Plant Molecular Biology **29**: 721–733, 1995.

Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*

Stephanie C. Ems[1], Clifford W. Morden[2], Colleen K. Dixon[3], Kenneth H. Wolfe[4], Claude W. dePamphilis[5] and Jeffrey D. Palmer*

*Department of Biology, Indiana University, Bloomington, IN 47405, USA (*author for correspondence); present addresses: [1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [2]Department of Botany/H.E.B.P., University of Hawaii at Manoa, Honolulu, HI 96822, USA; [3]Eli Lilly and Company, Inc., Lilly Corporate Center, Indianapolis, IN 46285, USA; [4]Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland; [5]Department of Biology, Vanderbilt University, Nashville, TN 37235, USA*

Received 17 May 1995; accepted in revised form 11 August 1995

Key words: *Epifagus*, maturase, nonphotosynthetic plant, RNA editing, splicing, transcription

Abstract

Expression of the vestigial plastid genome of the nonphotosynthetic, parasitic flowering plant *Epifagus virginiana* was examined by northern analysis and by characterization of cDNAs. Probes for each of 12 plastid genes tested hybridized to all lanes of northern blots containing total RNA prepared from stems and fruits of *Epifagus* and from leaves of tobacco. Certain transcript patterns in *Epifagus* plastids are highly complex and similar to those of tobacco operons. In contrast, genes such as *rps2*, which have become orphaned in *Epifagus* as a result of evolutionary loss of formerly cotranscribed genes, show simpler transcript patterns in *Epifagus* than in tobacco. Sizing and sequencing of cDNAs generated by reverse transcriptase-PCR for three genes, *rps12*, *rpl2*, and *clpP*, show that their transcripts are properly *cis*- and/or *trans*-spliced at the same five group II intron insertion sites used in photosynthetic plants. A single, conventional C→U edit in *rps12* was found among the total of 1401 nucleotides of cDNA sequence that was determined for the three genes. An octanucleotide sequence identical to a putative guide RNA of plant organelles and perfectly complementary to the *rps12* edit site itself was identified just 200 bp upstream of the edit site. These data, together with previous results from the complete sequencing of the *Epifagus* plastid genome, provide compelling evidence that this degenerate genome is nonetheless expressed and functional. Analysis of the putative maturase MatK, encoded by the group II intron of *trnK* in photosynthetic land plants but by a freestanding gene in *Epifagus*, leads us to hypothesize that it acts 'in *trans*' to assist the splicing of group II introns other than the one in which it is normally encoded.

Introduction

Complete sequencing of the 70 kb plastid genome of the nonphotosynthetic, parasitic flowering plant *Epifagus virginiana* (beechdrops; Orobanchaceae) has shown that it contains only 42 intact genes [27, 41–44], compared to 113 plastid genes in tobacco, its closest, fully sequenced photosynthetic relative [32, 34, 40]. Although at least 38 of the 42 *Epifagus* plastid genes specify components of the gene expression apparatus, all four normally plastid-encoded RNA polymerase genes are missing or defunct, implying that plastid transcription in *Epifagus* is entirely dependent on a nuclear-encoded polymerase [27, 41]. All but one of these 38 gene expression genes encode components of the plastid translational apparatus. However, many plastid tRNA and ribosomal protein genes have also been lost from the *Epifagus* genome, again implying surprising reliance on nuclear gene products [27, 42]. Functions are known for only two of the four nongenetic genes in *Epifagus*, *clpP* (encoding a protease subunit) and *accD* (encoding a subunit of acetyl-CoA carboxylase).

Surprisingly, the one nontranslational gene thought to be involved in gene expression, which we previously named '*matK*', is a freestanding gene in *Epifagus* [41]. In contrast, in all examined photosynthetic land plants *matK* is contained within the group II intron of the tRNA-lysine (UUU) gene (*trnK*) and is thought to encode a 'maturase', or splicing factor, that assists in the splicing of the intron in which it residues [2, 3, 11, 21, 28, 31, 33, 37, 38]. In particular, Mohr *et al.* [26] have shown that all *matK* genes contain a ca. 100 amino acid domain (termed domain 'X') that is present in virtually all group II intron maturases. We [41] and others [8, 22, 26] have suggested that, uniquely among group II maturases, the *matK* product (MatK) assists splicing of multiple, distantly related group II introns.

Molecular evolutionary evidence suggests that the *Epifagus* plastid genome is expressed and functional [41]. First, deletions are highly nonrandom; for example, 95% of photosynthetic sequences have been deleted, whereas only 20% of ribosomal protein sequences and none of the ribosomal RNA sequences are missing. Second, whereas 71 genes are entirely or largely deleted, many of the 42 remaining intact genes are so large (e.g., 2216 and 1738 codons) that their maintenance must result from selection for their continued function.

While the evolutionary arguments in favor of plastid genome function in *Epifagus* are strong, evidence in the form of detectable gene products is minimal. Thus far, this consists of a single northern blot of *Epifagus* total RNA, to which a probe for the 16S rRNA gene from tobacco chloroplasts hybridized to transcripts of 1.5 kb (interpreted as plastid 16S rRNA) and 1.8 kb (interpreted as cross-hybridization to mitochondrial 18S rRNA) [4]. A major goal of the present study is to remedy this situation by determining whether or not transcripts can be detected from a significant number of *Epifagus* plastid genes. A second goal is to examine the pattern of plastid transcripts in *Epifagus*, and to make direct comparisons to tobacco, in order to investigate the effects of the many gene and promoter deletions [41] on the expression of plastid operons of varying complexity. A final goal is to determine whether group II introns are properly spliced in *Epifagus* plastids and to relate this to the putative maturase activity of its freestanding *matK* gene.

Materials and methods

Above-ground stems and fruits of mature plants of *Epifagus virginiana* were collected from locations in Michigan, Indiana and Tennessee. Leaves of tobacco (*Nicotiana tabacum*) were harvested from greenhouse-grown plants. Field-collected plants were frozen in liquid nitrogen upon collection and stored at −70 °C prior to nucleic acid isolation. Total DNA was isolated using the modified CTAB method of Doyle and Doyle [5]. Total RNA was isolated either as described previously [4] or according to the following procedure. Frozen tissue (5–7 g) is ground in liquid nitrogen using a chilled mortar and pestle, and then placed in a −70 °C freezer for 5 min to drive off the liquid nitrogen. The frozen powder is added to a hot (60 °C) phenol emulsion [10 ml phenol plus 12 ml extraction buffer (100 mM Tris pH 8.0, 20 mM EDTA, 0.4% SDS, 0.5 M NaCl, 0.1% 2-mercaptoethanol)] and stirred at 60 °C for 5 min. Ten ml of chloroform/isobutanol (24:1) is added, and stirring is continued for another 5 min. The solution is centrifuged in an RNase-free tube at 10000 r.p.m. for 10 min at room temperature, and the aqueous phase is transferred to a new centrifuge tube containing 10 ml of chloroform/isobutanol (24:1). The solution is mixed well and then centrifuged as above. The aqueous phase is adjusted to 2 M NH_4 OAc and precipitated with 1 volume of isopropanol at room temperature for 15 min. The precipitate is collected by centrifugation, washed twice with 70% ethanol, dried, and resuspended in 0.8 ml of DEPC-treated dH_2O.

