Conservation of selection on matK following an ancient loss of its flanking intron

Aaron M. Duffy a,⁎, Scot A. Kelchner b, Paul G. Wolf a

a Department of Biology, Utah State University, Logan, UT 84322, USA
b Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA

A R T I C L E   I N F O
Article history:
Received 10 October 2008
Received in revised form 29 January 2009
Accepted 10 February 2009
Available online 21 February 2009
Received by G. Theissen

Keywords:
Group II intron
Secondary structure
trnK
Intron-encoded protein
Fern

A B S T R A C T
The chloroplast gene trnK and its associated group II intron appear to be absent in a large and ancient clade that includes nearly 90% of fern species. However, the maturase protein encoded within the intron (matK) is still present and located on the boundary of a large-scale inversion. We surveyed the chloroplast genome sequence of clade-member Adiantum capillus-veneris for evidence of a still present but fragmented trnK intron. Lack of signature structural domains and sequence motifs in the genome indicate loss of the trnK intron through degradation in an ancestor of the clade. In plants, matK preferentially catalyzes splicing of the trnK intron, but may also have a generalist function, splicing other group II introns in the chloroplast genome. We therefore tested whether a shift in selective constraint has occurred after loss of the trnK intron. Using previously unavailable sequences for several ferns, we compared matK sequences of the intron-less fern clade to sequences from seed plants and ferns with the intron and found no significant differences in selection among lineages using multiple methods. We conclude that matK in ferns has maintained its apparently ancient and generalized function in chloroplasts, even after the loss of its co-evolved group II intron. Finally, we also present primers that will allow amplification and nucleotide sequencing of the phylogenetically useful matK gene in additional fern taxa.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

The plant chloroplast gene matK has long sparked the interest of molecular evolutionary biologists. Its open reading frame is associated with a group II intron (Fig. 1A) that interrupts the coding sequence of tRNA^Lys(UUU) and it shows a much faster rate of sequence evolution than many other chloroplast genes (Wolfe et al., 1992; Hilu and Liang, 1997). The discovery of matK’s presence in the highly reduced plastomes of nonphotosynthetic plants (Wolfe et al., 1992; Ems et al., 1995) surprised many researchers and pointed to its probable role as a maturase that catalyzes the splicing reactions of more than one group II intron in the chloroplast genome (e.g., Ems et al., 1995; Liere and Link, 1995; Vogel et al., 1999).

Many known group II introns possess their own intron-encoded protein (IEP), which assists in splicing its host intron (Toor et al., 2001; Hausner et al., 2006). However, nearly all of the 20 or so group II introns in plant plastomes show severe degradation of their maturase open reading frames. This condition suggests that plant chloroplast group II introns no longer need to maintain their own splicing co-factor, an observation that many consider to be a strong indication of matK’s role as a generalist maturase (reviewed by Hausner et al., 2006).

The purported generalist function of matK would be unusual for an intron maturase. Introns and their IEPs are thought to have co-evolved: similar phylogenetic relationships are found among IEP sequences as are found among their intron RNA structures (Toor et al., 2001). Hence, there is a strong likelihood that the two components of a complete intron sequence (the intron itself, and its IEP open reading frame) are indelibly linked in terms of structure and function. Any shift in the intron’s primary sequence or secondary and tertiary structure would likely correlate with a change of its IEP sequence (and function) if it is to successfully pass through the filters of natural selection.

Interestingly, the catalysis link between host intron and IEP continues to be strong in matK, even though it may also function as a generalist splicing co-factor for many chloroplast introns. Vogel et al. (1997, 1999) have shown in vivo that matK is required for trnK intron splicing in barley, and it will preferentially catalyze this reaction over the presumably less specific splicing of additional chloroplast introns, particularly those of structural subclass IIA (Liere and Link, 1995).

