

1 **Conserved gene clusters in the scrambled plastomes of IRLC legumes (Fabaceae:**  
2 **Trifolieae and Fabeae).**

3

4 Saemundur Sveinsson<sup>1\*</sup>, Quentin Cronk<sup>2</sup>

5

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

9

10 Author for correspondence:

11 *Saemundur Sveinsson*

12 *Tel: 354-4335219*

13 *Email: saemundur@lbhi.is*

14

Total word count (excluding summary, references and legends):	2884	No. of figures:	3 (Figs 1 and 2 in colour)
Summary:	179	No. of Tabela:	3
Introduction:	397	No. of Supporting Information files:	0
Materials and Methods:	1038		
Results:	663		
Discussions:	719		
Acknowledgements:	61		

15

16

17      Summary

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

33

34      Key words (5-8): Fabaeae, IRLC, massively parallel sequencing , plastid operons, plastome  
35      evolution, plastome rearrangements, Trifolieae.

36

## 37 **Introduction**

38 The plastid genome, also known as the plastome, refers to the total genetic information of a  
39 single plant organelle, the plastid, which takes many developmental forms, the most notable  
40 being the chloroplast (Bock, 2007). Plastid genomes are circular structures of double stranded  
41 DNA, usually consisting of about 100-120 genes and are around 120-160 kb long in  
42 photosynthesizing plants (Bock, 2007). Their size, structure and gene content are highly  
43 conserved across land plants (Wicke *et al.*, 2011). However there are exceptions, such as the  
44 Geraniaceae and Campanulaceae, which are two angiosperm families known to contain  
45 species with highly rearranged plastomes (Haberle *et al.*, 2008; Guisinger *et al.*, 2011). A  
46 dominating feature of plastid genomes is the presence of a large inverted repeat (Wicke *et al.*,  
47 2011; Zhu *et al.*, 2015) separated by a small single copy region that is variable in orientation  
48 (Walker *et al.* 2015). However some plant groups have lost one copy of the repeat, one being  
49 a clade within papilionoid legumes (Fabaceae), known as the inverted repeat lost clade  
50 (IRLC) (Wojciechowski *et al.*, 2000).

51 Plants obtained their plastid organelles through an endosymbiosis event with a  
52 cyanobacteria-like organism, about 1.5 – 1.6 billion years ago (Margulis, 1970; Hedges *et al.*,  
53 2004). Its bacterial origin gives the plastid genome many prokaryotic features, such as small  
54 (70S) ribosomes and the absence of mRNA 3' polyA tails (see Stern *et al.*, 2010 for a  
55 review). An additional ancestral feature of the plastid genome is the organization of its  
56 coding region into multiple gene clusters, or operons (Sugita & Sugiura, 1996; Sugiura *et al.*,  
57 1998). These gene clusters are stretches of the plastome consisting of several genes that are  
58 transcribed into di- or polycistronic units, which are then processed before translation (Stern  
59 *et al.*, 2010). Several such clusters have already been identified in the plastid (Adachi *et al.*,  
60 2012; Ghulam *et al.*, 2013; Stoppel & Meurer, 2013).

61 Several legume genera within the IRLC are known to harbour highly rearranged  
62 plastomes, as a result from multiple translocations and/or inversions: *Trifolium* (Cai *et al.*,  
63 2008; Sabir *et al.*, 2014; Sveinsson & Cronk, 2014), *Pisum* (Palmer & Thompson, 1982),  
64 *Lathyrus* (Magee *et al.*, 2010), *Lens* and *Vicia* (Sabir *et al.*, 2014). The aim of this study is to  
65 analyse these rearrangements in these genera within IRLC, in order to investigate whether  
66 they can be used to study the organization of plastid genomes into operons.

## 67 **Material and Methods**

### 68 **Source of plant material**

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

76

### 77 **Illumina sequencing**

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

87

### 88 **Plastid genome assemblies and annotation**

89 Trimmomatic v.0.3 (Lohse *et al.*, 2012) was used to trim and remove low quality Illumina  
90 reads, with the following flags: LEADING:20 TRAILING:20  
91 SLIDINGWINDOW:4:15MINLEN:36. High quality reads were used in all subsequent  
92 analysis and singlet reads, i.e. reads without a paired end, were discarded. We used the *de*  
93 *novo* method implemented in CLC Genomic Workbench v.7.0.2 to generate assemblies for  
94 each species, using the default settings. Contigs of plastid origin were identified by a  
95 BLASTN search (Altschul *et al.*, 1997) to a plastid genome of a closely related species.  
96 These were generally the largest and most high coverage contigs in the *de novo* assembly and  
97 always had an E-value of 0 when blasted to the reference plastome. Regions with nucleotides

98 scored as Ns were manually resolved by retrieving sequence information directly from the  
99 quality-trimmed reads. For most species, the *de novo* assembly returned a single large plastid  
100 contig. When needed, multiple contigs containing plastid sequence were joined by hand,  
101 using information from the quality-trimmed reads. The quality of each plastome assembly  
102 was verified by visually by inspecting a BWA mem pileup, v. 0.7.5a (Li & Durbin, 2009), of  
103 paired end reads using Tablet v.1.13.12.17 (Milne *et al.*, 2013). We made sure that the  
104 connections between manually joined contigs were supported by paired-end read mapping.  
105 Finally all plastome assemblies were annotated using DOGMA (Wyman *et al.*, 2004).