RNA samples were denatured with glyoxyl and dimethyl sulfoxide and electrophoresed in 1.4% agarose gels (using a Jordan Scientific recirculating gel apparatus) according to the procedure of Sambrook *et al.* [30]. RNA was transferred from gels to MagnaGraph nylon filters using a Stratagene Posiblotter. Filters were UV-crosslinked for 23 s using a DNA transfer lamp from Photodyne and baked at 80 °C for 1 h. Hybridization probes were made by random-prime labeling [30] with ^{32}P-dATP. Filters were prehybridized for 4–16 h and hybridized overnight in 1 M NaCl, 1% SDS at 60 °C. Filters were washed in 2 × SSC, 5% SDS two times for 5 min at room temperature and then three times for 30 min at 60 °C. Filters were stripped of probe by two 5 min boiling washes with 0.1 × SSC, 0.1% SDS, 0.2 M Tris-HCl.

PCR reactions were carried out using Taq polymerase from Promega, dNTPs from Pharmacia, and the following pairs of primers: for *rpl2* the 5′ primer sequence was ATGGCGATACATTTATACAAAACTTCTAC [primer coordinates, according to the *Epifagus* coordinate system ([41]; also see GenBank accession number M81884) are 21406–21378] and the 3′ primer was AGCCAACGCTTAGATCCGGCTCTACC (coordinates 20111–20136); for *rps12* the 5′ primer was AAAAAGACAGCCAA(AT)C(CA) (12833–12817) and the 3′ primer was ATAAGGGCTAAAATCAC (31297–31313); for 5′ *clpP* the 5′ primer was ATGCCTATTGGTGTNCCNAA (14802–14783) and the 3′ primer was AAAGATCCCATTG(AT)NGCNGC (13844–13863); for 3′ *clpP* the 5′ primer was CTATATCTCAGTAT(AT)GA(AG)GA (14004–13985) and the 3′ primer was CATAAAAACATC(AG)CT(AG)TCCAT (13055–13075). Reaction volumes and conditions varied depending on use of thermocyclers from either Perkin Elmer Cetus or Idaho Technology. Some reverse-transcriptase-PCR (RT-PCR) reactions were performed using the Perkin Elmer Cetus GeneAmp RNA PCR kit. In general, the reverse transcription step consisted of incubations at 42 °C for 15 min, 99 °C for 5 min, and 5 °C for 25 min, while PCR consisted of 30 cycles of 95 °C for 1 min, 50 °C for 1 min, and 72 °C for 2 min. RNase (RNace-It) and DNase I (RNase-free) were from Stratagene. PCR products were analyzed on 2% NuSieve (FMC)/1% agarose gels and sequenced directly by standard methodology.

Results

Transcript patterns in Epifagus *plastids*

Northern blot analyses were carried out for 12 *Epifagus* plastid genes to determine whether these genes are expressed, to evaluate the complexity of their transcript patterns, and to make comparisons to the situation in tobacco, its closest relative among well-studied photosynthetic plants. All hybridizations were carried out against filters containing total RNA from stems and fruits of *Epifagus* and from green leaves of tobacco (e.g. Fig. 1). Preliminary experiments [4] showed that, for most genes, these *Epifagus* organs contain only about one-fiftieth as much plastid RNA as a proportion of total cellular RNA as do tobacco leaves. Hence, to obtain similar signals across lanes, we loaded 5 μg of each *Epifagus* RNA but only 0.1 μg of tobacco RNA.

Probes for all 12 *Epifagus* genes tested hybridized to one or more discrete transcripts, suggesting that they are all transcribed. These genes include all four rRNA genes (23S, 16S, 5S, 4.5S), four ribosomal protein genes (*rpl2*, *rps2*, *rps12*, *rps14*), two large ORFs (*ORF2216*, *ORF1738*), *accD*, and *clpP*. The 12 genes can be roughly grouped into three classes according to the complexity of their transcript patterns in *Epifagus* and the similarity of these patterns between *Epifagus*

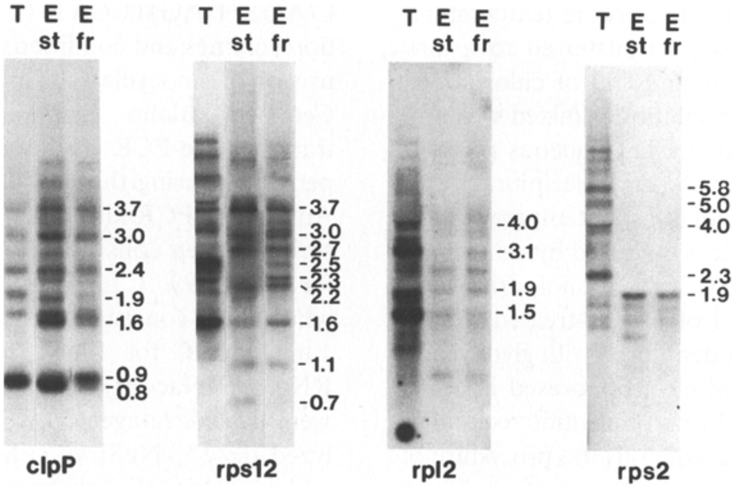

Fig. 1. Northern blot analysis of plastid transcript patterns in *Epifagus* and tobacco. A 5 μg portion of total RNA from stems (st) and fruits (fr) of *Epifagus* (E) and 0.1 μg of total leaf RNA from tobacco (T) were electrophoresed in agarose gels, transferred to filters, and hybridized with PCR probes (see Fig. 2 for probe locations) specific for each of the four genes indicated below the hybridization panels. Transcript sizes (in kb) are based on the 0.24–9.5 kb RNA ladder of GIBCO BRL. The *clpP* and *rps12* probes were hybridized successively to one filter, while the *rpl2* and *rps2* probes were hybridized to a different filter, made from a gel that had been run less far.

and tobacco. The two large ORFs detected one or two large (≥6 kb) transcripts of similar size in both species (data not shown). In contrast, two ribosomal protein genes (*rps2*, Fig. 1; *rps14*, data not shown) gave far simpler transcript patterns in *Epifagus* than in tobacco. The remaining eight genes all displayed generally similar and highly complex transcript patterns in both species. Examples of these latter two classes of patterns are illustrated in Fig. 1 and described below.