The unusual role of matK led us to question how selective constraints might vary for matK sequences in cases where the gene is no longer associated with its principal target, the trnK intron. This condition was observed in the chloroplast genome of the fern Adiantum capillus-veneris, which possesses matK and shows evidence of matK transcription (Wolf et al., 2004), yet appears to lack trnK and
its intron (Wolf et al., 2003; Fig. 1B). Genome mapping studies indicate that the loss of trnK and its intron is associated with an ancient inversion event in the ancestor of a large clade of leptosporangiate ferns (Hasebe and Iwatsuki, 1992, Stein et al., 1992; Roper, 2007). This is an old lineage (~265 mya), which includes nearly 90% of the approximately 11000 extant fern species (Pryer et al., 2004). This genome structure would also explain the failure to obtain matK sequence from ferns: the PCR primers used in other plants are located in the missing trnK exons (e.g. Hilu and Liang, 1997; Hilu et al., 2003; Hauser et al., 2006).

Cases of trnK intron loss with matK retention have been inferred only twice before in plants: once in the highly reduced chloroplast genome of the achlorophyllous parasitic plant Epifagus virginiana (Wolfe et al., 1992; Ems et al., 1995) and once in the chloroplast genome of Cuscuta reflexa, a parasitic plant with reduced photosynthetic activity (Funk et al., 2007). However, these are recent changes that may not shared with other extant taxa of each lineage, whereas the fern example is likely due to a very old event resulting in an intron-less matK in the majority of extant fern species. Ferns therefore present an opportunity to study the possible shift of selective constraints on an EIP (matK) after isolation from its co-evolved intron (the trnK intron) in a well-sampled and ancient clade of plants.

In this study, we use computational methods to establish whether the trnK intron is indeed absent from the A. capillus-veneris chloroplast genome or is instead divided but still functional, by searching for conserved intron sequence elements and intron-specific secondary structures. We then test whether any of matK’s major protein domains have experienced a shift in selective constraints after the loss of the trnK intron. We did this by: (1) obtaining matK sequences for additional fern taxa with and without a contiguous trnK intron, (2) comparing patterns of nucleotide and amino acid conservation across matK sequences in ferns and also seed plants, and (3) comparing rates of nonsynonymous to synonymous nucleotide substitutions in these groups using several methods. We also present primers for amplifying and sequencing a portion of the matK gene in ferns that are missing the trnK intron.

2. Materials and methods

2.1. Search for trnK intron in Adiantum

The trnK intron is not present in its expected location in the A. capillus-veneris chloroplast genome (Wolf et al., 2003), although this observation alone does not confirm its complete absence. Recombination involving group II introns has led to many known cases in which intron fragments are dislocated in a genome yet retain their function through trans-splicing mechanisms (Chapdelaine and Bonen, 1991; Bonen, 1993; Ems et al., 1995; Knoop et al., 1997; Jarrell et al., 1988; Malek and Knoop, 1998; Qiu and Palmer, 2004). Hence, it is important to our study that we first confirm an absolute loss of a conserved trnK intron sequence in the A. capillus-veneris chloroplast genome prior to interpreting any shift in patterns of selection among matK sequences that might be indicative of a change in the protein’s function. We devised a method that first tests for the uniqueness of sequence within highly conserved structures of the trnK intron and then identifies several small sequence elements that would assist in locating intron structural fragments within the A. capillus-veneris chloroplast genome.

We constructed estimated secondary structure models for the trnK intron/matk sequence regions from Petliia borealis, Marchantia polymorpha, Sphagnum platyphyllum, Cycas panzhihuaensis, Pinus thunbergii, and Atropa belladonna (Supplementary Table 1; see Hauser et al., 2006 for additional trnK intron secondary structure models) using the domain-by-domain folding strategy of Kelchner (2002). The resulting RNA structures were then compared to identify conserved structural elements. Sequences of these structural elements (Table 1) were used to search the intergenic spacer regions immediately upstream and downstream of the matK ORF in A. capillus-veneris and all positive matches were explored using localized RNA folding by Mfold (Zuker, 2003) to survey for possible II intron secondary structures. Sequences from both the upstream and downstream noncoding regions surrounding the A. capillus-veneris matK ORF were also folded with Mfold using an arbitrary “sliding window” approach with multiple sequence lengths shifting outward from the matK region in an attempt to identify helices that could be homologous with trnK intron domains. In particular, we looked for evidence of trnK intron domains 1 through 4, which would likely be proximal upstream of the matK ORF (Fig. 1).

