Expression and Function of the Chloroplast-encoded Gene matK.

Michelle Marie Barthet

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Biological Sciences

K. W. Hilu, Chair
E. Beers
G. Gillaspy
J. Sible
R. A. Walker

February 9, 2006
Blacksburg, Virginia

Keywords: MatK, chloroplast, maturase, fast-evolving, Orchidaceae

Expression and Function of the chloroplast-encoded gene matK.

Michelle Marie Barthet

ABSTRACT

The chloroplast *matK* gene has been identified as a rapidly evolving gene at nucleotide and corresponding amino acid levels. The high number of nucleotide substitutions and length mutations in *matK* has provided a strong phylogenetic signal for resolving plant phylogenies at various taxonomic levels. However, these same features have raised questions as to whether *matK* produces a functional protein product. *matK* is the only proposed chloroplast-encoded group II intron maturase. There are 15 genes in the chloroplast that would require a maturase for RNA splicing. Six of these genes have introns that are not excised by a nuclear imported maturase, leaving MatK as the only candidate for processing introns in these genes. Very little research has been conducted concerning the expression and function of this important gene and its protein product. It has become crucial to understand *matK* expression in light of its significance in RNA processing and plant systematics. In this study, we examined the expression, function and evolution of MatK using a combination of molecular and genetic methods. Our findings indicate that *matK* RNA and protein is expressed in a variety of plant species, and expression of MatK protein is regulated by development. In addition, *matK* RNA levels are affected by light. Furthermore, genetic analysis has revealed that although MatK has a high rate of amino acid substitution, these substitutions are not random but are constrained to maintain overall chemical structure and stability in this protein. We have also identified an alternate start codon for *matK* in some plant species that buffers

indels (insertions and deletions) in the open reading frame (ORF) that are not in multiples of three in the gene sequence. Usually, indels not in multiples of three result in frame shifts that destroy the reading frame. Our results indicate that an out-of-frame *matK* start codon in some orchids compensates for these otherwise deleterious indels. This research represents the first in-depth analysis of *matK* gene expression and contributes to several fields of biology including plant systematics, genetics and gene expression.

ACKNOWLEDGMENTS

I would like to especially thank Dr. Khidir Hilu for giving me the opportunity to work on such an amazing project and giving me the confidence and fortitude to continue in my career as a molecular biologist. Dr. Hilu has given me guidance in every aspect of graduate development, from research design to growth as an academic professor. He has been a great advisor and mentor. I would like to thank the members of my committee, Dr. Jill Sible, Dr. Richard Walker, Dr. Eric Beers, and Dr. Glenda Gillaspy for all their help and support with every step of this project. They have been tremendous source of support and comfort and have made my time here a great educational experience as well as a lot of fun.

I would like to thank past and present members of the Hilu lab. Special thanks to Sheena Friend, who has not only been a great lab mate but also a great friend, and Sunny Drysdale. I thank Sabrina Majumder and Chelsea Black, the two most dedicated undergraduates I've ever met, for all their help with various aspects of my research. Thanks to Rohit Kumar and Jenny Whitten, who has kept all of us in the lab young. My thanks to our colleagues Dr. Dietmar Quandt, Thomas Börsch, and Elke Doring for their help in plant collections and understanding plant systematics.

Special thanks to my parents, Dr. Joseph and Connie Barthet, and all my family who have supported and encouraged me throughout my academic career. Thanks to all my extended family, Dr. Bill and Joyce Parker, who have been there for me and my husband during these stressful and difficult times. I wish to especially thank Dr. Scott Parker for his advise concerning statistical calculations, reading through numerous drafts of all the chapters in this dissertation, and for always being a wonderful husband.

DEDICATION

This dissertation is dedicated to my parents who have encouraged my interests in biological sciences since I was a young child.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

Chapter 1

LITERATURE REVIEW

INTRODUCTION

The *mat*K gene has two unique features that underscore its importance in molecular biology and evolution: its fast evolutionary rate and its putative function as a group II intron maturase. *mat*K is a chloroplast-encoded gene nested between the 5' and 3' exons of *trnK*, tRNA-lysine (Sugita, Shinozaki, and Sugiura 1985) in the large single copy region of the chloroplast genome. The mode and tempo of *matK* evolution is distinct from other chloroplast genes. Rate of nucleotide substitution in *matK* is three times higher than that of the large subunit of Rubisco (*rbcL*) and six fold higher at the amino acid substitution rate (Johnson and Soltis 1994; Olmstead and Palmer 1994), denoting it as a fast- or rapidly-evolving gene. This high nucleotide and amino acid substitution rate provides high phylogenetic signal for resolving evolutionary relationships among plants at all taxonomic levels (Hilu and Liang 1997; Soltis and Soltis 1998; Hilu et al. 2003). The resolution achieved with sequences of *matK* is equal to using up to eleven other genes combined (Hilu and Liang 1997; Hilu et al. 2003).