Transcript patterns for clpP *and* rps12

Throughout photosynthetic land plants, including tobacco, exon 1 of *rps12* is located between *clpP* and *rps20* and is thought or known to be cotranscribed with them (Fig. 2; [14, 15, 39]). Exons 2 and 3 of *rps12* are located in the large inverted repeat, far away from exon 1, and are instead cotranscribed with *rps7* (Fig. 2). The mature mRNAs from each of these four genes are created by a complex series of events, including independent transcription of two small operons, transcript clipping between coding regions, three *cis*-splicing events (of intron 2 of *rps12* and of both introns of *clpP*), and the *trans*-splicing of intron 1 of *rps12* [10, 15–17, 45]. All four of these genes are retained as intact, conserved open reading frames in *Epifagus* ptDNA, and their organization is also conserved (Fig. 2; [41, 42]). Probes for two of the four genes (*clpP* and *rps12*) gave highly complex patterns of transcripts for both *Epifagus* and tobacco (Fig. 1). The *clpP* patterns are more similar between the two species, both qualitatively and quantitatively, than are the *rps12* patterns. The six most abundant *clpP* transcripts are of the same or similar sizes in the two species. The smallest of these, the putative monocistronic, fully spliced *clpP* mRNA of 800–900 nucleotides (ClpP is 197 amino acids), is, by at least a factor of two, the most abundant transcript in both species. Relatively few of the many *rps12* transcripts are of the same size in the two species, and even for these there are significant quantitative differences. Thus, in tobacco there are two predominant *rps12* transcripts, of 1.6 kb (the putative 'mature dicistronic *rps12*/*rps7* mRNA'; [10]) and of 2.5 kb, whereas in *Epifagus* these are just two of a multitude of bands of roughly equal abun-

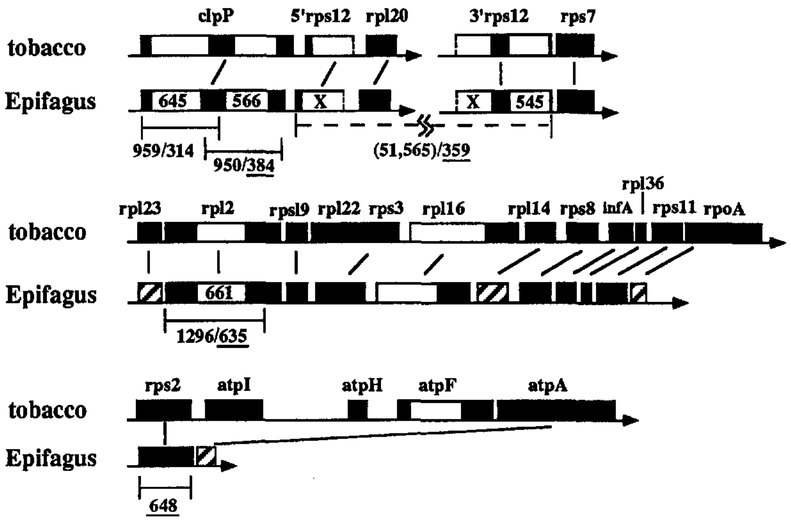

Fig. 2. Comparative organization in *Epifagus* and tobacco of three sets of genes whose expression is examined in Figs. 1 and 3. Filled boxes represent exons, open boxes show introns, and hatched boxes indicate pseudogenes. Angled lines between the *Epifagus* and tobacco maps connect homologs. Arrows indicate direction of transcription of genes thought to be contranscribed. The sizes of the four *cis*-spliced *Epifagus* introns whose splicing is examined in Fig. 3 are indicated in bp within the intron boxes. The two 'halves' of the *trans*-spliced intron in *Epifagus rps12* are indicated with an 'X'. The horizontal lines below the *Epifagus* maps indicate regions that were PCR-amplified to analyze plastid splicing (Fig. 3) and/or generate probes for northern hybridizations (Fig. 1). The vertical slashes at the ends of these lines indicate the positions of the PCR primers (see Materials and methods for primer coordinates and sequences). The numbers below these lines indicate the expected sizes of PCR products using either genomic DNA as template (before the slashes) or cDNA as template (after the slashes). For *rps2*, only genomic PCR was done. The four underlined PCR product sizes indicate those products used as probes in Fig. 1. The dashed line and parentheses below *rps12* denote the fact that, since intron 1 of *rps12* is *trans*-spliced, its flanking exons, as located in the same orientation, are 51,565 bp apart in the genome. *Epifagus* mapping data are from [41]; tobacco mapping and expression data are from [10, 17, 29, 32, 34, 36, and 45].

dance. As expected for contranscribed genes (Fig. 2), *clpP* and *rps12* hybridized to some of the same-sized larger bands, i.e., transcripts of 3.7 and 3.0 kb (Fig. 1).

These comparisons between species are complicated by the more subtle differences evident in transcript patterns between *Epifagus* RNAs made from stems and fruits (Fig. 1). The major *clpP* transcript in stems and fruits is 0.8 and 0.9 kb, respectively, while each tissue appears to contain a small amount of the other RNA species. These differences most likely reflect organ-specific differences in the extent of clipping at the 5′ and/or 3′ termini of the spliced, monocistronic *clpP* mRNA. In the case of *rps12*, certain transcripts appear to be uniquely present in either stems (0.7 and 2.0 kb) or fruits (2.5, 3.1, 3.3 kb), while the other six or seven transcripts appear to be present in both organs.

Transcript patterns for rp12

In most photosynthetic land plants, including tobacco (Fig. 2), *rpl2* is part of a large (ca. 12 genes, ca. 8 kb) operon of mostly ribosomal protein genes [7, 14, 34, 36]. This gene cluster is reduced in size by about 1.5 kb in *Epifagus* as the result of the deletion of part or all of four of these genes (Fig. 2). As expected, both *Epifagus* and tobacco show complex transcript patterns for *rpl2*, which also contains a single intron of about 660 bp in both species. As with *rps12*, certain *rpl2* transcripts (of 1.5, 1.9, and 3.1 kb) predominate in tobacco, whereas all *rpl2* transcripts are of roughly equal abundance in *Epifagus*. Compared to transcript levels for the same gene in tobacco, *rpl2* is notably less highly expressed in *Epifagus* than are any of the other genes examined (Fig. 1 and data not shown). In spite of these quantitative differences, several *rpl2* transcripts (1.5–3.1 kb in size) appear to be the same size in *Epifagus* and tobacco.