The nucleotide sequence of segment 5 of the trnK intron is generally well conserved among land plants (Supplementary Fig. 1) and should be recognizable if it is present in the chloroplast genome. The 34 nucleotides of domain 5 from P. borealis, M. polymorpha, S. platyphyllum, C. panzhihuaensis, P. thunbergii, and A. belladonna were used individually for BLAST searches against the GenBank database to verify

![Fig. 1. The trnK/matK region as found in most angiosperms (A) and in the fern Adiantum capillus-veneris (B) roughly to scale. Transcription proceeds from left to right. The tether and domain X are relatively conserved functionally regions of the matK protein. Numbers 1 through 6 designate the location of domains 1 through 6 of the trnK intron, respectively.](image-url)
that these nucleotides would identify domain 5 in other land plant trnK sequences, against the Angiopteris evecta chloroplast genome sequence to verify that they would identify domain 5 in a fern, and directly against the A. capillus-veneris chloroplast genome sequence as an attempt to locate trnK domain 5 in this genome. Each BLAST (blastn) search employed a low complexity filter with an expect threshold of 10 and word size of 11, and the return limit was set at the maximum 1000 matches from a Eukaryota virtual database.

2.2. DNA extraction, amplification, and sequencing

Taxa were chosen to represent most major fern lineages (Supplementary Table 1). Genomic DNA for new sequences was extracted using Qiagen DNEasy kits or a CTAB method (Doyle and Doyle, 1987). Published seed plant matK primers failed to amplify matK in fern taxa, and attempts to design universal fern matK primers using the matK sequences of A. capillus-veneris and A. evecta were not successful due to the lack of highly conserved regions. Sequences for matK in Osmanda cinnamomea, Marsilea mutica, and Dicksonia antarctica were obtained by amplifying large fragments using primers in flanking genes and primer walking. Using these sequences and the matK sequences of A. capillus-veneris and A. evecta, two sets of fern matK primers were designed: one set for the more recently derived "modern" ferns and one set for the more basal "early" ferns (Fig. 2). These primers were used to amplify and sequence portions of matK from Lygodium japonicum and Pteridium aquilinum. Because these primers are located within matK, to obtain the complete matK sequences, flanking regions were amplified and sequenced using these primers combined with primers in flanking genes. We also used the published matK sequences of the lycophyte Huperzia lucidula, the seed plants Pinus koraiensis, Amborella trichopoda, Nymphaea alba, Magnolia dealbata, Helianthus annuus, Triticum aestivum, and Cucumis pectinata, and the monilophytes Pilosotum nudum, and Ophioglossum petiolatum from GenBank (Supplementary Table 1). The matK sequences of Isoetes engelmannii and Equisetum arvense were extracted from their unpublished complete chloroplast genomes (K. Karol, personal communication).

2.3. Phylogenetic analysis

Nucleotide sequences were converted to amino acid sequences and aligned using clustalW with manual refinement informed by the nucleotide sequence and following the alignment principles of Kelchner (2000). The alignment was then converted back to nucleotide sequences for subsequent analyses. A single nucleotide gap was opened after nt 149 in Huperzia to shift the reading frame so that the rest of the sequence would align. To match known RNA editing in Adiantum (Wolf et al., 2004), ACG (T) was changed to ATG (M) at nt 5, TCA (S) was changed to TTA (L) at nt 497, TCC (S) was changed to TTC (T) at nt 559 in Isoetes, TGA was changed to CGA (R) and TAA was changed to CAA (Q) at nt 256 and 262 of Dicksonia and at nt 262 and 268 of Pteridium, and TGA was changed to CGA (R) at nt 250 and 256 of Marsilea. Making these changes extended the ORFs to approximately the same length as other matK sequences. Sections of the alignment containing indels that could not be aligned unambiguously were not included in subsequent analyses (Fig. 3).