In addition to the high rate of substitution, *matK* also displays varying number and size of indels (insertions and deletions) (Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003). Most indels identified in *matK* have been found in multiples of three, conserving the reading frame (Hilu and Liang 1997; Hilu and Alice 1999; Whitten, Williams, and Chase 2000; Hilu et al. 2003). However, the presence of indels, the high substitution rate, and premature stop codons found in some plant families (Kores et al. 2000; Kugita et al. 2003) raises the question of whether a gene with these features maintains stable protein structure and function.

Previous studies at the nucleotide and amino acid level indicated that *matK* does not have homogenous substitution rates across its ORF but, instead, exhibits varying rates of substitution

(Hilu and Liang 1997). One highly conserved region of 448 bp is located in the 3' end of *matK* (Hilu and Liang 1997). Sequence analysis indicated that this region displayed homology to domain X of mitochondrial group II intron maturases (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). Although *matK* contains many indels throughout the reading frame (Hilu and Liang 1997; Hilu and Alice 1999; Hilu et al. 2003), domain X was found to lack indels (Hilu and Liang 1997; Hilu and Alice 1999). Based on this lack of indels, it was suggested that functional constraint exists in domain X (Hilu and Liang 1997).

matK is the only gene found in the chloroplast genome of higher plants that contains this putative maturase domain (Neuhaus and Link 1987). There are 16 group II introns nested within 15 chloroplast genes (Kohchi et al. 1988; Ems et al. 1995; Maier et al. 1995), which would require a maturase for intron splicing and proper protein translation. Maturases are splicing factors that aid in splicing and folding group II introns (Vogel et al. 1997). Studies of the white barley mutant *albostrians*, a chloroplast ribosomal mutant, demonstrated that although some group II introns were processed by an imported nuclear maturase, there were at least six plastid genes (*trnK*, *atpF, trnI*, *trnA*, *rpl2*, and *rps12 cis*) with group II introns that would require a chloroplast maturase for splicing (Vogel et al. 1997; Vogel, Borner, and Hess 1999). Western blot analysis indicated that a protein of approximately 60 kDa is produced by the *matK* gene in barley (Vogel, Borner, and Hess 1999). Identification of a MatK protein product and demonstration of a lack of splicing for some group II introns in the *albostrians* mutant suggest a potential functional role for MatK as a group II intron maturase in the chloroplast.

I hypothesized that matK is an expressed gene in the chloroplast genome and that it functions as a maturase to process and splice out group II introns. I have used both molecular techniques and bioinformatics to address questions regarding the expression and function of

matK. To investigate the molecular expression of *matK*, I first examined the RNA transcript produced by this gene and determined if this transcript can be separated from the precursor transcript of *trnK*. Second, a synthesized peptide antigen was used to produce an antibody against MatK to corroborate the RNA data with protein expression. Third, I examined the influence of light and development on *matK* RNA and protein levels.

In order to gain information on how a gene with such a high evolutionary rate is able to maintain stable protein structure and could still function in the chloroplast, we examined the evolution of MatK at the chemical level using bioinformatics. Chemically conserved amino acid replacement, the most prominent form of amino acid replacement in functionally or structurally important regions of protein-coding genes, would act as silent mutations to minimize the impact on protein structure and/or function (Clarke 1970; Graur 1985; Graur and Li 1988; Wolfe and dePamphilis 1998). Previous study by Cheng et al. (2005) indicated that functionally important sites may tend to rely less on the direct amino acid sequence, and more on chemical conservation. Using amino acid and chemical composition and variability, we were able to determine regions of structural verses functional constraint in MatK and compare this constraint with that of a slow evolving gene, a pseudogene, and another group II intron maturase. This analysis also provided information regarding chemical composition and variability throughout green plants. Further, transmembrane domains were predicted in the putative MatK amino acid sequence, which could suggest a membrane-associated location for this protein. In addition, I identified a unique out-of-frame alternate start codon for MatK in the Orchidaceae (orchid family), a plant family previously though to contain *matK* as a pseudogene in several species (Kores et al. 2000; Goldman et al. 2001).

Utilizing both molecular biology and bioinformatics, we have provided a detailed analysis of *matK* expression, putative functional roles for MatK protein in development and photosynthesis, and demonstrated regions of structural and functional importance in the protein. Information from this study will provide insight into how mode and tempo of nucleotide and amino acid substitution can impact protein structure of a rapidly evolving gene, provide insight into the molecular expression of the only putative chloroplast encoded maturase, assess its potential function in the chloroplast machinery, and further promote the utility of this gene in plant systematics and molecular evolution.