Transcript patterns for rps2

In striking contrast to *clpP*, *rps12* and *rpl2*, and to the four ribosomal RNA genes, which also show similarly complex transcript patterns in both species (data not shown), the *rps2* transcript pattern is much simpler in *Epifagus* than in tobacco (Fig. 1). In *Epifagus*, there is a major *rps2* transcript of 1.9 kb, a few smaller transcripts, and no larger ones. In tobacco, however, there are many *rps2* transcripts, most of which are larger than 2 kb in size. This difference in transcript complexity is an expected consequence of gene loss in *Epifagus* ptDNA. In photosynthetic plants, *rps2*, including tobacco (Fig. 2), is upstream of, and cotranscribed with, a cluster of four ATPase genes (*atpI*, *atpH*, *atpF*, and *atpA*; [12, 29]), one of which contains an intron. The first three of these genes are completely absent from *Epifagus* ptDNA, while only a short remnant of *atpA* is present (Fig. 2).

Splicing and editing of Epifagus *plastid transcripts*

To investigate potential splicing and editing of *Epifagus* plastid transcripts, cDNAs generated from three genes that contain five of the six surviving *Epifagus* plastid introns were analyzed by sizing and sequencing. cDNAs were made by reverse transcription using the 3′ primers given in Materials and methods, followed by PCR amplification using the same 3′ primers in combination with the indicated 5′ primers (primer locations are shown in Fig. 2). Three cDNAs were designed to contain a single intron insertion site (for the single *rpl2* intron, and for each of the two *clpP* introns), while the *rps12* cDNA was designed to encompass the positions of both its *cis*- and *trans*-spliced introns.

Agarose gel analysis shows that the predominant products from RT-PCR of RNA templates (Fig. 3, lanes 3, 4) are smaller than those from DNA templates (Fig. 3, lanes 1, 2) by amounts corresponding precisely to the sizes of the relevant introns (cf. Figs. 2 and 3). For example, the major cDNA product for *rpl2* is about 640 bp (Fig. 3, lanes 3, 4), essentially as expected for a spliced transcript (expected size 635 bp; Fig. 2),

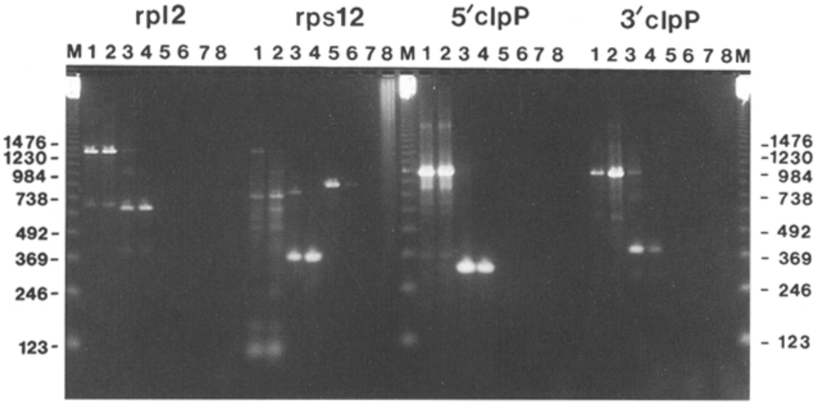

Fig. 3. PCR analysis of three intron-containing plastid genes and their transcripts. Primer pairs (see Fig. 2 and Materials and methods for primer locations and sequences) flanking either the single intron in *rpl2*, both the *cis*- and *trans*-spliced introns in *rps12*, or the first (5′) or second (3′) intron in *clpP* were used in PCR amplification of *Epifagus* total DNA (lanes 1, 2 and 6), total RNA (lanes 3, 4, 7 and 8), or no template (lane 5). *Taq* polymerase was included in all reactions, whereas a prior reverse transcription step was done for lanes 3, 4 and 8. Lanes 2, 6 and 8 contained RNAase, and lanes 4, 6 and 8 contained DNAase. Sizes (in bp) are shown for some of the bands from the 123 bp size ladder (lanes marked 'M') from GIBCO BRL.

and smaller than the major genomic product of about 1300 bp (Fig. 3, lanes 1, 2; expected size = 1296 bp) by 661 bp, the predicted length of its intron. Similar results hold for the 5′ and 3′ *clpP* amplifications. For *rps12*, the major cDNA product, of about 360 bp (Fig. 3, lanes 3, 4), again corresponds well to the size expected for the spliced gene product (359 bp). The genomic amplifications (Fig. 3, lanes 1, 2), on the other hand, yielded a large array of presumably nonspecific products owing to the *trans*-spliced organization of its first intron.

In general, the control lanes show that there is no template activity in the reverse transcriptase and PCR reagents (Fig. 3, lanes 5), that production of the DNA-template products of lanes 1 and 2 is DNase-sensitive (lanes 6), and that production of the RNA-template products of lanes 3 and 4 is either reverse transcriptase-dependent (lanes 7) or RNase-sensitive (lanes 8). In addition, lanes 2 show that production of the DNA-template products is RNase-resistant, and lanes 4 that production of the RNA-template products is DNase-resistant. The appearance of an 840 bp band in the negative control lanes 5 and 6 for *rps12* is, however, unexpected and inexplicable. Overall, these experiments lead us to conclude that significant levels of spliced RNA products are present for all three genes.

The four cDNA PCR products shown in Fig. 3 were completely sequenced directly, without cloning (data not shown). Comparison to the genomic sequences [41] establishes that the five group II introns of these three genes are located and spliced out at exactly the same positions as in the homologous genes of tobacco and other examined land plants [25, 34]. Genomic and cDNA exon sequences were identical for the regions compared in *rpl2* and *clpP*. However, a single difference, indicative of RNA editing, was found for *rps12*. Coding region position 221 is C in the *Epifagus* plastid genome, but T in the corresponding cDNA. This difference was observed, without any evidence of partial editing (i.e., all of the cDNA PCR products gave homogeneous sequencing ladders), in all five preparations of *Epifagus* RNA and DNA analysed from the same plants collected from five locations in Indiana, Michigan and Tennessee. We thus conclude that most or all *Epifagus rps12* transcripts, or at least the spliced ones, have undergone a C→U editing event. This edit changes codon 74 from TCA (serine) in the gene to UUA (leucine) in the mRNA. Overall, then, only a single edit was found among the total of 1401 bp of cDNA sequence determined for these three genes (this total is the sum of the nine cDNA exon lengths reported in Table 1, minus the overlap in *clpP* exon 2 between the 5′ and 3′ *clpP* cDNAs, and minus the PCR primers).