106

### 107 **Phylogenetic analysis**

108 Due to the extensive rearrangements observed in the plastomes (see Sabir *et al.*, 2014;  
109 Sveinsson & Cronk, 2014), we restricted our plastome phylogenetic analysis to protein  
110 coding genes. We used a custom phylogenetic pipeline, plast2phy, that extracted protein  
111 coding regions from DOGMA annotated plastomes, aligned individual gene with Mafft v.  
112 7.0.5(-auto flag) (Kato & Standley, 2013), trimmed alignment gaps using trimAl v.1.2 (-  
113 automated1 flag) (Capella-Gutiérrez *et al.*, 2009) and finally generated a concatenated  
114 alignment of all genes. The pipeline, Plast2phy, written in Python, is available at  
115 <https://github.com/saemi/plast2phy>. Model of base substitution were tested for the  
116 concatenated matrix using jModelTest v.2.1.1 (Guindon & Gascuel, 2003; Darriba *et al.*,  
117 2012). Using the Akaike information criterion (AIC), we determined the GTR+G+I model  
118 optimal for the concatenated plastome alignment. We analysed the dataset under maximum  
119 likelihood (ML; Felsenstein, 1973) using GARLI (Zwickl, 2006). We ran GARLI v. 2.0 with  
120 default settings, using ten independent searches and 100 bootstrap replicates. Bootstrap  
121 consensus was calculated using SumTrees v. 3.3.1 in the DendroPy package (Sukumaran &  
122 Holder, 2010). Trees from phylogenetic analysis were drawn using FigTree v.1.4.0  
123 (<http://tree.bio.ed.ac.uk/software/figtree/>), rooted with *Cicer aretinum*.

124

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

126 The progressive alignment method, implemented in the MAUVE v.2.3.1 package (Darling *et al.*,  
127 2010), was used with the default parameters to identify locally collinear blocks (LCBs)  
128 among the plastid genomes listed in Table 1. In this study, a LCB represents a region within a  
129 plastid genome that can occur in different orientation and/or order among the studied plastid

130 genomes, but is free from any internal rearrangements (see Darling *et al.*, 2010). These  
131 regions are therefore putatively orthologous in nature. I used two programs, projectAndStrip  
132 and makeBadgerMatrix (downloaded from <http://gel.ahabs.wisc.edu/mauve/snapshots/>, on 11  
133 November 2014) to generate a LCB boundary file from the MAUVE alignment. The LCB  
134 boundary file contained information on where the LCBs start and end in each of the analysed  
135 plastome. I used the chickpea plastome (*Cicer arietinum*) [NCBI Reference Sequence  
136 NC\_011163] (Jansen *et al.*, 2008), as the reference plastome. We visualized the observed  
137 plastid rearrangements using Circos v.0.66 (Krzywinski *et al.*, 2009), were a custom Python  
138 script was to generate Circos input files and executing the program for each of plastid  
139 genome (Figs. 1 and 2). Figure 2 was generated by manually combining all the Circos maps  
140 with a phylogenetic cladogram using inkscape ([www.inkscape.org](http://www.inkscape.org)). Information regarding  
141 the size of LCBs and lengths of protein coding genes in Table 2 are based on the reference  
142 plastome of *Cicer arietinum* [NC\_011163]. Putative unannotated tRNAs were identified  
143 using MITOS (Bernt *et al.*, 2013).

144

#### 145 **Relationship among divergence time and plastid rearrangements**

146 In order to investigate the relationship among species divergence and the number plastid  
147 rearrangements, we estimated two relevant parameters in a pairwise manner. Firstly we  
148 estimated the reversal distance using GRIMM v. 2.0.1 (Tesler, 2003). Reversal distance is the  
149 minimum number of reversal steps for two genomes to become completely syntenic (Tesler,  
150 2003). GRIMM uses the LCB boundary file, described in the previous section, as its input  
151 file. Secondly we used the synonymous substitution rate (Ks) between pairs of plastid  
152 genomes as an indicator for the divergence between species. Ks values were calculated using  
153 MEGA v.6.0 (Tamura *et al.*, 2013), from a concatenated alignment file of the plastid protein  
154 genes. Divergence times can be estimated using published estimates of plastid mutation rates,  
155 which range from 1.1 – 2.9 silent substitutions per billion years (Wolfe *et al.*, 1987). The  
156 reversal distance was plotted against divergence time using R v. 3.0.2 (R Core Team, 2014)  
157 and ggplot2 (<http://ggplot2.org/>), and their relationship visualized using a smoothing curve,  
158 using the following command: `geom_smooth(degree=1, shape=2/3, method='loess',`  
159 `level=.95)`.

