Localized hypermutation and associated gene losses in legume chloroplast genomes

Alan M. Magee,1 Sue Aspinall,2 Danny W. Rice,3 Brian P. Cusack,1 Marie Sémon,4 Antoinette S. Perry,1 Saša Stefanović,5 Dan Milbourne,6 Susanne Barth,6 Jeffrey D. Palmer,3 John C. Gray,2 Tony A. Kavanagh,1 and Kenneth H. Wolfe1,7

1Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland; 2Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom; 3Department of Biology, Indiana University, Bloomington, Indiana 47405, USA; 4Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, CNRS, INRA, UCB Lyon 1, Ecole Normale Supérieure de Lyon, 69364 Lyon Cedex 07, France; 5Department of Biology, University of Toronto, Mississauga, Ontario L5L 1C6, Canada; 6Teagasc Crops Research Centre, Oak Park, Carlow, Ireland

The genome organization and gene content of chloroplast DNA (cpDNA) are highly conserved among most flowering plant species (Palmer 1985; Sugiyama 1992; Jansen et al. 2007). The chloroplast genome of the most recent common ancestor of all angiosperms contained 113 different genes (four rRNA genes, 30 tRNA genes, and 79 protein genes), and this content has been retained in many angiosperms (Kim and Lee 2004). Rates of synonymous nucleotide substitution in chloroplast genes are generally low (a few fold lower than plant nuclear genes) and relatively homogeneous within a genome except for a threefold difference in rate between the large inverted repeat (IR) and single-copy regions (Wolfe et al. 1998). Lineage-specific variation in the chloroplast genome in legumes has been reported for rpl22 in legumes (Gantt et al. 1991); for infA in several lineages, including almost all rosids (Millet et al. 2001); and for rpl32 in two families of Malpighiales (Cusack and Wolfe 2007; Ueda et al. 2005). In addition, Ueda et al. (2008) identified gene substitution as the mechanism of loss of the rps16 gene from cpDNA in Medicago and Populus. The loss of rps16 from cpDNA is compensated by dual targeting (to chloroplasts as well as mitochondria) of mitochondrial ribosomal protein S16, which is encoded by a nuclear gene. Several other examples of losses of genes from cpDNA in photosynthetic angiosperms have been reported, and it is striking that the few species in which gene losses have occurred tend also to be those whose chloroplast genomes are highly rearranged relative to the ancestral angiosperm organization (Jansen et al. 2007). As with angiosperm mitochondrial genomes (Adams and Palmer 2003), most of the genes that have been lost from chloroplast genomes during recent evolution have coded for ribosomal proteins (Jansen et al. 2007). There have been no published reports of the loss of genes coding for components of photosystems I or II (psa genes), the electron transfer chain (pet genes), or the chloroplast ATP synthase (atp genes) from cpDNA in any angiosperms except parasitic species (Wolfe et al. 1992; Funk et al. 2007; McNeal et al. 2007).

One group of angiosperms that is known to be relatively prone to cpDNA rearrangement and gene loss is the legume family
(Fabaceae) (Palmer et al. 1988). The large IR that is otherwise almost universally present in chloroplast genomes is absent from one large clade of legumes (the IR loss clade, or IRLC) (Wojciechowski et al. 2004), some of which also show other rearrangements of gene order. Chloroplast genomes in the IRLC species are also notable for having significant amounts of repetitive DNA, something not usually seen in angiosperm cpDNA (Milligan et al. 1989; Saski et al. 2005; Cai et al. 2008). Five instances of gene loss from the IRLC chloroplast genomes have been discovered. As well as the aforementioned gene transfers of rpl22 and infA and substitution of tps16 (Gantt et al. 1991; Millen et al. 2001; Ueda et al. 2008), it has been reported that that accD is completely absent from Trifolium subterraneum (subclower; Cai et al. 2008). Slot-blot hybridization experiments suggested that ycf4 and rps16 may have been lost independently multiple times in different lineages of legumes (Doyle et al. 1995).

In the course of this study, we reviewed all reported instances (in published papers or in GenBank annotations) of gene loss among the 103 complete angiosperm chloroplast genome sequences that are publicly available, and found that 27 different protein-coding genes have been lost in at least one lineage (Supplemental Table S1). We found that some reported gene losses are simply due to annotation errors; because of this, the numbers of losses we describe here are slightly different from those in Jansen et al. (2007). In particular, we noticed that the gene ycf4, which was originally not identified in the genome sequences of the legumes Glycine max (soybean; Saski et al. 2005), T. subterraneum (subclower; Cai et al. 2008), Cicer arietinum (chickpea), and Medicago truncatula (Jansen et al. 2008), is in fact present in the cpDNAs of all these species but is so divergent that it was not recognized by the DOGMA software (Wyman et al. 2004) used to annotate them. This discovery prompted us to investigate the rapid evolution of ycf4 and its surrounding region in legumes.