Discussion

General aspects of Epifagus *plastid genome expression*

The results presented in this paper firmly support our previous conclusion, based on molecular evolutionary evidence (see Introduction and [41]), that the highly reduced plastid genome of the nonphotosynthetic parasite *Epifagus virginia* is nonetheless a functional, expressed genome. All 12 *Epifagus* plastid genes examined are transcribed by the criterion of detection of stable RNAs in northern blot experiments using gene-specific probes. Moreover, eight of the genes display complex transcript patterns reminiscient of those found for the same genes in tobacco and other photosynthetic plants. This indicates that similar processes of transcription initiation and termination from multiple sites, transcript trimming and clipping, and *cis*- and *trans*-splicing occur in plastids of photosynthetic and permanently nonphotosynthetic plants. Direct evidence that splicing occurs in *Epifagus* plastids was obtained by sizing and sequencing cDNAs made by reverse transcriptase-based PCR. These experiments establish that all five examined introns (out of six total in *Epifagus* ptDNA) are spliced effectively *in vivo*. Finally, comparison of these cDNA sequences with published genomic sequences [41] reveals that RNA editing, a low-frequency but characteristic process of plastids in photosynthetic plants [18], occurs at similarly low levels in *Epifagus* plastids.

Overall, then, we conclude that the *Epifagus* genome is indeed transcribed, and that its transcripts are subject to the entire gamut of post-transcriptional processing events that are well characterized in photosynthetic plastids. The existence of such plastid-characteristic transcript patterns and processing events (e.g., group II intron splicing and RNA editing), together with the identity of the cDNAs to published plastid gene sequences, rules out the formal possibility that the highly reduced plastid genome in *Epifagus* is actually defunct and that the transcripts detected are instead the products of chloroplast genes transferred to the nucleus. Although we have not formally demonstrated expression of the *Epifagus* plastid genome at the level of translation, we feel this is a moot point. In our view, the molecular evolutionary evidence – the specific retention of a distinct subset of intact plastid genes (some quite large) against a background of deletion of most genes [41] – taken together with the evidence presented here for transcription and proper post-transcriptional processing of these retained genes, makes it a foregone conclusion that these transcripts are translated into functional plastid polypeptides.

Although seven of the nine genes examined that show complex transcript patterns in tobacco also show complex patterns in *Epifagus*, in most cases significant qualitative and quantitative differences were observed between the two species using the same gene probe. In other words, the case of *clpP*, in which both species show the same-sized predominant transcript, as well as several minor, higher-molecular-weight transcripts of the same or similar sizes, is unusual (Fig. 1). For *rps12* and *rpl2* (Fig. 1), there are fewer comigrating transcripts in tobacco and *Epifagus*, and in both cases processing appears to be less efficient in *Epifagus*, as reflected by a higher proportion of larger to smaller transcripts in the parasite than in tobacco leaf chloroplasts. This last difference may relate more to developmental than evolutionary factors, as it has been shown in maize, for example, that for most plastid genes the ratio of unspliced to spliced transcripts is much higher in nonphotosynthetic tissues (meristems and roots) than in green leaves [1]. In this respect, it is important to emphasize that the cDNA analyses shown in Fig. 3 should not be interpreted quantitatively, i.e., they should not be taken to mean that most or all of the transcripts from these three intron-containing genes have been spliced. This is because the much smaller spliced transcripts would be expected to out-compete any unspliced transcripts in both the reverse transcription and PCR-amplification reactions used to generate the cDNAs. Accurate measurement of amounts of spliced and unspliced transcripts requires other approaches, such as the S1-nuclease-protection assay used by Barkan [1].

Some of the qualitative differences seen in transcript patterns between *Epifagus* and tobacco are, however, an expected consequence of their evolutionary divergence, namely, the extensive gene loss leading to reduction of transcription unit size in *Epifagus*. At the extreme, far simpler transcript patterns, lacking the abundance of multiple larger transcripts observed in tobacco, were seen for *rps2* (Fig. 1) and *rps14* (data not shown). For both genes, all other genes (four and two, respectively) with which they are cotranscribed in photosynthetic plants are entirely or largely absent from *Epifagus* ptDNA (Fig. 2; [41]). Operon reduction (from 12 genes in tobacco to eight in *Epifagus*; see Fig. 2) is also probably responsible for some of the less impressive differences observed in transcript patterns produced by the *rpl2* probe (Fig. 1).

RNA editing in Epifagus *plastids*

A total of 1401 nucleotides of cDNA sequence was determined in this study for the intron-containing genes *rps12*, *rpl2* and *clpP*. Comparison to the corresponding published [41] plastid gene sequences revealed a single, conventional C→U edit. A similarly low incidence of RNA editing has been observed in both photosynthetic and nonphotosynthetic tissues of photosynthetic plants (e.g., [19, 23]; reviewed in [18]). It is worth noting that several putative edits were initially observed in comparing cDNA and genomic sequences from *Epifagus*, but further examination showed that all but one actually reflected population-level DNA variation. The *bona fide rps12* edit is the only difference that persisted when RNAs and DNAs were sequenced from the very same plant. This provides an important cautionary lesson for all studies purporting to find evidence of RNA editing. The *rps12* edit in *Epifagus* changes codon 74 from TCA (serine) to UUA (leucine). Plastid *rps12* genes from 14 other diverse land plants encode only leucine at this position (J. Allen and J. Palmer, unpublished), implying that the *rps12* edit observed in *Epifagus* evolved relatively recently.

Curiously, the C residue that is edited in *Epifagus rps12* is embedded in an octanucleotide sequence (ATAATTCA) that is perfectly complementary to part of one of several putative guide RNA consensus sequences identified in plastid and plant mitochondrial genomes by Maier *et al.* [24]. Furthermore, about 200 bp upstream of this edit site, located within intron 1 of *rps12*, is a perfect inverted repeat (TGAATTAT) of the edit-octanucleotide sequence. It is tempting to speculate that, prior to splicing, an intramolecular 8-bp stem forms between the edit site and its intronic complement, and that formation of this secondary structure in the *rps12* transcript guides the editing apparatus to this site.

That RNA editing is very rare in the three genes examined in this study, and probably in the *Epifagus* plastid genome as a whole, may be of some functional importance relative to two previous observations regarding the *Epifagus* plastid genome. First, inferences that virtually all *Epifagus* plastid gene products have faster rates of sequence evolution than those of photosynthetic plants [27, 41, 42, 44] are supported by the observation that RNA editing is infrequent and thus unlikely to 'correct' extensively modified gene sequences. Second, *Epifagus* plastid DNA encodes a grossly incomplete set of tRNA genes, but the choice of codons in *Epifagus* plastid proteins suggests full availability of all tRNAs [27, 42]. Minimal RNA editing for these three protein genes decreases the likelihood that a full set of tRNAs could somehow be specified from the *Epifagus* plastid genome by post-transcriptional modification of its reduced set of tRNA genes and supports our earlier hypothesis of extensive plastid import of cytoplasmic tRNAs [27, 42].