160

## 161 **Results**

### 162 **Phylogeny of species with rearranged plastomes**

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

175

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

177 The entire plastid genome in these species has been broken up multiple times by  
178 rearrangement but certain genomic blocks have never been broken up. MAUVE identified a  
179 total of 32 localized collinear blocks (LCB) (Fig. 2b) in the 23 analysed plastomes. Out of  
180 these 32 LCBs, 26 contained protein-coding genes and one LCB was made up of the plastid  
181 rRNA genes (see Table 2). These 26 blocks varied in size and contained gene clusters (GCs)  
182 that varied in the number of genes that they encompass (see Table 2). Nine of the blocks  
183 (gene clusters) contained only a single gene, five blocks were composed of two genes and the  
184 remaining 12 blocks consisted of more than 2 genes. The largest gene cluster (GC) is GC-1,  
185 13.8 kb in length, containing the following genes: *rpl32*, *ndhF*, *psbA*, *matK*, *rbcL*, *atpE* and  
186 *atpB* (Table 2). The smallest gene cluster detected was GC-8, about 1.2 kb in length,  
187 containing only a single gene, *petN*. Many of the gene clusters have previously been  
188 recognized as plastid operons (i.e. transcriptional units), such as GC-2, 6, 7, 18, 27 and 31  
189 (see Sugita & Sugiura, 1996). Several other gene clusters share extensive similarities with  
190 previously reported plastid operons but can differ in the presence or absence of a single gene,  
191 e.g. GC-11, 12 and 21. Gene containing LCBs cover about 98% of the total length of the

192 *Cicer arietinum* plastome. This suggests that the delimitation of these clusters is not random  
193 and is under functional constraint (see discussion).

194

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

214

## 215 **Discussion**

### 216 **The rearrangements of plastomes in IRLC legumes**

217 Plastid genomes of analysed *Trifolium* and the Fabae (*Lens*, *Vicia*, *Pisum* and *Lathyrus*)  
218 species are highly rearranged, as a result of multiple rounds of translocations and/or  
219 inversions (Fig. 2 and Fig. 2). These rearrangements have previously been reported (Palmer  
220 & Thompson, 1982; Cai *et al.*, 2008; Magee *et al.*, 2010; Sabir *et al.*, 2014; Sveinsson &  
221 Cronk, 2014). Plastid genomes tend to be quite conserved structurally across land plants (see

222 Wicke *et al.*, 2011). However, besides these IRLC legumes, there are other well-known  
223 exceptions, such as Geraniaceae (Guisinger *et al.*, 2011) and Campanulaceae (Haberle *et al.*,  
224 2008). The plastome rearrangements described here for the Fabaeae appear to be most similar  
225 to those reported in *Trachelium caeruleum* (Campanulaceae), since they do not involve  
226 proliferation of repeated elements, such as in certain *Trifolium* species (Sveinsson & Cronk,  
227 2014) or in the Geraniaceae (Guisinger *et al.*, 2011; Weng *et al.*, 2013). The functional cause  
228 of these rearrangements is not known. The stability of plastid genomes is maintained through  
229 recombinational mechanisms, which are controlled by a large number of nuclear genes (see  
230 Maréchal & Brisson, 2010). The loss of the inverted repeat may be involved in the genome  
231 instability but, if so, the relationship is not simple as some IRLC legumes such as *Medicago*  
232 and *Cicer* have well conserved typical legume gene orders. Whatever the functional changes  
233 that result from this unprecedented genome instability, it is clear that it offers a unique  
234 opportunity for the study the organization of inviolable transcriptional units within the plastid  
235 genomes of flowering plants. Against a background of extensive genome scrambling, blocks  
236 with conserved gene order stand out.

237

### 238 **Do conserved blocks in otherwise rearranged plastomes represent operons?**

239 The genespace of plastid genomes is organized into transcriptional units, similar to operons  
240 in the genome of their cyanobacterial ancestors (Stern *et al.*, 2010). However it is important  
241 to note that despite plastids being of bacterial origin, most aspects of the regulation of plastid  
242 gene expression are radically different from bacteria, mainly due to interactions with the  
243 nuclear genome (Stern *et al.*, 2010). Nevertheless, it is well established from functional  
244 studies that many plastid genes are organized into dicistronic or polycistronic operon-like  
245 units, i.e. co-regulated gene blocks, also known as transcriptional units (Sugita & Sugiura,  
246 1996). It is therefore reasonable to assume that any structural rearrangements that would  
247 break up these transcriptional units would be detrimental to the plastid and be selected  
248 against.