Ycf4 is a thylakoid protein that has been shown to play a role in regulating photosystem I assembly in cyanobacteria (Wilde et al. 1995) and to be essential for photosystem I assembly in Chlamydomonas (Boudreau et al. 1997; Onishi and Takahashi 2009). Experiments in Chlamydomonas indicate that Ycf4 is the second of three scaffold proteins that act sequentially during the assembly process, with Ycf4's roles being to stabilize an intermediate subcomplex consisting of the PsaaB heterodimer and the three stromal subunits PsacDE, and to add the Psaf subunit to this subcomplex (Ozawa et al. 2009). As well as the loss of ycf4 in P. sativum, several other previous studies have indicated that the evolution of ycf4 in legumes may be unusual. In soybean and Lotus japonicus, the Ycf4 protein, which is almost universally 184 or 185 amino acids long, has expanded to about 200 residues (Reverdatto et al. 1995; Kato et al. 2000). The gene also has a high rate of synonymous nucleotide substitution between the two species (Perry and Wolfe 2002). Phylogenetic trees for phaseoloid legumes constructed using ycf4 were incongruent with trees constructed using seven other genes, due to accelerated evolution of codon positions 1 and 2 in ycf4 (Stefanovic et al. 2009). In blot hybridizations to DNAs from 280 diverse angiosperms (as in Millen et al. 2001) using a ycf4 probe from tobacco, we observed (SS and JP, unpubl.) strong hybridization to all DNAs except those from the only Papilionoid legumes examined: Medicago (no signal from five species) and Vigna (considerably diminished signal). We show here that ycf4 is situated in a local mutation hotspot, in Lathyrus, and possibly in other legume species, resulting in dramatic acceleration of sequence evolution in some species and evolutionary gene losses in others.

**Results**

**Rapid evolution of ycf4 in legumes**

To investigate acceleration of the evolutionary rate of ycf4 in legumes, we compared its nonsynonymous and synonymous nucleotide substitution rates in different angiosperm lineages to the rates observed in two other, widely sequenced chloroplast genes, rbcL and matK. This analysis included new ycf4 sequence data from Lathyrus and other legumes, together with sequences from a previous phylogenetic study (GenBank [http://www.ncbi.nlm.nih.gov/Genbank/] accession nos. EU717431–EU717464; Stefanovic et al. 2009) and other database sequences. For each gene, we used a likelihood model to estimate the numbers of nonsynonymous (ds) and synonymous (ds) nucleotide substitutions that occurred on each branch of an angiosperm phylogenetic tree (see Methods). In the ds trees, ycf4 is seen to evolve much faster in most legumes than in other angiosperms (Fig. 1) but no similar acceleration is seen in legume rbcL or matK, which suggests that the acceleration is locus-specific, as well as lineage-specific. Within legumes, the first accelerated branch is the one leading to a large clade (Milletioid, Robinioid, and the IRLC; asterisk in Fig. 1), and the legumes that are outgroups to this branch do not show acceleration. This branch is also the first one on which the Ycf4 protein size expands above 200 amino acids (Fig. 1). Even faster periods of ds evolution are seen in the genera Desmodium and Lathyrus relative to other legumes. Ycf4 is a pseudogene in three of six Desmodium species we sequenced and in Clitoria ternatea (Supplemental Fig. S1C, left panel). In the de trees, some acceleration is seen in ycf4 of legumes relative to other angiosperms, particularly in Lathyrus, but again no similar acceleration is seen in legume rbcL or matK (Fig. 1). The genus Lathyrus also shows by far the greatest increases in Ycf4 size, reaching 340 residues in Lathyrus latifolius and Lathyrus clyclorhous.

Remarkably, there is less amino acid sequence conservation between the Ycf4 proteins of two species within the genus Lathyrus (31% identity between Lathyrus palustris and L. clyclorhous), than between tobacco and the cyanobacterium Synecocystis (45% identity). Nevertheless, ycf4 can be inferred to be functional in the four Lathyrus species in which it is intact (Fig. 2), for two reasons. First, even though the level of amino acid sequence conservation among Lathyrus species is very low, many of the sites in the C-terminal part of the protein (beginning at position 248 in Fig. 2) that are conserved among other land plants and cyanobacteria are also conserved in Lathyrus. Second, comparing ycf4 sequences among Lathyrus species shows that they have lower levels of nonsynonymous than synonymous nucleotide substitutions (ds/ ds < 1) (Table 1), which is a hallmark of sequences that are being constrained to code for proteins (Kimura 1977; Graur and Li 1999). We therefore infer that these long ycf4 genes in Lathyrus species are biologically functional. However, the level of constraint on Lathyrus ycf4 is lower than on other angiosperm ycf4s (e.g., de/ds = 0.15 between tobacco and spinach ycf4, compared with de/ds = 0.36–0.81 within the genus Lathyrus). Tests for positive (Darwinian) selection suggested that some Desmodium branches within the ycf4 tree have undergone adaptive evolution, and in separate analyses, site-specific tests for positive selection were significant for some codons in ycf4 when the whole legume tree was considered (data not shown). However, in view of the evidence that the whole region around ycf4 has a high mutation rate (see below), and
because we also found some $d_s/d_k$ values greater than 1 within the genus *Lathyrus* for two genes flanking *ycf4* (*cemA* and *accD* (Table 1), we suspect that the high $d_s/d_k$ values are artifacts stemming from a combination of an increased mutation rate and lessened constraints on protein sequences, rather than being indicative of positive selection on multiple adjacent genes.