MatK and group II intron splicing in plastids

We have shown that five of the six group II introns in *Epifagus* ptDNA are spliced properly (we have not examined the sixth, in *rpl16*). These results are significant in light of the curiously freestanding maturase gene, *matK*, present in *Epifagus* ptDNA. Although freestanding in *Epifagus*, *matK* is contained within the group II intron of *trnK* in all photosynthetic land plants (*trnK* is absent from *Epifagus* ptDNA). Mohr *et al.* [26] showed that MatK (the product of *matK*), although having lost most of the reverse transcriptase character of group II-encoded proteins, has retained a ca. 100 amino acid domain ('X') characteristic of nearly all group II maturases (i.e., splicing factors) and which they speculate binds to intron RNA. MatK protein was recently detected in potato chloroplasts [6] and has been shown to bind to *trnK* RNA [22]. Because matK is freestanding in *Epifagus*, we [41] and others [8, 22, 26] have hypothesized that, unlike other group II maturases, MatK has the novel [20] property of assisting splicing of introns that are distantly related to the one in which it normally resides.

To better understand the evolution and potential functional divergence of MatK in *Epifagus*, we have analyzed its sequence relatedness to MatK proteins from photosynthetic plants. *Epifagus* MatK is unusually divergent in two ways. First, as shown in Fig. 4, it is considerably shorter (439 amino acids) than all other MatK proteins (497–515 amino acids) due to an N-terminal truncation. Second, as is true of virtually all *Epifagus* plastid gene products [41, 42], MatK is changing faster than normal in *Epifagus*. *Epifagus* and tobacco are more closely related to each other than to any of the other species compared in Fig. 4, yet the amino acid identities between MatK of *Epi-*

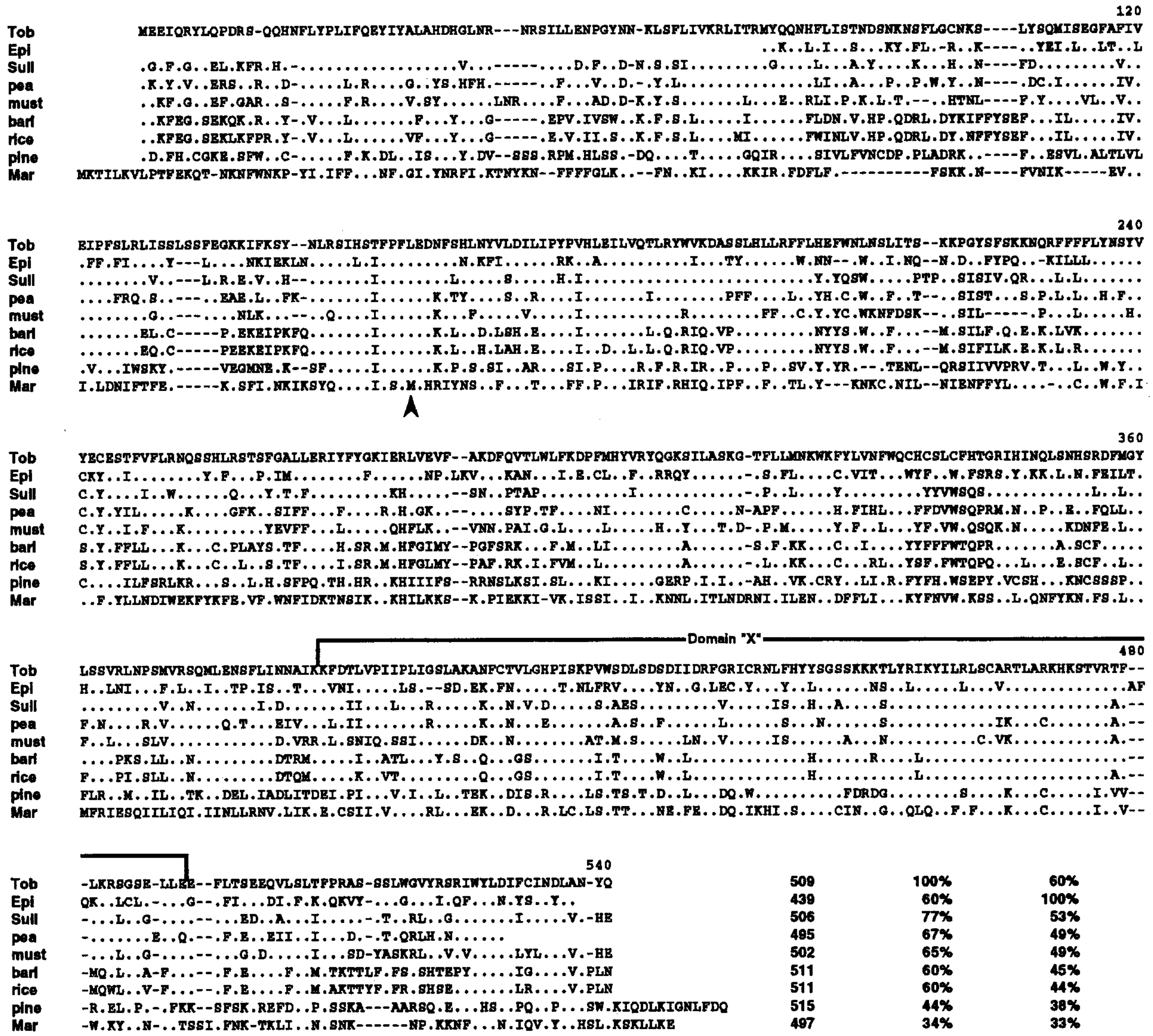

Fig. 4. Unusual divergence of MatK in *Epifagus*. An alignment of MatK amino acid sequences from tobacco (tob [33]), *Epifagus virginiana* (Epi [41]), *Sullivantia sullivantii* (Sull [13]), pea [2], mustard (must [28]), barley (barl [3, 31]; GenBank accession number X64129), rice [11], *Pinus contorta* (pine [21]), and *Marchantia polymorpha* (Mar [38]) was first made using CLUSTAL [9] and then improved by eye. Dots represent identical amino acids; dashes represent gaps. The bracketed region is the ca. 100 amino acid long domain X, a putative RNA-binding domain common to almost all group II intron maturases [26]. The length of each protein is given in the first column following the alignment, followed by the percent amino acid identity (gaps excluded) to MatK from tobacco (second column) and *Epifagus* (third column). This alignment contains several adjustments compared to published inferences of the MatK start site. Completion [41] of the partial *matK* sequence of Morden *et al.* [27] reveals that the 439 amino acid long *Epifagus* MatK probably starts at a position 200 bp downstream from the tobacco start site; translation of the *Epifagus* sequence from the tobacco start site yields a stop codon 60 codons downstream of this ATG as a consequence of a short frameshift mutation in *Epifagus*. Although Mohr *et al.* [26] extended the inferred length of *Epifagus* MatK at 505 amino acids by postulating an error in our published [41] *matK* sequence that would eliminate this frameshift, reinspection of our sequencing films confirms the frameshift, and we therefore stand by our original inference that *Epifagus* MatK is 439 amino acids in length. In accordance with the proposal of Tsudzuki *et al.* [37], the reported [38] 370 amino acid MatK (see arrow) of *Marchantia* has been extended to 497 amino acids via introduction of an extra T in a run of 12 T's at positions 26, 745–26, 756. The published reading frames of the sequences from mustard [28] and rice [11] begin 22 and 31 codons, respectively, before the start sites proposed in Fig. 3. However, the nucleotide sequence of this region is mustard contains five frameshift insertions of 1, 2 or 5 bp relative to tobacco,