249 Our results are in agreement with that assumption, as many of the gene clusters that we  
250 observe are known plastid polycistronic operons (Table 2 here; and Table 2 in Sugita &  
251 Sugiura, 1996). Examples of this are: (i) Gene Cluster 21 (GC-21, Table 2) that seems to  
252 correspond to the *psbB* operon, which has been extensively studied (Stoppel & Meurer,  
253 2013); (ii) GC-7 which contains the same genes as the *psbD/C/Z* operon, which has been

254 characterized in tobacco (Adachi *et al.*, 2012); (iii) GC-28 which contains all the rRNA  
255 genes, which are necessary to construct the plastid 70S ribosome. The numerous genes that  
256 are not associated with any other and freely translocate independently, are likely to represent  
257 single gene transcriptional units, i.e. monocistronic operons. Gene Cluster 1 (*rpl32-ndhF-*  
258 *psbA-matK-rbcL-atpB-atpE*) is of particular interest, as it contains seven cistrons that  
259 previously were thought to be transcribed independently (i.e. as monocistronic units) or  
260 belong to different operons (see Table 2 in Sugita & Sugiura, 1996). Our results are highly  
261 suggestive that that GC-1 is a conserved plastid operon, at least in the legume species  
262 analysed here. Six LCBs without any annotated protein-coding or RNA genes were also  
263 identified (varying in size between 177 and 689 nt (see Table 2; Supporting Information 2)).  
264 However, we identified putative unannotated tRNAs in these blocks (Table 2) and so it is  
265 possible that they too are under functional constraint.

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

272

## 273 **Acknowledgements**

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

279

280 **References**

- 281 **Adachi Y, Kuroda H, Yukawa Y, Sugiura M. 2012.** Translation of partially overlapping  
282 *psbD-psbC* mRNAs in chloroplasts: the role of 5'-processing and translational coupling.  
283 *Nucleic Acids Research* **40**: 3152–8.
- 284 **Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.**  
285 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.  
286 *Nucleic Acids Research* **25**: 3389–3402.
- 287 **Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritsch G, Pütz J,**  
288 **Middendorf M, Stadler PF. 2013.** MITOS: Improved de novo Metazoan Mitochondrial  
289 Genome Annotation. *Molecular Phylogenetics and Evolution* **69**: 313-319.
- 290 **Bock R. 2007.** Structure, function, and inheritance of plastid genomes. In: Bock R, editor.  
291 *Cell and Molecular Biology of Plastids*. Berlin, Germany: Springer, 29–63.
- 292 **Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl J V, Boore J,**  
293 **Jansen RK. 2008.** Extensive reorganization of the plastid genome of *Trifolium subterraneum*  
294 (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions.  
295 *Journal of Molecular Evolution* **67**: 696–704.
- 296 **Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009.** trimAl: a tool for automated  
297 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.
- 298 **Darling AE, Mau B, Perna NT. 2010.** Progressivemauve: Multiple genome alignment with  
299 gene gain, loss and rearrangement. *PLoS ONE* **5**. e11147.
- 300 **Darriba D, Taboada GL, Doallo R, Posada D. 2012.** jModelTest 2: more models, new  
301 heuristics and parallel computing. *Nature Methods* **9**: 772.
- 302 **Doyle JJ, Doyle JL. 1987.** A rapid DNA isolation procedure for small quantities of fresh leaf  
303 tissue. *Phytochemical Bulletin* **19**: 11–15.
- 304 **Felsenstein J. 1973.** Maximum likelihood and minimum-steps methods for estimating  
305 evolutionary trees from data on discrete characters. *Systematic Zoology* **22**: 240–249.

- 306 **Ghulam MM, Courtois F, Lerbs-Mache S, Merendino L. 2013.** Complex processing  
307 patterns of mRNAs of the large ATP synthase operon in *Arabidopsis* chloroplasts. *PLoS ONE*  
308 **8:** e78265.
- 309 **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, Adiconis X, Fan**  
310 **L, Raychowdhury R, Zeng Q, et al. 2011.** Full-length transcriptome assembly from RNA-  
311 Seq data without a reference genome. *Nature Biotechnology* **29:** 644–652.
- 312 **Guindon S, Gascuel O. 2003.** A simple, fast, and accurate algorithm to estimate large  
313 phylogenies by maximum likelihood. *Systematic Biology* **52:** 696–704.
- 314 **Guisinger MM, Kuehl J V, Boore JL, Jansen RK. 2011.** Extreme reconfiguration of  
315 plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon  
316 usage. *Molecular Biology and Evolution* **28:** 583–600.
- 317 **Haberle RC, Fourcade HM, Boore JL, Jansen RK. 2008.** Extensive rearrangements in the  
318 chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes.  
319 *Journal of Molecular Evolution* **66:** 350–361.
- 320 **Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004.** A molecular timescale of eukaryote  
321 evolution and the rise of complex multicellular life. *BMC Evolutionary Biology* **4:** 2.
- 322 **Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. 2008.** Complete plastid  
323 genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of  
324 *rps12* and *clpP* intron losses among legumes (*Leguminosae*). *Molecular Phylogenetics and*  
325 *Evolution* **48:** 1204–1217.
- 326 **Katoh K, Standley DM. 2013.** MAFFT Multiple Sequence Alignment Software Version 7:  
327 Improvements in Performance and Usability. *Molecular Biology and Evolution* **30:** 772–780.
- 328 **Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra**  
329 **MA. 2009.** Circos: an information aesthetic for comparative genomics. *Genome Research* **19:**  
330 1639–1645.
- 331 **LPWG (The Legume Phylogeny Working Group). 2013.** Legume phylogeny and  
332 classification in the 21st century : Progress, prospects and lessons for other species-rich  
333 clades. *Taxon* **62:** 217–248.
- 334 **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler  
335 transform. *Bioinformatics* **25:** 1754–1760.