We may have slightly overestimated the divergence of *Lathyrus* Ycf4 proteins because we inferred protein sequences from chloroplast DNA sequences, whereas some chloroplast transcripts are known to undergo mRNA editing (Stern et al. 2010). Editing in angiosperms involves C → U changes and typically occurs at 30–40 sites per genome (Tsudzuki et al. 2001; Inada et al. 2004). However, even extensive C → U editing could only marginally reduce the divergence in *Lathyrus* Ycf4. For example, if we assume that every possible C → U editing event that could increase the similarity between *L. palustris* and *L. cirrhosus* Ycf4 proteins actually occurs, their sequence identity only increases from 31% to 32%. Furthermore, no sites in *ycf4* are known to undergo mRNA editing in other species (Tsudzuki et al. 2001; Chateigner-Boutin and Small 2007; and our analyses of EST data from *M. truncatula*, *Lotus japonicus*, and *G. max*).

**Gene losses and repetitive DNA in legumes**

We sequenced the region flanking the *ycf4* locus in five *Lathyrus* species, *P. sativum* (pea) and *Vicia faba* (broad bean) and compared it to the available data for other legumes (Fig. 3). This comparison reveals a history of multiple gene losses and gene length changes within a small region of cpDNA. We identified *ycf4* pseudogenes in both *P. sativum* and *Lathyrus odoratus* (sweet pea), which must be the result of two separate losses of the gene (Fig. 3). The small photosystem I gene *psaI*, normally found immediately upstream of *ycf4*, is missing from a clade of four *Lathyrus* species but is present in *L. palustris*. Also in this region of the genome, the ribosomal protein gene *rps16* was lost from cpDNA in the common ancestor of the IRLC clade (Doyle et al. 1995), and *accD*, coding for a subunit of acetyl-CoA carboxylase, is missing from *T. subterraneum* cpDNA, which has become rearranged in this region (Cai et al. 2008). Both *ycf4* and *accD* show extensive length variation among the legume species that retain them (Fig. 3).

The expansion of the *accD* open reading frame is partly explained by the presence of numerous tandemly repeated sequences in this region of legume cpDNA. As reported previously (Nagano et al. 1991a; Smith et al. 1991), and shown by a dot-matrix plot in Supplemental Figure S2A, *P. sativum* *accD* contains several in-frame internal repeats of up to 37 codons long. *L. sativus* *accD* has a similarly repetitive structure, but the sections of the gene that are repeated are different in the two species (Supplemental Fig. S2B,C). There are tandem repeats in the intergenic DNA between *accD* and *ycf4* in *L. latifolius* (Supplemental Fig. S2D), and a tandem repeat of 15 codons is located within the 5′ end of *L. sativus* *ycf4* (Supplemental Fig. S2B). All the repeats are species-specific, which suggests that these minisatellite-like sequences have a high turnover rate. However, some other species, such as *L. odoratus*, do not contain tandem repeats in this region, and the expanded size of *ycf4* in most *Lathyrus* species is not primarily due to the accumulation of repeats.

**Sequences of the *P. sativum* and *L. sativus* chloroplast genomes**

To establish whether the patterns of evolution seen around the *ycf4* locus are atypical of the rest of the genome, we sequenced the chloroplast genome of *L. sativus* (grass pea; 121,020 bp) and completed the genome sequence of *P. sativum* cpDNA (pea; 122,169 bp).

![Image](https://example.com/image.jpg)
Both of these genomes lack the IR. They have rearrangements of gene order relative to the ancestral angiosperm order, as represented by tobacco, and also relative to each other. The gene order in *P. sativum* can be obtained from the tobacco order by eight inversion steps (Palmer et al. 1988), beginning with a 50-kb inversion that is shared by most legumes and that placed *rps16* beside *accD*. The first three inversions occurred before the separation of the lineages giving rise to *P. sativum* and *L. sativus*, after which there were five more inversions specific to *P. sativum*, and three more inversions specific to *L. sativus* (Supplemental Fig. S3). None of the inversions in *P. sativum* or *L. sativus* is shared with the highly rearranged cpDNA of *T. subterraneum* (Cai et al. 2008), other than the initial 50-kb inversion (Supplemental Fig. S3E).

The *L. sativus* genome sequence shows that it shares four gene losses that have already been reported in *P. sativum*: *infA*, *rps16*, *rpl22*, and *rpl23* (Gantt et al. 1991; Nagano et al. 1991a,b; Millen et al. 2001); whereas *L. sativus ycf4* is intact. The status of *rpl23* in *P. sativum* has been unclear because it contains a 190-bp