We performed Bayesian Metropolis coupled Markov chain Monte Carlo analyses using MrBayes v3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The data were partitioned by codon position and each partition was assigned its own model of nucleotide substitution (GTR + I + gamma) as determined using MrModeltest 2.2 (Nylander, 2004), a modified version of Modeltest 3.6 (Posada and Crandall, 1998). Using a random starting tree, we performed three separate runs with four chains each for 1250000 generations, sampling every 1000 generations. We plotted the log probability of observing the data by generation to detect stationarity and discarded the first 250 samples as "burnin." We pooled the post-burnin trees from each run and calculated a majority-rule consensus tree.

2.4. Tests of relaxed selective constraint

We compared the patterns of nucleotide and amino acid conservation along the length of the matK sequence in ferns with the trnK intron, ferns without the trnK intron, and seed plants. By focusing on the pattern of conservation rather than on absolute levels of conservation, we reduce bias due to different sample sizes, relationships between the groups, and whether taxa are equivalent between groups.

Several methods are available for detecting shifts in selection using the ratio of nonsynonymous to synonymous substitution rates (dN/dS; reviewed by Yang and Bielawski, 2000). If matK has been under relaxed constraint in the chloroplast genomes of ferns since the loss of trnK and its intron, we should be able to detect an increase in dN/dS ratios relative to ferns that have retained trnK and its intron. Those differences might be expected to be most evident in the active regions of the gene: domain X, which is associated with splicing activity, and the "tether region," which represents part of a reverse-transcriptase (RT) domain (Mohr et al., 1993; Haussner et al., 2006; Fig. 1). We tested for shifts in selection specifically on these domains, as well as on the gene as a whole. However, when using these much smaller subsets of the data, no differences in the pattern of substitution rates among the lineages were found and no statistical tests found significance that was not also present for the entire matK sequence, so only the results for the entire matK sequence will be presented.

![Fig. 2. "Universal" fern matK primers and their approximate annealing positions within a representative fern matK sequence. "Modern" fern primers were designed to anneal to regions conserved within Marsilea, Dicksonia, and Adiantum matK sequences. "Early" fern primers were designed to anneal to regions conserved between Angiopteris, Osmanda and Marsilea.](image_url)
We used the program package PAML (Yang, 1997) to evaluate models that allow ratios of nonsynonymous to synonymous substitution rates to vary among lineages in order to test whether levels of selective constraint vary among the lineages of ferns without the trnK intron, ferns with the trnK intron, and the rest of the tree. This method (Yang, 1998) requires a priori hypotheses of which branches on the tree vary, and models are compared using likelihood ratio tests. We constructed several models (Table 2) ranging from the simplest, with all 3 lineages assigned a single dN/dS ratio (Model A), to the most general, with each of the three lineages assigned a separate ratio (Model D). PAML and another program package, HYPHY (Kosakovsky Pond et al., 2005), were both used to evaluate a free branch model, which assigns a separate ratio to each branch of the tree.

We also performed HYPHY GA Branch Analysis, using a genetic algorithm to assign dN/dS ratios. Each branch on the phylogenetic tree was assigned to a dN/dS ratio class with the optimal number of classes determined from the data (Kosakovsky Pond and Frost, 2005). We then looked for patterns in the way branches were assigned to ratio classes that might suggest differences between lineages.