- 336 **Li L, Stoeckert CJ, Roos DS. 2003.** OrthoMCL: identification of ortholog groups for  
337 eukaryotic genomes. *Genome Research* **13**: 2178–2189.
- 338 **Liu L, Yu L, Pearl DK, Edwards S V. 2009.** Estimating species phylogenies using  
339 coalescence times among sequences. *Systematic Biology* **58**: 468–477.
- 340 **Liu L, Yu L. 2010.** Phybase: an R package for species tree analysis. *Bioinformatics* **26**: 962–  
341 963.
- 342 **Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. 2012.** RobiNA: a  
343 user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids*  
344 *Research* **40**: W622–627.
- 345 **Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S,**  
346 **Milbourne D, Barth S, Palmer JD, et al. 2010.** Localized hypermutation and associated  
347 gene losses in legume chloroplast genomes. *Genome Research* **20**: 1700–1710.
- 348 **Maréchal A, Brisson N. 2010.** Recombination and the maintenance of plant organelle  
349 genome stability. *The New Phytologist* **186**: 299–317.
- 350 **Margulis L. 1979.** Origin of Eukaryotic Cells. New Haven, CT, USA: Yale University Press.
- 351 **Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D.**  
352 **2013.** Using Tablet for visual exploration of second-generation sequencing data. *Briefings in*  
353 *Bioinformatics* **14**: 193–202.
- 354 **Palmer JD, Thompson WF. 1982.** Chloroplast DNA rearrangements are more frequent  
355 when a large inverted repeat sequence is lost. *Cell* **29**: 537–550.
- 356 **R Core Team. 2014.** R: A language and environment for statistical computing. R Foundation  
357 for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. [WWW  
358 document] URL <http://www.R-project.org/> [accessed 23 October 2014].
- 359 **Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, Jansen R, Ruhlman**  
360 **T. 2014.** Evolutionary and biotechnology implications of plastid genome variation in the  
361 inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal* **12**: 743–754.
- 362 **Stamatakis A. 2014.** RAxML version 8: A tool for phylogenetic analysis and post-analysis  
363 of large phylogenies. *Bioinformatics* **30**: 1312–1313.

- 364 **Stern DB, Goldschmidt-Clermont M, Hanson MR. 2010.** Chloroplast RNA metabolism.  
365 *Annual Review of Plant Biology* **61**: 125–155.
- 366 **Stoppel R, Meurer J. 2013.** Complex RNA metabolism in the chloroplast: an update on the  
367 *psbB* operon. *Planta* **237**: 441–449.
- 368 **Sugita M, Sugiura M. 1996.** Regulation of gene expression in chloroplasts of higher plants.  
369 *Plant Molecular Biology* **32**: 315–326.
- 370 **Sugiura M, Hirose T, Sugita M. 1998.** Evolution and mechanism of translation in  
371 chloroplasts. *Annual Review of Genetics* **32**: 437–459.
- 372 **Sukumaran J, Holder MT. 2010.** DendroPy: a Python library for phylogenetic computing.  
373 *Bioinformatics* **26**: 1569–1571.
- 374 **Sveinsson S, McDill J, Wong GKS, Li J, Li X, Deyholos MK, Cronk QCB. 2014.**  
375 Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using  
376 transcriptomics. *Annals of Botany* **113**: 753–761.
- 377 **Sveinsson S, Cronk Q. 2014.** Evolutionary origin of highly repetitive plastid genomes within  
378 the clover genus (*Trifolium*). *BMC Evolutionary biology* **14**: 228.
- 379 **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013.** MEGA6: Molecular  
380 Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**: 2725–  
381 2729.
- 382 **Tesler G. 2003.** Efficient algorithms for multichromosomal genome rearrangements. *Journal*  
383 *of Computer and System Sciences* **65**: 587–609.
- 384 **Walker JF, Jansen RK, Zanis MJ, Emery NC. 2015.** Sources of inversion variation in the  
385 small single copy (SSC) region of chloroplast genomes. *American Journal of Botany*.
- 386 **Wasmuth JD, Blaxter ML. 2004.** prot4EST: translating expressed sequence tags from  
387 neglected genomes. *BMC Bioinformatics* **5**: 187.
- 388 **Weng M-L, Blazier JC, Govindu M, Jansen RK. 2013.** Reconstruction of the ancestral  
389 plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats  
390 and nucleotide substitution rates. *Molecular Biology and Evolution*: 1–15.

391 **Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011.** The evolution  
392 of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant*  
393 *Molecular Biology* **76**: 273–297.

394 **Wojciechowski, M. F., M. J. Sanderson, K. P. Steele, and A. Liston. 2000.** Molecular  
395 phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: a supertree  
396 approach. In: Herendeen PS, Bruneau A, eds. *Advances in Legume Systematics*. Kew, UK:  
397 Royal Botanic Gardens, 277-298.