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

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. palustris vs. L. latifolius</td>
<td></td>
</tr>
<tr>
<td><em>ycf4</em></td>
<td>109</td>
</tr>
<tr>
<td><em>accD</em></td>
<td>112</td>
</tr>
<tr>
<td><em>cemA</em></td>
<td>150</td>
</tr>
<tr>
<td><em>matK</em></td>
<td>176</td>
</tr>
<tr>
<td>L. palustris vs. L. sativus</td>
<td></td>
</tr>
<tr>
<td><em>rbcL</em></td>
<td>200</td>
</tr>
<tr>
<td>L. odoratus vs. L. latifolius</td>
<td></td>
</tr>
<tr>
<td><em>rcbI</em>-<em>atpB</em> spacer</td>
<td>747</td>
</tr>
<tr>
<td><em>trnF</em>-<em>trnL</em> spacer</td>
<td>475</td>
</tr>
<tr>
<td><em>trnS</em>-<em>trnG</em> spacer</td>
<td>616</td>
</tr>
<tr>
<td>L. cirrhosus vs. L. latifolius</td>
<td></td>
</tr>
<tr>
<td><em>ycf4</em></td>
<td>109</td>
</tr>
<tr>
<td><em>accD</em></td>
<td>112</td>
</tr>
<tr>
<td><em>cemA</em></td>
<td>150</td>
</tr>
<tr>
<td><em>matK</em></td>
<td>176</td>
</tr>
<tr>
<td>Sequence</td>
<td>Sites</td>
</tr>
<tr>
<td><em>rbcL</em>-<em>atpB</em> spacer</td>
<td>747</td>
</tr>
<tr>
<td><em>trnF</em>-<em>trnL</em> spacer</td>
<td>475</td>
</tr>
<tr>
<td><em>trnS</em>-<em>trnG</em> spacer</td>
<td>616</td>
</tr>
</tbody>
</table>

For protein-coding genes the synonymous divergence (*d_S*), its standard error (SE), and the nonsynonymous-to-synonymous ratio (*d_N/d_S*, also called ω) are shown. For intergenic regions, divergence (K) was calculated using Kimura’s two-parameter method. NA, Not applicable (gene not present); ND, not determined.

*Average number of sites compared across the reported species pairs. The *ycf4*, *accD*, *cemA*, and *rbcL* comparisons are all not full-length. For *ycf4*, only the relatively conserved section between position 164 in Figure 2 and the C terminus was compared.

*Sequence data from Kenicer et al. (2005).*
program (Yang 2007). For most loci, the divergence between these species is less than 0.1 substitutions per site (median \( d_S = 0.055 \)).

Sites of gene loss coincide with a mutation hotspot

We measured synonymous divergence in each protein-coding gene between the \( P. \) sativum and \( L. \) sativus chloroplast genomes using \( d_S \) (black circle symbols in Fig. 4), calculated by the yn00
chloroplast genomes, we wondered how these rates compared with the
rate in the nuclear genome. Relatively few nuclear genes have
been sequenced from Lathyrus species, so we generated new
expressed sequence tag (EST) data from Lathyrus odoratus (sweet-
pea; see below) and identified putatively orthologous nuclear
genes between these and database sequences from P. sativum.
Among 56 putative orthologs, the median $d_S$ is 0.131 (Supplemental
Table S2), which is 2.4 times higher than the median $d_S$ (0.055) for
chloroplast genes compared between P. sativum and L. sativus. Thus
in comparisons between Lathyrus and P. sativum, as in other flow-
ering plant comparisons, the synonymous divergence in most parts
of the chloroplast genome is lower than in the nuclear genome. The
synonymous divergence in ycf4, however, is at least 10 times greater
than in the nuclear genome (the ratios of the $d_S$ values given above,
$1.084/0.131 = 8.3$ and $1.522/0.131 = 11.6$, are underestimates
of the actual ratio because the numerators involve a shorter di-
vergence time).

Transfer of Trifolium accD to the nucleus

We suspect that ycf4 and psaI have been transferred to the nuclear
genome in the Lathyrus species that lack them in cpDNA, because
these species are fully photosynthetic and must have a functional
photosystem I. However, we were unable to find nuclear copies of
these genes. We made numerous unsuccessful attempts (see
Methods) to amplify ycf4 and psaI by PCR from genomic DNA of
L. odoratus (which lacks both of them in its cpDNA) and has a
smaller ycf4 pseudogene than P. sativum). We then made cDNA
from young green leaves of L. odoratus and sequenced 8702 ESTs.
None of the ESTs were derived from a nuclear ycf4 or psaI, even though we did find ESTs corresponding to seven of the nine other
nuclear-encoded subunits of photosystem I (Jolley et al. 2005), and
to the older nuclear-transferred genes infA and rpl22. Psal is a very
small protein (34–40 amino acids) that is conserved between cya-
nobacteria and land plants and is physically located toward the
exterior of photosystem I in P. sativum, where it interacts strongly
with PsaH (Jolley et al. 2005; Amunts et al. 2007). It seems unlikely
that photosystem I in Lathyrus could function efficiently without
Psal, although tobacco plants with a psaI knockout do not show
a mutant phenotype under standard growth conditions (MA
Schöttler and R Bock, pers. comm.). Most of the small membrane-
spanning subunits of photosystem I appear to be nonessential, and
knockout lines do not display visible mutant phenotypes (Varotto
et al. 2002; Jensen et al. 2007; Schöttler et al. 2007). However,
the loss of individual small membrane-spanning subunits usu-
ally affects the assembly of other subunits and results in lower