fagus and other angiosperms (e.g., *Epifagus*-pea, 49%) are much lower than between tobacco and the same plants (i.e., tobacco-pea, 67%; see values at end of Fig. 4). Despite its rapid sequence divergence and unusual N-terminal truncation, *Epifagus* MatK is nonetheless highly colinear with other plastid MatK proteins (Fig. 4).

If MatK is indeed a maturase in *Epifagus*, then which of the six group II introns still present in its genome might it act on? Two lines of evidence suggest that the single group II intron in *rpl2* is a good candidate for being a site of MatK action in *Epifagus*. First, Hess *et al.* [8] showed that the *rpl2* intron is the only one of seven group II introns examined whose splicing is blocked in four different nuclear mutants of barley, each of which is severely deficient in its level of plastid ribosomes. They concluded that a functioning plastid translational apparatus is necessary, most likely to produce MatK, in order for splicing of the *rpl2* intron to occur. Second, although still distantly related overall, the *rpl2* intron is nonetheless the most closely related plastid intron in sequence to the intron in *trnK* (the host gene for *matK* in photosynthetic plants), based both on formal phylogenetic analysis (data not shown) and on the classification scheme of Michel *et al.* [25]. A second candidate for MatK-assisted splicing in *Epifagus* is the *cis*-spliced intron 2 of *rps12*, whose splicing has also been shown to be blocked in plastid ribosome-deficient mutants of barley (T. Börner, W. R. Hess and T. Hüebschmann, pers. comm.). Of the other four group II introns present in *Epifagus*, two (the *trans*-spliced intron 1 of *rps12* and the single intron in *rpl16*) are spliced in the barley mutants (and are therefore probably not dependent on MatK for their splicing; [8], T. Börner, W. R. Hess and T. Hüebschmann, pers. comm.), while the two *clpP* introns could not be analyzed because they are absent from the barley *clpP* gene (S. Downie and J. Palmer, unpublished). In summary, we suggest that there are two good candidates for MatK action in *Epifagus*: the single intron in *rpl2* and the second intron in *rps12*.

Acknowledgements

We are grateful to Thomas Börner, Wolfgang Hess, Thomas Hüebschmann, Karsten Liere, and Gerhard Link for communicating their unpublished results; Thomas Börner, L. Michelle Bowe, and B. Swalla for their comments on the manuscript; and John Mullet for providing a clone of barley *matK* for our reanalysis of its sequence (see GenBank accession number X64129). This work was supported by funds from the NIH (GM-35087) to J.D.P., the Alfred P. Sloan Foundation (90-3-5) to K.H.W. and the NSF (DEB-9120258) to C.W.D.

References

1. Barkan A: Tissue-dependent plastid RNA splicing in maize: transcripts from four plastid genes are predominantly unspliced in leaf meristems and roots. Plant Cell 1: 437–445 (1989).
2. Boyer SK, Mullet JE: Pea chloroplast tRNA^{Lys} (UUU) gene: transcription and analysis of an intron-containing gene. Photosyn Res 17: 7–22 (1988).
3. Boyer SK, Mullet JE: Sequence and transcript map of barley chloroplast *psbA* gene. Nucl Acids Res 16: 8184 (1988).
4. de Pamphilis CW, Palmer JD: Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature 348: 337–339 (1990).
5. Doyle JJ, Doyle JL: A rapid DNA isolation procedure for small quantities of fresh tissue. Phytochem Bull 19: 11–15 (1987).
6. du Jardin P, Portetelle D, Harvengt L, Dumont M, Wathelet B: Expression of intron-encoded maturase-like polypeptides in potato chloroplasts. Curr Genet 25: 158–163 (1994).
7. Fukuzawa H, Kohchi T, Sano T, Shirai H, Umesono K, Inokuchi H, Ozeki H, Ohyama K: Structure and organization of *Marchantia polymorpha* chloroplast genome. III.

whereas after the proposed ATG start all insertion/deletion events are in multiples of 3 bp [28]. Similarly, the upstream domain of rice is 95% identical to that of barley except that barley has a 1 bp deletion relative to the rice sequence 26 bp upstream of the proposed start site of *matK*. Resequencing (see GenBank accession number X64129) of the beginning of the barley *matK* gene extends the length of its protein from 504 [3, 31] to 511 amino acids.

Gene organization of the large single copy region from *rbcL* to *trnI* (CAU). J Mol Biol 203: 333–351 (1988).