398 **Wolfe KH, Li WH, Sharp PM. 1987.** Rates of nucleotide substitution vary greatly among  
399 plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of*  
400 *Sciences of the United States of America* **84**: 9054–9058.

401 **Wyman SK, Jansen RK, Boore JL. 2004.** Automatic annotation of organellar genomes with  
402 DOGMA. *Bioinformatics* **20**: 3252–3255.

403 **Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2015.** Evolutionary dynamics of the plastid  
404 inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New*  
405 *Phytologist*.

406 **Zwickl DJ. 2006.** *Genetic algorithm approaches for the phylogenetic analysis of large*  
407 *biological sequence datasets under the maximum likelihood criterion*. PhD thesis, University  
408 of Texas, Austin, TX, USA.

409

410

436 Table 1. Information regarding the plastid genomes used in this study. Details regarding the  
 437 Illumina sequencing and voucher details are presented where applicable. An asterisk (\*)  
 438 indicates species sequenced in this study.

<b>Species (USDA seed accession)</b>	<b>GenBank accession</b>	<b>Herbarium voucher (UBC<sup>2</sup>)</b>
<i>Cicer arietinum</i> L. (NA)	NC_011163	NA
<i>Medicago truncatula</i> Gaertn. (NA)	NC_003119	NA
* <i>Trifolium strictum</i> L. (PI 369147)	KJ788292	V241491
<i>T. grandiflorum</i> Schreb. (NA)	NC_024034	NA
<i>T. aureum</i> Pollich (NA)	NC_024035	NA
* <i>T. boisseri</i> Guss. (PI 369022)	KJ788284	V241490
* <i>T. glanduliferum</i> Boiss. (PI 296666)	KJ788285	V241492
* <i>Vicia sativa</i> L. (PI 293436)	KJ850242	NA
* <i>Lens culinaris</i> Medik. (PI 592998)	KJ850239	NA
* <i>Pisum sativum</i> L. (W6 32866)	KJ806203	V241498
* <i>Lathyrus clymenum</i> L. (RP <sup>3</sup> )	KJ850235	V241501
* <i>L. tingitanus</i> L. (RP <sup>3</sup> )	KJ850238	V241502
<i>L. sativus</i> L. (NA)	NC_014063	NA
* <i>L. odoratus</i> L. (RP <sup>3</sup> )	KJ850237	V241503

* <i>L. pubescens</i> Hook. & Arn. (RP <sup>3</sup> )	KJ806200	V241505
* <i>L. inconspicuus</i> L. (W6 2817)	KJ850236	V241504
* <i>L. davidii</i> Hance (RP <sup>3</sup> )	KJ806192	V241506
* <i>L. palustris</i> L. (NA)	KJ806198	V241511
* <i>L. japonicus</i> Willd.(NA)	KJ806194	NA
* <i>L. littoralis</i> Endl.(NA)	KJ806197	NA
* <i>L. graminifolius</i> (S.Watson) T.G.White (DLP <sup>4</sup> accession: 920239)	KJ806193	V241507
* <i>L. ochroleucus</i> Hook.  (NA)	KJ806199	V241489
* <i>L. venosus</i> Muhl. ex Willd.  (NA)	KJ806202	V241509

439 <sup>1</sup>SD: Standard Deviation

440 <sup>2</sup>UBC Herbarium, Vancouver BC Canada

441 <sup>3</sup>Roger Parsons Sweet Peas, Chichester UK

442 <sup>4</sup>Desert Legume Project Germplasm, AZ USA

443

444

445 Table 2. Details regarding the gene content, positional boundaries, length of locally collinear  
 446 blocks identified within the analysed plastome. An asterisk (\*) marks putative tRNAs  
 447 unannotated in the *Cicer* plastome.

LCB ID	Genes residing in the LCB	LCB boundary <sup>2</sup>	LCB length (coding % <sup>1</sup> )
GC-01	<i>rpl32-ndhF-psbA-matK-rbcL-atpB-atpE</i>	121898 - 125304 and 1 - 10457	13,862 (60.0)
GC-02	<i>ndhC-ndhK-ndhJ</i>	10458 - 13156	1,524 (56.5)
GC-03	<i>rps4</i>	13157 - 15877	606 (22.3)
GC-04	<i>ycf3</i>	15878 - 18242	1,977 (83.6)
GC-05	* <i>trnE</i> , * <i>trnI</i>	18243 - 18420	177 (-)
GC-06	<i>psaA-psaB-rps14</i>	18421 - 24458	4,761 (78.9)
GC-07	<i>psbZ-psbC-psbD-psbM</i>	24459 - 31398	2,778 (40.0)
GC-08	<i>petN</i>	31399 - 32561	96 (8.3)
GC-09	* <i>trnQ</i> , * <i>trnF</i>	32562 - 32875	313 (-)
GC-10	* <i>trnC</i> , * <i>trnF</i> , * <i>trnK</i> , * <i>trnW</i> , * <i>trnV</i>	32876 - 33397	521 (-)
GC-11	<i>rpoB-rpoC1</i>	33398 - 40310	6,050 (87.5)
GC-12	<i>rpoC2-rps2-atpI-atpH-atpF-atpA</i>	40311 - 51303	8,586 (78.1)
GC-13	* <i>trnQ</i> , * <i>rrnL</i> , * <i>trnE</i> , * <i>rrnL</i>	51304 - 51635	331 (-)
GC-14	<i>psbI-psbK</i>	51636 - 54492	297 (10.4)
GC-15	<i>accD-psaI</i>	54493 - 58824	1,488 (34.4)
GC-16	<i>cemA</i>	58825 - 60169	690 (51.3)
GC-17	<i>petA</i>	60170 - 61921	963 (55.0)
GC-18	<i>psbJ-psbL-psbF-psbE</i>	61922 - 63839	612 (31.9)
GC-19	<i>petL-petG-psaJ-rpl33-rps18-rpl20</i>	63840 - 67436	1,236 (34.4)