**Figure 4.** Sequence divergence between the P. sativum and L. sativus chloroplast genomes. The x-axis lists genes or exons in the order in which they occur in the L. sativus genome. Black filled circles show $d_S$ (number of synonymous substitutions per synonymous site) for each orthologous protein gene pair, calculated using yn00 (Yang 2007). White and gray filled circles show divergence ($K$) for each intergenic region or intron, respectively, calculated by Kimura’s two-parameter method (Kimura 1983). Vertical bars, $d_S$ or $K$ ± 1 SE. Because ycf4 is a pseudogene in P. sativum and psaI is not present in L. sativus, the $d_S$ value plotted for ycf4 is for a comparison between L. sativus and L. palustris, and the $d_S$ value plotted for psaI is for a comparison between P. sativum and L. palustris (see text). No divergence values are plotted for intergenic regions that are not flanked by the same genes in the two species or that are shorter than 100 bp.

**Figure 5.** Sequence divergence between L. latifolius and L. ciri
cosus in the accD-ycf4-cemA region (left) and the atpB-rbcL region (right). Vertical tickmarks indicate the locations of each nucleotide substitution, catego-
rized according to whether it occurs at codon position 1, 2, or 3; or in
intergenic DNA; and as a transversion (Tv; tickmarks above the horizontal lines) or a transition (Tt; tickmarks below the horizontal line). The total
numbers of each type of substitution are shown on the right. Supple-
mental Figure S5 shows the nucleotide sequence alignment summarized
in the left panel.
efficiencies of excitation transfer and electron transfer (Varotto et al. 2002; Jensen et al. 2007; Schöttler et al. 2007), which would be evolutionarily deleterious. The only other known cases of loss of psaI from plastid DNA are in the parasitic species Cuscuta gronovii and Cuscuta obtusiflora which have reduced levels of photosynthesis but retain all other photosynthesis genes (Funk et al. 2007; McNeal et al. 2007), and in the nonphotosynthetic parasite Epifagus (Wolfe et al. 1992).

Although we have direct evidence for association between gene losses and a mutation hotspot only in the genus Lathyrus, it is intriguing that other species in the IRLC legume clade show evolutionary losses of other genes that neighbor ycf4 and psaI (Fig. 3). The loss of rps16 in the common ancestor of the IRLC clade can be explained in terms of gene substitution by the nuclear gene for mitochondrial RPS16, as already demonstrated for Medicago (Ueda et al. 2008), and so does not necessitate a gene transfer to the nucleus. Rps16 has been lost on multiple independent occasions during land plant evolution (Supplemental Table S1; Ohayama et al. 1986; Tsuzuki et al. 1992; Ueda et al. 2008), so it is possible that its multiple losses in legumes are simply the result of relatively easy and/or early substitution by the mitochondrial gene.

The other IRLC legume gene loss in the neighborhood of ycf4 and psaI is the loss of accD in Trifolium (Fig. 3). AccD codes for a subunit of acetyl-CoA carboxylase, which functions in lipid synthesis and is an essential chloroplast gene in tobacco (Kode et al. 2005). The loss in Trifolium is one of five separate known instances of loss of accD in angiosperm cpDNAs (Supplemental Table S1). In grass species—the only case that has been studied in detail—the protokaryotic multisubunit carboxylase in the plastid has been completely replaced by a nuclear-encoded single-chain carboxylase of eukaryotic ancestry (Konishi et al. 1996; Gornicki et al. 1997). We identified an evolutionary transfer of accD to the nucleus in Trifolium. Using high-throughput EST sequence data from Trifolium repens (white clover), we found a cDNA structure consisting of a fusion between a gene for plastid lipoamide dehydrogenase (LPD2) and accD (Supplemental Fig. S6A-D). We confirmed the presence of a fused mRNA by reverse transcriptase PCR and Sanger sequencing (Supplemental Fig. S6E).

In plastids, lipoamide dehydrogenase is a component of pyruvate dehydrogenase, a complex that makes acetyl-CoA (Lutziger and Oliver 2000; Drea et al. 2001). The T. repens nuclear transcript codes for a predicted protein of 805 amino acids, with residues 1–512 (including a transit peptide) derived from LPD2 and residues 513–805 derived from accD. By comparison to the known genomic structures of LPD genes in M. truncatula, we infer that in T. repens the accD sequence has replaced the final two exons (exons 14 and 15) of its LPD2 gene, with the point of fusion occurring at the third codon of exon 14. We did not find any evidence for alternative splicing of the LPD2–accD fusion to form two products, as occurs with the SOO–ppl32 fusion in mangrove trees (Cusack and Wolfe 2007). The fusion to accD probably rendered LPD2 unable to code for functional lipoamide dehydrogenase, because the fusion protein lacks some conserved residues normally provided by exons 14 and 15, but T. repens retains and expresses a paralogous gene LPD1 that also codes for plastid lipoamide dehydrogenase (Supplemental Fig. S6A). We found the transferred gene in T. repens, but we presume that the transfer is shared by other Trifolium species, including the two that have been demonstrated to have no accD in their cpDNAs (T. subterraneum and Trifolium pratense) (Doyle et al. 1995; Cai et al. 2008). We also found database ESTs for a nuclear accD in T. pratense (red clover), but they are too short to confirm that this species also has the LPD2–accD fusion. Phylogenetic analysis indicates that the T. repens and T. pratense nuclear sequences have a monophyletic origin and that the transfer of accD to the nucleus occurred within the IRLC clade (Supplemental Fig. S6F), consistent with the change in LPD2 gene structure that occurred after Trifolium diverged from Medicago (Supplemental Fig. S6A). The Trifolium nuclear accD gene is transcribed in both T. repens and T. pratense, is predicted to have a functional transit peptide in T. repens (TargetP cTP score 0.976) (Emmanuelsson et al. 2000), and shows evidence of selection to maintain its AccD-coding function (dω/ds = 0.26 between T. repens and T. pratense in the accD region of the transcript). Moreover, the Trifolium nuclear mRNAs code for a leucine residue at a site that undergoes an essential Ser → Leu mRNA edit in P. sativum plastids (Supplemental Fig. S6D; Sasaki et al. 2001; Inada et al. 2004).