8. Hess WR, Hoch B, Zeltz P, Hübschmann T, Kössel H, Börner T: Inefficient *rpl2* splicing in barley mutants with ribosome-deficient plastids. Plant Cell 6: 1455–1465 (1994).
9. Higgins DG, Sharp PM: Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5: 151–153 (1988).
10. Hildebrand M, Hallick RB, Passavant CW, Bourque DP: *Trans*-splicing in chloroplasts: the *rps12* loci of *Nicotiana tabacum*. Proc Natl Acad Sci USA 85: 372–376 (1988).
11. Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M: The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. Mol Gen Genet 217: 185–194 (1989).
12. Hudson GS, Mason JG, Holton TA, Koller B, Cox GB, Whitfeld PR, Bottomley W: A gene cluster in the spinach and pea chloroplast genomes encoding one CV_1 and three CV_0 subunits of the H^+-ATP synthase complex and the ribosomal protein S2. J Mol Biol 196: 283–298 (1987).
13. Johnson LA, Soltis DE: *matK* DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Syst Bot 19: 143–156 (1994).
14. Kanno A, Hirai A: A transcription map of the chloroplast genome from rice (*Oryza sativa*). Curr Genet 23: 166–174 (1993).
15. Kohchi T, Ogura H, Umesono K, Yamada Y, Komano T, Ozeki H, Ohyama K: Ordered processing and splicing in a polycistronic transcript in liverwort chloroplasts. Curr Genet 14: 147–154 (1988).
16. Kohchi T, Umesono K, Ogura Y, Komine Y, Nakahigashi K, Komano T, Yamada Y, Ozeki H, Ohyama K: A nicked group II intron and *trans*-splicing in liverwort, *Marchantia polymorpha*. Nucl Acids Res 16: 10025–10036 (1988).
17. Koller B, Fromm H, Galun E, Edelman M: Evidence for *in vivo trans* splicing of pre-mRNAs in tobacco chloroplasts. Cell 48: 111–119 (1987).
18. Kössel H, Hoch B, Maier RM, Igloi GL, Kudla J, Zeltz P, Freyer R, Neckermann K, Ruf S: RNA editing in chloroplasts of higher plants. In: Brennicke A, Kück (eds) Plant Mitochondria, pp. 93–102, VCH Chemie, Weinheim (1993).
19. Kuntz M, Camara B, Weil JH, Schantz R: The *psbL* gene from bell pepper (*Capsicum annuum*): plastid RNA editing also occurs in non-photosynthetic chromoplasts. Plant Mol Biol 20: 1185–1188 (1992).
20. Lambowitz AM, Belfort M: Introns as mobile genetic elements. Annu Rev Biochem 62: 587–622 (1993).
21. Lidholm J, Gustafsson P: A three-step model for the rearrangement of the chloroplast *trnK-psbA* region of the gymnosperm *Pinus contorta*. Nucl Acids Res 19: 2881–2887 (1991).
22. Liere K, Link G: RNA binding activity of the *matK* protein encoded by the chloroplast *trnK* intron from mustard (*Sinapis alba* L.). Nucl Acids Res 23: 917–921 (1995).
23. Maier RM, Hoch B, Zeltz P, Kössel H: Internal editing of the maize chloroplast *ndhA* transcript restores codons for conserved amino acids. Plant Cell 4: 609–619 (1992).
24. Maier RM, Neckermann K, Hoch B, Akhmedov NB, Kössel: Identification of editing positions in the *ndhB* transcript from maize chloroplasts reveals sequence similarities between editing sites of chloroplasts and plant mitochondria. Nucl Acids Res 20: 6189–6194 (1992).
25. Michel F, Umesono K, Ozeki H: Comparative and functional anatomy of group II catalytic introns – a review. Gene 82: 5–30 (1989).
26. Mohr G, Perlman PS, Lambowitz AM: Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucl Acid Res. 21: 4991–4997 (1993).
27. Morden CW, Wolf KH, dePamphilis CW, Palmer JD: Plastid translation and transcription genes in a nonphotosynthetic plant: intact, missing, and pseudo genes. EMBO J 10: 3281–3288 (1991).
28. Neuhaus H, Link G: The chloroplast $tRNA^{Lys}$ (UUU) gene from mustard (*Sinapis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Curr Genet 11: 251–257 (1987).
29. Ohto C, Torazawa K, Tanaka M, Shinozaki K, Sugiura M: Transcription of ten ribosomal protein genes from tobacco chloroplasts: a compilation of ribosomal protein genes found in the tobacco chloroplast genome. Plant Mol Biol 11: 589–600 (1988).
30. Sambrook J, Fritsch EF, Maniatis T: Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1989).
31. Sexton TB, Jones JT, Mullet JE: Sequence and transcriptional analysis of the barley ctDNA region upstream of *psbD-psbC* encoding *trnK*(UUU), *rps16*, *trnQ*(UUG), *psbK*, *psbI*, and *trnS*(GCU). Curr Genet 17: 445–454 (1990).
32. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M: The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5: 2043–2049 (1986).
33. Sugita M, Shinozaki K, Sugiura M: Tobacco chloroplast $tRNA^{Lys}$ (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. Proc Natl Acad Sci USA 82: 3557–3561 (1985).

34. Sugiura M: The chloroplast genome. Plant Mol Biol 19: 149–168 (1992).
35. Sugiura M, Shinozaki K, Tanaka M, Hayashida N, Wakasugi T, Matsubayashi T, Ohto C, Torazawa K, Meng BY, Hidaka T, Zaita N: Split genes and *cis/trans* splicing in tobacco chloroplasts. In: von Wettstein D, Chua NH (eds) Plant Molecular Biology, pp. 65–76. Plenum, New York (1987).
36. Tanaka M, Wakasugi T, Sugita M, Shinozaki K, Sugiura M: Genes for the eight ribosomal proteins are clustered on the chloroplast genome of tobacco (*Nicotiana tabacum*): similarity to the S10 and *spc* operons of *Escherichia coli*. Proc Natl Acad Sci USA 83: 6030–6034 (1986).
37. Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, Sugiura M: Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ*, *trnK*, *psbA*, *trnI*, and *trnH* and the absence of *rps16*. Mol Gen Genet 232: 206–214 (1992).
38. Umesono K, Inokuchi H, Shiki Y, Takeuchi M, Chang Z, Fukuzawa H, Kohchi T, Shirai H, Ohyama K, Ozeki H: Structure and organization of *Marchantia polymorpha* chloroplast genome II. gene organization of the large single copy region from *rps12'* to *atpB*. J Mol Biol 203: 299–331 (1988).
39. Weglöhner W, Subramanian AR: Nucleotide sequence of a region of maize chloroplast DNA containing the 3′ end of *clpP*, exon 1 of *rps12* and *rpl20* and their cotranscription. Plant Mol Biol 18: 415–418 (1992).
40. Wolfe KH, Morden CW, Palmer JD: Ins and outs of plastid genome evolution. Curr Opin Genet Devel 1: 523–529 (1991).
41. Wolfe KH, Morden CW, Palmer JD: Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89: 10648–10652 (1992).
42. Wolfe KH, Morden CW, Ems SC, Palmer JD: Rapid evolution of the plastid translational appartus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J Mol Evol 35: 304–317 (1992).
43. Wolfe KH, Morden CW, Palmer JD: Small single-copy region of plastid DNA in the non-photosynthetic angiosperm *Epifagus virginiana* contains only two genes: differences among dicots, monocots and bryophytes in gene organization at a non-bioenergetic locus. J Mol Biol 223: 94–104 (1992).
44. Wolfe KH, Katz-Downie DS, Morden CW, Palmer JD: Evolution of the plastid ribosomal RNA operon in a nongreen parasitic plant: accelerated sequence evolution, promoter deletion, and tRNA pseudogenes. Plant Mol Biol 18: 1037–1048 (1992).
45. Zaita N, Torazawa K, Shinozaki K, Sugiura M: *Trans* splicing in vivo: joining of transcripts from the 'divided' gene for ribosomal protein S12 in the chloroplasts of tobacco. FEBS Lett 210: 153–156 (1987).