GC-20	<i>5'rps12-clpP</i>	67437 - 69967	1,578 (52.2)
GC-21	<i>psbB-psbT-psbN-psbH-petB-petD</i>	69968 - 75752	4,667 (80.7)
GC-22	<i>rpoA-rps11-rpl36-rps8</i>	75753 - 79071	1,932 (58.2)
GC-23	<i>rpl14-rpl16-rps3-rps19-rpl2-rpl23</i>	79072 - 85227	4,567 (74.2)
GC-24	<i>ycf2</i>	85228 - 91848	5,772 (87.2)
GC-25	<i>*trnE, *trnL2, *trnM</i>	91849 - 92152	303 (-)
GC-26	<i>ndhB</i>	92153 - 95069	2,225 (76.3)
GC-27	<i>3'rps12-rps7</i>	95070 - 96289	468 (29.1)
GC-28	<i>16S-23S-4.5S-5S ribosomal RNA</i>	96290 - 106075	4,524 (46.2)
GC-29	<i>ycf1-rps15-ndhH-ndhA-ndhI-ndhG</i>	106076 - 117316	10,216 (90.9)
GC-30	<i>ndhE-psaC-ndhD</i>	117317 - 119928	2,046 (78.4)
GC-31	<i>ccsA</i>	119929 - 121207	972 (76.1)
GC-32	<i>*trnM, *trnL1, *trnQ, *trnW</i>	121208 - 121897	689 (-)

448 <sup>1</sup>Coding percentage (%) was calculated as the proportion of gene length compared to the total  
 449 length of the LCB.

450 <sup>2</sup>The plastid genome of *Cicer arietinum* was used as a reference.