Discussion

The genomic region around ycf4 in Lathyrus is a dramatic hotspot for point mutations. It is difficult to quantify the factor by which its mutation rate is increased relative to the rest of the genome, but comparisons of synonymous site divergence indicate an increase of at least 20-fold, both in comparisons between P. sativum and L. sativus (Fig. 4) and among Lathyrus species (Table 1). Between L. latifolius and L. cirrhousus, the increase may be even greater (Fig. 5; Table 1). Even a 20-fold mutation rate increase only goes partway toward explaining how the protein sequence divergence between L. palustris and L. cirrhousus (with a divergence time of <10 Myr) (Kenicer et al. 2005) exceeds that between other angiosperms and cyanobacteria (separated by >1000 Myr); a relaxation of selective constraints on the Ycf4 protein in legumes must be involved too. Although there have been previous reports that the variance of synonymous substitution rates among genes in many eukaryotic genomes is greater than expected by chance (e.g., Baer et al. 2007; Fox et al. 2008), there are few if any precedents for the phenomenon that we describe here—a sharply localized mutation rate acceleration of great magnitude in one specific region of a genome. The existence of the hotspot violates the common assumption that the point mutation rate is approximately constant in all regions of the same genome (Kimura 1983), which underpins the silent molecular clock hypothesis (Ochman and Wilson 1987). Our results bear some similarities to the “mutation showers” (transient localized hypermutation events) that have been found in some studies on the genomic distribution of spontaneous mutations (Drake 2007; Wang et al. 2007; Nishant et al. 2009). As well as being a mutation hotspot, ycf4 and its neighbors also appear to be a hotspot for the formation and turnover of minisatellite sequences in Lathyrus.

The previous study most relevant to our findings is that of Erixon and Oxelman (2008), who reported somewhat similar results for the chloroplast clpP gene in Silene and Oenothera species. For some interspecies comparisons in their study, both dS and dD were elevated in clpP compared with other chloroplast genes, although the dS elevations were at most fivefold for clpP, compared with at least 20-fold for ycf4 in Lathyrus. Also, insertions of repetitive amino acid sequence regions occurred in some of the fast-evolving taxa. Locus-specific rate accelerations affecting both dS and dD were reported in cpDNA of Geraniaceae, but in this case, the accelerations occurred in numerous genes (Guisinger et al. 2008). In all IR-containing cpDNAs, the synonymous rate is higher in single-copy genes than in IR-located genes, probably due to a copy-number effect during DNA repair (Wolfe et al. 1987; Birky and Walsh 1992; Perry and Wolfe 2002). Dramatic accelerations of
synonymous rates have been found in the mitochondrial genomes of some plants, such as *Plantago*, *Pelargonium*, and certain *Silene* species (Cho et al. 2004; Parkinson et al. 2005; Mower et al. 2007; Sloan et al. 2009). Most of these mitochondrial accelerations appear to affect all genes in the genome similarly, but among-gene rate heterogeneity was found within the mtDNAs of a few species (Mower et al. 2007), including a 40-fold difference in synonymous rates between *atp9* and three other mitochondrial genes in *Silene* (Sloan et al. 2009). Because plant mitochondrial genomes are relatively large and do not show much gene order conservation, most studies have only examined individual genes so the sizes of the genomic regions affected by rate acceleration are not known.

Apart from these organellar examples, there are very few precedents for a mutation rate change that is so pronounced over such a short physical distance. One early study (Martin and Meyerowitz 1986) reported a 2-kb region of noncoding DNA near the glue gene cluster of three *Drosophila* species, which contained an abrupt boundary between a conserved region and a nonconserved region with a 10-fold elevated substitution rate, but this report has not been followed up with more extensive analyses based on complete genome sequence data. An abrupt boundary of evolutionary rates also occurs on the mammalian X chromosome at the junction between the pseudoautosomal region and the X-specific region. The pseudoautosomal part of the gene *Fxy*, which spans this junction in laboratory mice, has a synonymous rate about 60 times faster than the X-specific part of the gene, which spans this junction in laboratory mice, has a synonymous rate about 60 times faster than the X-specific part of the gene, probably because the high recombination rate in the pseudoautosomal part leads to high levels of biased gene conversion (Perry and Ashworth 1999; Duret and Galtier 2009).