3. Results

3.1. Search for trnK intron in Adiantum

Using the domain-by-domain folding method on the trnK intron/matK sequence regions of land plants, we recovered core secondary structural models consistent with well supported group II intron models (Michel et al., 1989) and identified several conserved sequence elements related to structures shared across taxa (Table 1). Most of these short (4–7 nucleotides) elements returned positive matches when used in localized sequence searches upstream and downstream of the matK ORF in A. capillus-veneris, but none was

Fig. 3. Representative matK nucleotide alignment including all fern sequences, one seed plant, and one lycophyte sequence. Nucleotide sequences were translated to amino acid sequences and aligned using ClustalW. Regions with indels could not be aligned unambiguously and were removed from the alignment and are not shown. Shaded portions of the alignment represent the tether region including RT subdomains V, VI, and VII (amino acids 207–270) and domain X (amino acids 293–393).
embedded in sequences that could be folded into RNA secondary structures resembling those of the trnK intron from which the element was derived. We found no evidence of potentially homologous helical structures and their associated conserved sequence elements that would indicate a trnK intron fragment. We conclude that matK is no longer associated with a divided, but functional, trnK intron in the chloroplast genome of A. capillus-veneris.

BLAST searches using domain 5 sequences from each of six land plant taxa returned trnK intron sequences (but not other group II intron sequences) from nearly all the major plant groups including the fern, *Angiopteris evecta*. This is consistent with our expectation that the nucleotide sequence of domain 5 in the trnK intron is both unique to the trnK intron and highly conserved in land plant chloroplast genomes. These BLAST searches did not return a domain 5 sequence for the trnK intron in *A. capillus-veneris*. When these six domain 5 sequences were used for BLAST searches directly against the *A. capillus-veneris* chloroplast genome, only a few 11-nucleotide complements were returned with low expect values (0.045 to 1.9), but none demonstrated a secondary structure that was potentially homologous to domain 5 of a group II intron. None of the sequences identified a sequence in *A. capillus-veneris* that resembled the trnK intron in either its primary or secondary structure, strongly favoring the hypothesis that the trnK intron, or recognizable fragments thereof, is no longer present in this chloroplast genome.

### 3.2. Phylogenetic analysis

The *matK* consensus tree topology (Supplementary Fig. 2) is similar to published phylogenetic hypotheses with just a few differences, primarily on branches that are not strongly supported in analyses using other genes. *Equisetum* is placed as sister to the rest of the monilophytes, rather than sister to *Angiopteris* as hypothesized by Pryer et al. (2004) and Qiu et al. (2006) or sister to the core leptosporangiates (*Marsilea, Dicksonia, Adiantum, and Pteridium*) as hypothesized by Magallón and Sanderson (2005). Within the seed plants, relationships differ somewhat from published hypotheses (Hilu et al., 2003; Magallón and Sanderson, 2005; Qiu et al., 2006). Despite these differences, the relationships for the comparisons we want to make are consistent with published hypotheses — the ferns and the seed plants form monophyletic sister groups, and the ferns without the trnK intron (hereafter referred to as the “K-minus clade”) form a well supported monophyletic group nested within ferns with the trnK intron (hereafter referred to as the “K-plus ferns”). The consensus tree topology could be biased by the small number of taxa included and also by the limitations of any single-gene phylogenetic analysis, so a constraint tree based on published topologies was used for subsequent analyses. Results of analyses using both topologies were compared to ensure that topology differences did not affect our inferences (data not shown).

#### 3.3. Tests of relaxed selective constraint

The proportion of variable nucleotide alignment sites for K-plus ferns, the K-minus clade, and seed plants was lower than the percent total amino acid variation (K-plus ferns: 70% vs. 81%; K-minus clade: 67% vs. 82%; seed plants: 58% vs. 74%). This is consistent with values reported for other plant groups (Hilu and Liang, 1997) and suggests a lack of strong functional constraint on the gene as a whole since nucleotide variation typically translates into lower amino acid variation in functionally constrained genes (Graur and Li, 2000). However, any direct comparison of the percentage variation between K-plus ferns, the K-minus clade, and seed plants is probably inappropriate as these values are likely biased by the different sample sizes used and the phylogenetic relationships among the groups.