451

452 Table 3. Reversal distance matrix, produced by GRIMM, between each major plastotypes  
453 (see Fig. 2b).

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

454

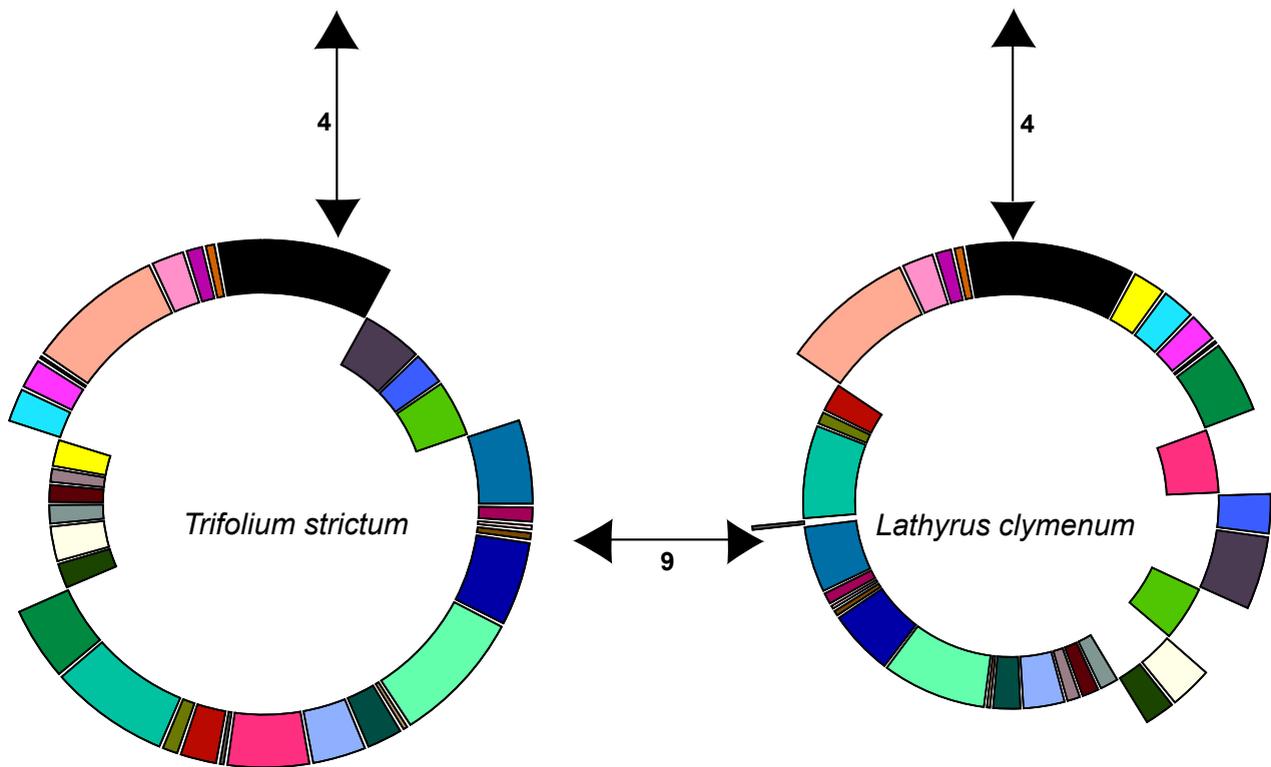
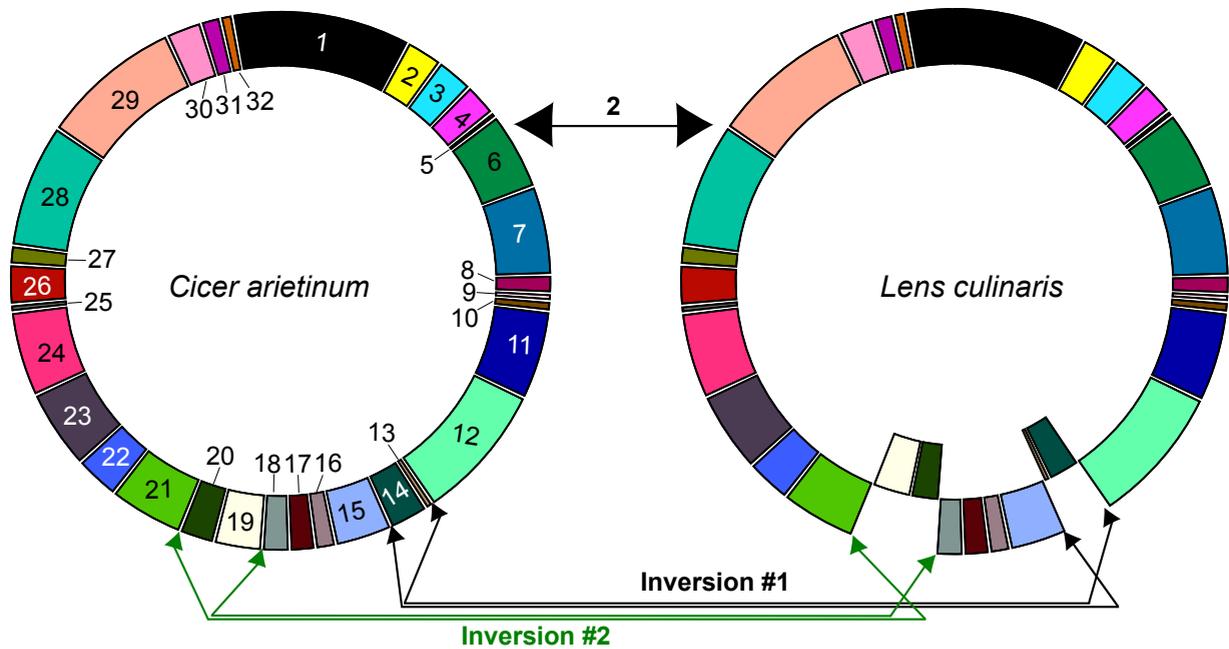
## 411 **Figure legends**

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

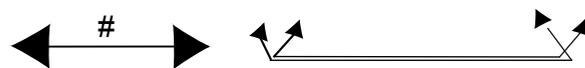
419  
420 Figure 2 – Plastid rearrangements of the investigated species shown in a phylogenetic  
421 context. (a) A cladogram showing the phylogenetic relationship among the species in this  
422 study. The phylogeny was constructed from a concatenated alignment of plastid protein  
423 coding genes. All nodes except one were retrieved with a 100% bootstrap support. (b) Visual  
424 representation of the rearranged plastid genomes of the species in this study. Locally  
425 collinear blocks (LCBs) are represented as coloured boxes. The letter in the middle of each  
426 circle refers to the major plastotype of each species, shown after the species names in  
427 parenthesis in the cladogram (a).

428  
429 Figure 3 – The relationship between the pairwise species divergence and the reversal distance  
430 of the plastid genome analysed in this study. Synonymous substitutions rates (Ks) of plastid  
431 protein coding genes are used as a proxy for species divergence time, and are shown on the x-  
432 axis. The reversal distance (extent of rearrangements) is shown on the y-axis. The black line  
433 is a smooth curve that illustrates the relationships among these two variables. 95%  
434 confidence intervals are represented as grey areas around the line.

435



**Legend**



Reversal distance

Inversion event

LCB