Is the chloroplast hypermutation phenomenon unique to *Lathyrus*? At present, *Lathyrus* is the only legume genus for which we have extensive sequence data from more than one species, so we are unable to say whether the same hotspot is present in legumes outside this genus. Therefore the only gene losses we can potentially attribute directly to hypermutation are those of *ycf4* in *L. odoratus* and of *psaI* in the ancestor of four *Lathyrus* species. *Ycf4* is also evolving fast in *Desmodium* and has been lost in three species of that genus. Losses of *ycf4* in *P. sativum*, of *accD* in *Trifolium*, and the older loss of *rps16* in the ancestor of the IRLC clade are suggestive but we have no direct evidence that these loci were fast-evolving prior to the gene losses. It is possible that a hotspot has existed throughout legume evolution and was the cause of the *ycf4* acceleration seen in the common ancestor of *Millettioides*, *Robinioides*, and the IRLC (Fig. 1) but that the exact location of the hotspot (and its associated tandem repeat sequences) has varied somewhat among lineages, affecting *ycf4* in some taxa, but *accD* or *psaI* in others. We do not know the molecular basis for the increases in either the point mutation rate or the length mutation rate, but we speculate that they might be connected. We suggest that a correlation between the two rates could develop if, for some reason, the genomic region around *ycf4* was subject to repeated DNA breakage and repair (cf. Guisinger et al. 2008; Yang et al. 2008). In this regard, it is interesting to note that only a few angiosperm species have cpDNAs that are highly rearranged relative to the canonical gene order, but among these, there are several independent lineages that are both highly rearranged and contain rapidly-evolving protein genes (Jansen et al. 2007). These lineages include *Jasminum* (acceleration of *accD*; Lee et al. 2007), *Silene* (acceleration of *clpP*; Erixon and Oxelman 2008), and now *Lathyrus* (acceleration of *ycf4*). The phylogenetic diversity of these lineages suggests that hypermutable regions may exist in other angiosperm cpDNAs, and our findings may go some way toward explaining the apparent bursts of organelle-to-nucleus gene transfer seen in some angiosperms.

It is likely that many factors dictate whether a gene can be lost from an organelle genome. One property that is common to the gene transfer and gene substitution processes is that they both involve a phase during which the organelle gene and the nuclear gene coexist in the same species (Timmis et al. 2004). Analogous to a gene duplication, this two-gene phase can be resolved either by losing the organelle copy (resulting in a successful transfer of function) or by losing the nuclear copy (restoring the status quo). Intermediates in this process, and sister lineages where the two-gene phase was resolved in opposite ways, have been identified (Adams et al. 1999). Brandvain and Wade (2009) have shown theoretically that the ratio between the point mutation rates in the organelle and nuclear copies has a profound influence on the direction in which the two-gene phase is resolved. If the organelle mutation rate is lower than the nuclear mutation rate, as is true for most plant mitochondrial and chloroplast genes, then gene transfer will not occur unless there is a benefit to relocating the gene. By contrast, if the organelle rate exceeds the nuclear rate, then gene transfer is predicted to occur even in the absence of any benefit (Brandvain and Wade 2009). Therefore, in a genome such as *Lathyrus* cpDNA, in which the mutation rate exceeds the nuclear rate only in one hypermutable region, we should expect to see more transfers, substitutions, or losses of genes from the hypermutable region than from the rest of the genome. This argument provides a plausible explanation for the losses of *ycf4* and *psaI* seen in some *Lathyrus* species and, perhaps more generally, for the cluster of losses from the *rps16-accD-psaI-ycf4* region seen in other legume cpDNAs.

### Methods

#### Plant material

Seeds of *Lathyrus sativus* (cv. Ciccheria Marchigiana) were purchased from B&K World Seeds. Seeds of *L. cirrhosus* (accession no. LAT17) were obtained from the Leibniz Institute of Plant Genetics and Crop Plant Research. Other *Lathyrus* species were purchased from Thompson & Morgan. Additional sequencing of *P. sativum* cpDNA was done using cv. Feltham First.

#### Nucleotide sequencing

The *P. sativum* (pea) chloroplast genome sequence was completed by S.A. and J.C.G. using the chain termination method (Sanger et al. 1977) with fluorescent dideoxynucleotides on PCR products amplified from cloned *PstI* fragments (Palmer and Thompson 1981), from cpDNA extracted from isolated chloroplasts, or from total DNA extracted from shoots of 8-d-old seedlings. Chloroplasts were isolated by the high-salt method (Boorkmans et al. 1984), and DNA was extracted by the CTAB (hexadecyltrimethylammonium bromide) method (Milligan 1989). Previously published regions were not resequenced except at the borders or where there were discrepancies between publications. Newly sequenced regions were completed on both strands, and all the *PstI* sites used for cloning were confirmed by sequencing spanning PCR fragments. At the *ycf4* locus, there is a 2-bp deletion in the sequence reported by Nagano et al. (1991a), relative to the sequence reported by Smith et al. (1991), both of which were obtained from the same cloned 17.3-kb *PstI* fragment from *P. sativum* cv. Alaska. We confirmed that this 2-bp deletion exists, in both *cv. Alaska* and *cv. Feltham First*. This correction means that the reported ORF157 (Smith et al. 1991) does not exist.
The \textit{L. sativus} (grasspea) chloroplast genome sequence was determined by A.M.A., T.A.K., and K.H.W. Approximately 150 seeds were grown on soil in the greenhouse. Seedling shoots were harvested at 7 d post-germination, and cpDNA was prepared according to the method described by Milligan (1989) except that chloroplast lysis and cpDNA recovery procedures were modified. Chloroplasts were lysed by adding a 1/5 volume of 10% CTAB (Sigma-Aldrich) and heating for 20 min at 70°C. This was followed by a chloroform extraction, treatment with RNaseA (10 \( \mu \)g/mL), and isopropanol precipitation of cpDNA. A plasmid library of nebulated fragments was constructed from 5 \( \mu \)g of cpDNA by GATC-Biotech. The genome sequence was assembled from 1536 Sanger shotgun sequence reads with primer-walking to close gaps.