Whereas the percentage variation and absolute number of variable nucleotides and amino acids may be biased, the pattern of variation or distribution of relatively conserved or variable regions along the sequence can be compared directly. The pattern of nucleotide and amino acid variation along the length of the *matK* sequence is similar in K-plus ferns, the K-minus clade, and seed plants (Fig. 4), and is similar to the pattern described by Hilu et al. (2003) in the *matK* sequence for a large sample of angiosperms. The three groups have highly variable sequences but a region with less variability in one group tends to have less variability in the other groups as well. All three groups have regions of relatively conserved sequence located approximately 300 nucleotides (100 amino acids) from the 5′ end of our alignment, and approximately 200 nucleotides (65 amino acids) from the 3′ end of our alignment, with the latter region corresponding to part of domain X. These same low variability regions have been identified in other plant groups as well, including the nonphotosynthetic, trnK-lacking *Epipagus* (Hilu and Liang, 1997; Young and DePamphilis, 2000). All three groups also show a slight reduction in sequence variability within the tether region, but it is less conserved than the other two regions.

Likelihood ratio tests using PAML did not find significant differences between models varying in the levels of heterogeneity in the dN/dS ratio among lineages (Table 2, Table 3). The model with one ratio for the K-minus clade and one for the rest of the tree (Model B) does not fit the data significantly better than a single ratio model (Model A), suggesting that the dN/dS ratio for the K-minus clade is not different from the rest of the tree. This is the case even if the dN/dS ratio for the K-plus ferns is allowed to vary (Model C vs. Model D). To allow for the possibility that purifying selection on *matK* was
diminished following the inversion and loss of the intron but then increased again, we tested specifically for a change in $d_N/d_S$ on the branch where the inversion that separated $matK$ from $trnK$ and its intron occurred (Model E). We tested whether a model with one ratio for that single branch and another ratio for the rest of the tree fits the data better than the one ratio model (Model A) (Table 2, Table 3). The $d_N/d_S$ ratios for nearly all branches were less than one, suggesting mild purifying selection on the gene as a whole, consistent with tests of selection on $matK$ in other taxa using protein side-chain composition (Barthet and Hilu, 2008). There is no obvious increase in $d_N/d_S$ ratios on the branches associated with the K-minus clade. In general, values for the branches associated with the K-minus clade fall between those for the K-plus fern lineages and those for the seed plant lineages. There are three branches in the tree with $d_N/d_S$ values greater than one, but these all have very short branch lengths so it is not clear whether the high ratios are valid or are due to chance. However, these large or infinite $d_N/d_S$ values are not assigned to branches associated with the K-minus clade. These results do not suggest that the $matK$ sequence of the K-minus clade is under relaxed selective constraint or positive selection relative to the other groups. If anything, the K-minus clade is under slightly stronger purifying selection than the seed plants and slightly weaker purifying selection than the K-minus ferns.

We used a free ratio model to infer a separate $d_N/d_S$ ratio for each branch of the tree (Fig. 5). The free model analysis was performed using both PAML and HYPHY, and though the two software packages gave slightly different values for $d_N$ and $d_S$, the ratios were identical. Likelihood ratio tests show that the free ratio model fits the data significantly better than the one ratio model (Model A) (Table 2, Table 3). The likelihood ratio tests show that the free ratio model fits the data significantly better than the one ratio model (Model A) (Table 2, Table 3). The $d_N/d_S$ ratios for nearly all branches were less than one, suggesting mild purifying selection on the gene as a whole, consistent with tests of selection on $matK$ in other taxa using protein side-chain composition (Barthet and Hilu, 2008). There is no obvious increase in $d_N/d_S$ ratios on the branches associated with the K-minus clade. In general, values for the branches associated with the K-minus clade fall between those for the K-plus fern lineages and those for the seed plant lineages. There are three branches in the tree with $d_N/d_S$ values greater than one, but these all have very short branch lengths so it is not clear whether the high ratios are valid or are due to chance. However, these large or infinite $d_N/d_S$ values are not assigned to branches associated with the K-minus clade. These results do not suggest that the $matK$ sequence of the K-minus clade is under relaxed selective constraint or positive selection relative to the other groups. If anything, the K-minus clade is under slightly stronger purifying selection than the seed plants and slightly weaker purifying selection than the K-minus ferns.

