tr>
<tr>
<td>GC-03</td>
<td>rps4</td>
<td>13157 - 15877</td>
<td>606 (22.3)</td>
</tr>
<tr>
<td>GC-04</td>
<td>ycf3</td>
<td>15878 - 18242</td>
<td>1,977 (83.6)</td>
</tr>
<tr>
<td>GC-05</td>
<td>*trnE, *trnI</td>
<td>18243 - 18420</td>
<td>177 (-)</td>
</tr>
<tr>
<td>GC-06</td>
<td>psaA-psaB-rps14</td>
<td>18421 - 24458</td>
<td>4,761 (78.9)</td>
</tr>
<tr>
<td>GC-07</td>
<td>psbZ-psbC-psbD-psbM</td>
<td>24459 - 31398</td>
<td>2,778 (40.0)</td>
</tr>
<tr>
<td>GC-08</td>
<td>petN</td>
<td>31399 - 32561</td>
<td>96 (8.3)</td>
</tr>
<tr>
<td>GC-09</td>
<td>*trnQ, *trnF</td>
<td>32562 - 32875</td>
<td>313 (-)</td>
</tr>
<tr>
<td>GC-11</td>
<td>rpoB-rpoC1</td>
<td>33398 - 40310</td>
<td>6,050 (87.5)</td>
</tr>
<tr>
<td>GC-12</td>
<td>rpoC2-rps2-atp1-atpH-atpF-atpA</td>
<td>40311 - 51303</td>
<td>8,586 (78.1)</td>
</tr>
<tr>
<td>GC-13</td>
<td>*trnQ, *rrnL, *trnE, *rrnL</td>
<td>51304 - 51635</td>
<td>331 (-)</td>
</tr>
<tr>
<td>GC-14</td>
<td>psbI-psbK</td>
<td>51636 - 54492</td>
<td>297 (10.4)</td>
</tr>
<tr>
<td>GC-15</td>
<td>accD-psa1</td>
<td>54493 - 58824</td>
<td>1,488 (34.4)</td>
</tr>
<tr>
<td>GC-16</td>
<td>cemA</td>
<td>58825 - 60169</td>
<td>690 (51.3)</td>
</tr>
<tr>
<td>GC-17</td>
<td>petA</td>
<td>60170 - 61921</td>
<td>963 (55.0)</td>
</tr>
<tr>
<td>GC-18</td>
<td>psbJ-psbL-psbF-psbE</td>
<td>61922 - 63839</td>
<td>612 (31.9)</td>
</tr>
<tr>
<td>GC-19</td>
<td>petL-petG-psa3-rpl33-rps18-rpl20</td>
<td>63840 - 67436</td>
<td>1,236 (34.4)</td>
</tr>
<tr>
<td>GC-20</td>
<td>5'-rps12-clpP</td>
<td>67437 - 69967</td>
<td>1,578 (52.2)</td>
</tr>
<tr>
<td>GC-21</td>
<td>psbB-psbT-psbN-psbH-petB-petD</td>
<td>69968 - 75752</td>
<td>4,667 (80.7)</td>
</tr>
<tr>
<td>GC-22</td>
<td>rpoA-rps11-rpl36-rps8</td>
<td>75753 - 79071</td>
<td>1,932 (58.2)</td>
</tr>
<tr>
<td>GC-23</td>
<td>rpl14-rpl16-rps3-rps19-rpl2-rpl23</td>
<td>79072 - 85227</td>
<td>4,567 (74.2)</td>
</tr>
<tr>
<td>GC-24</td>
<td>ycf2</td>
<td>85228 - 91848</td>
<td>5,772 (87.2)</td>
</tr>
<tr>
<td>GC-26</td>
<td>ndhB</td>
<td>92153 - 95069</td>
<td>2,225 (76.3)</td>
</tr>
<tr>
<td>GC-27</td>
<td>3'-rps12-rps7</td>
<td>95070 - 96289</td>
<td>468 (29.1)</td>
</tr>
<tr>
<td>GC-28</td>
<td>16S-23S-4.5S-5S ribosomal RNA</td>
<td>96290 - 106075</td>
<td>4,524 (46.2)</td>
</tr>
<tr>
<td>GC-29</td>
<td>ycf1-rps15-ndhH-ndhA-ndh1-ndhG</td>
<td>106076 - 117316</td>
<td>10,216 (90.9)</td>
</tr>
<tr>
<td>GC-30</td>
<td>ndhE-psaC-ndhD</td>
<td>117317 - 119928</td>
<td>2,046 (78.4)</td>
</tr>
<tr>
<td>GC-31</td>
<td>ccsA</td>
<td>119929 - 121207</td>
<td>972 (76.1)</td>
</tr>
</tbody>
</table>

1 Coding percentage (%) was calculated as the proportion of gene length compared to the total length of the LCB.

2 The plastid genome of *Cicer arietinum* was used as a reference.
Table 3. Reversal distance matrix, produced by GRIMM, between each major plastotypes (see Fig. 2b).

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>12</td>
<td>9</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>6</td>
<td>56</td>
<td>30</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Figure legends

Figure 1 – Example of rearrangements and reversal distances between plastid genomes. A comparison of the order and orientation of locally collinear blocks (LCBs) among four species: *Cicer arietinum*, *Lens culinaris*, *Trifolium strictum* and *Lathyrus clymenum*. LCBs are represented as differently coloured boxes and the numbers refer to the LCB IDs in Table 2. Inverted LCBs are positioned on an inner circle. Reversal distances (see Material and Methods) between species is shown on arrowed lines. Two plastid inversion events between *C. arietinum* and *L. culinaris* are highlighted.

Figure 2 – Plastid rearrangements of the investigated species shown in a phylogenetic context. (a) A cladogram showing the phylogenetic relationship among the species in this study. The phylogeny was constructed from a concatenated alignment of plastid protein coding genes. All nodes except one were retrieved with a 100% bootstrap support. (b) Visual representation of the rearranged plastid genomes of the species in this study. Locally collinear blocks (LCBs) are represented as coloured boxes. The letter in the middle of each circle refers to the major plastotype of each species, shown after the species names in parenthesis in the cladogram (a).

Figure 3 – The relationship between the pairwise species divergence and the reversal distance of the plastid genome analysed in this study. Synonymous substitutions rates (Ks) of plastid protein coding genes are used as a proxy for species divergence time, and are shown on the x-axis. The reversal distance (extent of rearrangements) is shown on the y-axis. The black line is a smooth curve that illustrates the relationships among these two variables. 95% confidence intervals are represented as grey areas around the line.