We tried unsuccessfully to amplify \textit{ycf4} and \textit{psaL} by PCR from \textit{L. odoratus} genomic DNA using 16 and 9 primer combinations, respectively, and a range of amplification conditions. These primers were designed based on amino acid residues conserved among known Fabaceae \textit{Ycf4} and \textit{Psal} proteins, but primer design for these genes is difficult due to the fast rate of \textit{ycf4} evolution and the short length of \textit{psaL}, as well as the high A+T content of the region. To obtain EST data from \textit{L. odoratus}, we isolated poly(A) mRNA from leaves of 3-d-old seedlings. A normalized cDNA library was constructed by GATC-Biotech, and the 3’ ends of 8702 cDNAs were sequenced by Agencourt Biosciences. ESTs were assembled into contigs, and putative orthologs between these contigs and \textit{P. sativum} sequence data from GenBank were identified according the method of Sémon and Wolfe (2008).

The other new sequence data indicated in Figures 3 and 5 were generated by PCR amplification and sequencing (by primer walking) of at least three independent cloned products for each region. The cDNA region that normally contains \textit{ycf4} was PCR amplified from \textit{L. latifolius}, \textit{L. cirsus}, \textit{L. odoratus}, and \textit{L. palustris} using primers designed from the \textit{P. sativum accD} (5’-AACGAGCCACAGG TCAASTAAAATGG-3’) and \textit{cemA} (5’-GACGGGATACAGGATTATAAATCAG-3’) genes. The \textit{atpB-rbcL} region from \textit{L. latifolius}, \textit{L. cirsus}, and \textit{L. palustris} was amplified with primers 5’-TGRAAAACTCATACGATCAGGGG-3’ and 5’-TATGATCTCACCAGGA CATACG-3’. \textit{T. repens} mRNA sequences coding for \textit{LpD1} and the \textit{LPD2-accD} fusion gene were identified among 700,000 ESTs obtained by high-throughput pyrosequencing of flower, leaf, and stolon mRNA from the inbred line S (7S.4.6.3.3.4.4.10) (DM and SB, unpubl.) and assembled manually. The structure of the \textit{LPD2-accD} junction was confirmed by reverse transcriptase-PCR from \textit{LpD1} leaf mRNA (commercial variety Nusalir) and Sanger sequencing.

Computational methods

Sequence divergence for most analyses was calculated using yno0 from the PAML package (Yang 2007) for coding regions and Kimura’s two-parameter method (Kimura 1983) for noncoding regions. Gene sequences were aligned by reverse-translation of ClustalW alignments of the corresponding protein sequences. Noncoding sequences were aligned using ClustalW with manual adjustment for regions around \textit{ycf4}. For the analysis in Figure 1 and Supplemental Figure S1, we first constructed a maximum likelihood phylogeny (in PAUP) from \textit{matK} sequences using the HKY substitution model with a four category gamma rate distribution. The transition/transversion ratio and shape parameter were estimated iteratively until the topology converged. This analysis included legume \textit{ycf4} sequences from Stefanovic et al. (2009): GenBank [http://www.ncbi.nlm.nih.gov/Genbank/] accession nos. EU717431–EU717464). The \( d_0 \) and \( d_\infty \) branch lengths for the \textit{matK}, \textit{ycf4}, and \textit{rbcL} trees were estimated based on the PAML/codeML free-ratio model, using the fixed topology obtained from the above

matK ML analysis. Dot-matrix plots were made using DNAMAN (http://www.lynnnon.com) (Huang and Zhang 2004).

Acknowledgments

We thank Gavin Conant for help with Figure S3, Shusei Sato for \textit{Tritiumol} cDNA clones, Greg Kenicer for prepublication access to sequence data, and Ralph Bock for discussion. S.A. and J.C.G. thank Chris Maddren (Department of Genetics, University of Cambridge) for DNA sequencing. This study was supported by Science Foundation Ireland (K.H.W., T.A.K.), European Commission FP5 Plastid Factory (J.C.G., T.A.K.), US National Institutes of Health (J.D.P.), and the METAcyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. (J.D.F., D.W.R.).

References


Received June 20, 2010; accepted in revised form September 20, 2010.