The HYPHY Genetic Algorithm Branch Analysis assigned each branch to one of four classes ranging from $d_N/d_S = 0.139$ to $0.531$ (Fig. 6). None of the branches associated with the K-minus clade was assigned to the highest $d_N/d_S$ class. All but one of these branches were assigned to the 2nd highest (and most common) class. Within all the ferns, the only two branches assigned to the highest $d_N/d_S$ class were the branch rooting Angiopteris, Equisetum, and the Leptosporangiate ferns, and the branch leading to Ophioglossum. The rest of the
If *matK* was functioning primarily as a splice factor for *trnK* and its intron, once *trnK* was lost, we would expect *matK* to be released from this particular selective constraint and to either accumulate substitutions and become degraded, or optimize to its role as a generalist catalyst of splicing IIA introns, or less likely, to gain new function. However, the similar distributions of conserved nucleotides and amino acids in the *matK* sequences of seed plants, K-plus ferns, and the K-minus clade suggest that substitutions are not being accumulated differently in the *matK*s that exist without the *trnK* intron. Additionally, analyses of dN/dS ratios using several methods — including the entire coding sequence as well as specifically focusing on the functional domains — do not support the hypothesis that *matK* is under relaxed selective constraint in taxa without *trnK* and its intron, as would be expected if the gene was no longer functional in these taxa. If anything, the dN/dS analyses suggest that *matK* in the ferns without the *trnK* intron has been under a similar level of selective constraint to that of the seed plants, whereas in the ferns with the *trnK* intron, *matK* is slightly more constrained.

The presence of *matK* cDNAs in *A. capillus-veneris* (Wolf et al., 2004) and another fern in the K-minus clade (Barthet and Hilu, 2007) suggests that *matK* continues to function despite the absence of the *trnK* intron. If *matK* were performing a new role in the genomes of ferns without the *trnK* intron, we would expect to see evidence of positive selection as the sequence responded to different selective forces, but there is no evidence of positive selection in any of the analyses. So, even after the loss of *trnK* and its intron, *matK* is apparently performing a largely similar function, consistent with the hypothesis that *matK* is involved in splicing other group IIA introns in the chloroplast (Vogel et al., 1999). Other genes in the *A. capillus-veneris* chloroplast genome with group II introns are *atpF, clpF, rpl2, rps12, trnA, trnI*, and *trnV* (Wolf et al., 2003; Funk et al., 2007).

The hypothesized inversion (Hasebe and Iwatsuki, 1992; Stein et al., 1992; Roper, 2007) that appears to have disrupted the intron is shared by members of a clade that diverged approximately 265 mya (Pryer et al., 2004) and includes nearly 90% of all fern species. This indicates that the *matK* condition in these taxa is both ancient and stable and contrasts with the other examples of inferred maturase isolation — the reduced chloroplast genomes of *E. virginiana* (Wolf et al., 1992; Ems et al., 1995) and *C. reflexa* (Funk et al., 2007), which are members of much younger angiosperm clades.

Because the genomes of nearly 90% of all extant fern species do not have the typical *trnK/matK* arrangement, and many of the primers used to amplify *matK* in other plant groups are within *trnK* or its intron, ferns have been underrepresented in previous analyses of *matK*. The high rate of nucleotide substitution in *matK* makes it difficult to design primers that are conserved across large taxonomic groups, and taxon-specific primers are often required (Hilu et al., 2003). A recent proposal of two options for three-gene land plant DNA barcodes included *matK* in both options, but noted that additional work was needed to improve the performance of the primer sets (Chase et al., 2007). We have so far been unable to design a single set of primers to amplify *matK* across all the fern taxa used in this study, but designed two sets: one in sequence conserved across the early diverging fern lineages and one in sequence conserved across the more recently derived lineages within the K-minus clade (Fig. 2). There is some overlap in the taxa that each set will amplify, but so far the “early” fern primers have been tested successfully on *Osmunda, Lygodium, Gleichenia, Marsilea*, and *Dicksonia* and the “modern” fern primers have been tested successfully on *Adiantum, Dicksonia*, and *Pteridium*. Because these primers are located within the *matK* sequence they cannot amplify the entire *matK* gene. But, by combining these primers with knowledge of the gene order surrounding *matK* in ferns with and without *trnK* (Fig. 1), the entire *matK* sequence may be amplified using published *cblB* or *rps16* primers for the upstream flank and *ndhB* (K-minus clade) or *psba* (K-plus ferns) for the downstream flank (Roper, 2007). The primers designed for this study should allow *matK* to be amplified and sequenced for most fern taxa.

---

4. Discussion

Our investigation of the atypical *trnK/matK* condition in the *A. capillus-veneris* chloroplast genome (which is shared by other ferns in the K-minus clade) leads us to conclude that the *trnK* intron has been lost even though its IEP, *matK*, is retained. The position of the *matK* ORF at the border of an inferred inversion event had allowed for the possibility that the *trnK* intron was still functional in the genome as a *trans*-spliced, divided intron. If so, *matK* would be flanked by one or more of the *trnK* intron domains, and the highly conserved, diagnostic domain 5 sequence would be detectable in the genome. However, we found no evidence for any conserved *trnK* intron secondary structure or sequence motifs flanking the *matK* ORF, or for conserved *trnK* domain 5 sequence elements in the complete *A. capillus-veneris* chloroplast genome. These observations favor the hypothesis that the *trnK* intron was disrupted by an inversion event in an ancestor of the higher ferns, failed to persist by *trans*-splicing mechanisms in the chloroplast genome, and was subsequently degraded and lost.
4.1. Evolutionary significance of maturase isolation

Close functional association of introns and their encoded maturases is common among prokaryotic, protist, and fungal group II introns (Lambowitz and Perlman, 1990; Moran et al., 1995; Saldanha et al., 1999; Huang et al., 2005), many of which appear to have co-evolved with their IEPs (Costa et al., 1997; Toor et al., 2001; Zimmerly et al., 2001). Liere and Link (1995) established that matK binds preferentially to trnK intron pre-mRNAs in the mustard Sinapis alba when given the opportunity. Although matK is typical in that it maintains a tight functional association with its encoding intron, it is also unusual in its apparent contribution to splicing reactions of other group IIA introns (Vogel et al., 1999). One plausible history for the advent of matK’s generalist nature is given here. It is thought that mobile group II introns invaded ancestral chloroplast genomes sometime prior to the divergence of the charophytes and embryophytes (Toor et al., 2001; Sanders et al., 2003). As most extant mobile group II introns use IEPs for retrotranspositioning (Zimmerly et al., 2001), it is likely that each invading intron carried a maturase ORF in its domain 4 and one of these was the ancestor of matK. Consistent with this hypothesis are the observations that matK is always associated with the trnK intron when both are present, and that chloroplast genomes that lack group II introns also lack matK (Turtle et al., 1999; Lemieux et al., 2000).

After their establishment in plastid genomes, most chloroplast group II introns seem to have shifted from a specific interaction with their IEPs to a more relaxed association with a variety of splice-assisting proteins such as those encoded by nuclear genes crs1 and crs2 (Jenkins et al., 1997), plastid ribosomes (Hess et al., 1994), and other as yet to be determined factors (Vogel et al., 1996; Matsuura et al., 2001; Toor et al., 2001). Such associations would have lessened the constraints on chloroplast IEPs, eventually leading to the degradation of ORFs in nearly all group II introns. Although there are at least four protein-assisted pathways for splicing different assemblages of introns in plant chloroplasts, introns belonging to subclades IIA may continue to rely on matK (Jenkins et al., 1997; Vogel et al., 1999). The presence of matK in the reduced chloroplast genomes of E. virginiana and C. reflexa, and our example of its preservation after trnK intron loss in higher ferns, seem to support this model.

To our knowledge, the higher ferns represent only the third reported case of intron loss with IEP retention — and the first where the intron loss was attributable to an inversion. The expected co-dependence of a group II intron and its IEP makes such an event remarkable. However, plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single...