LITERATURE REVIEW

Structure of the matK gene

For most land plants, the *mat*K gene is nested between the two exons of *trn*K*,* tRNA-lysine (Figure 1.1). The *mat*K ORF is approximately 1500 bp in most angiosperms (Hilu, Alice, and Liang 1999), corresponding to around 500 amino acids for the translated protein product. The structure of this gene includes indels of various length and number (Hilu and Alice 1999; Hilu et al. 2003). For example, the *Epifagus mat*K gene contains a 200-bp deletion at the 5' end compared to tobacco *matK* (Ems et al., 1995). Nucleotide substitution rates are not evenly distributed across the *matK* ORF, but instead *matK* has regions displaying high mutation rates (Hilu and Liang 1997). The third codon position tends to have a slightly higher mutation rate than the first and second codon position, suggesting neutral or purifying selection in this gene (Young and dePamphilis 2000). Predicted amino acid sequence analysis from various plants

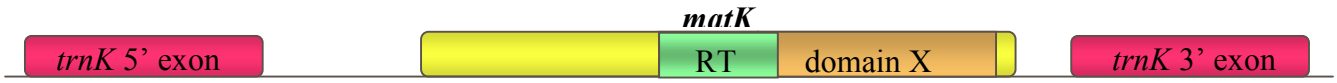


Figure 1.1. The location of *matK* between the 5' and 3' exons of *trnK*. The *matK* ORF is highlighted in yellow, region of the reverse transcriptase (RT) domain in the *matK* ORF is highlighted in green and domain X is highlighted in brown. The *trnK* exons are highlighted in red.

indicated a highly conserved region close to the 3' end of this gene that lacks indels (Hilu and Liang 1997). This region contains 448 bp, is called domain X (Figure 1.1), and has similarity to a conserved functional domain found in mitochondrial group II intron maturases (Sugita et al., 1985; Neuhaus and Link, 1987).

Group II introns

Introns can be classified into one of three groups: I, II, or III. Group I introns are considered primarily self-splicing but an accessory protein factor for intron excision is required in some cases (Saldanha et al. 1993; Geese and Waring 2001). Group II introns often require a maturase for excision and can only self-spice under non-physiological conditions (Saldanha et al. 1993; Noah and Lambowitz 2003). Group III introns are a modified form of group II introns (Mohr, Perlman, and Lambowitz 1993). Group II introns can be further subdivided based on structural characteristics into IIA and IIB (Michel, Umesono, and Ozeki 1989). The cellular location of group I and group II introns is very similar with both being dispersed in mitochondria and chloroplast genomes and can be found in mRNA, tRNA and rRNA genes of each of these organelles (Ferat and Michel 1993; Cho et al. 1998; Vogel, Borner, and Hess 1999; Bhattacharya et al. 2002; Rudi, Fossheim, and Jakobsen 2003). Unlike group II introns, group I introns have also been found in nuclear genes (Saldanha et al. 1993). For group I introns, the splicing reaction involves a guanine as the attacking group to break the phosphodiester bond on the 5' side of the intron (Alberts et al. 1994). Group II introns, however, use adenine as the attacking nucleotide to form the lariet structure (Alberts et al. 1994; Kelchner 2002). Group II introns can be either autocatalytic, the ancestral characteristic, or require splicing factors to form the lariat structure and be spliced out of the RNA transcript (Alberts et al. 1994).

Group II intron evolution

Group II introns have been found in fungal and plant mitochondria as well as chloroplasts (Mohr, Perlman, and Lambowitz 1993). In addition, this group of introns has also been identified in the proteobacterium *Azotobacter vinelandii* and the cyanobacterium *Calothrix,* bacteria related to the probable ancestor of the mitochondria and chloroplast, respectively (Ferat and Michel 1993). Thus, group II introns are an ancestral character of organelle evolution. Several group II introns contain an open reading frame (Sugita, Shinozaki, and Sugiura 1985; Mohr, Perlman, and Lambowitz 1993; Moran et al. 1995; Saldanha et al. 1999). This open reading frame often encodes a reverse-transcriptase/maturase that is capable of transposing the intron into a new location (Mohr, Perlman, and Lambowitz 1993; Moran et al. 1995). A study using *Saccharomyces cervisiae* demonstrated this mobility by showing that mitochondrial group II introns can be inserted in *cox1* alleles that were originally missing these introns (Moran et al. 1995). This mobility provided an indication of the evolution of ORFs in these introns. The ORF not only contained domain X for maturase activity, but also a reverse-transcriptase (RT)-like domain (Mohr, Perlman, and Lambowitz 1993). Phylogenetic analysis of the RT domain of intron-encdoed proteins (IEPs) indicated sequence homology to the RT domains of retroviruses, such as HIV-1 (Blocker et al. 2005), and non-long–terminal repeat (LTR) retroelements (Mohr, Perlman, and Lambowitz 1993; Moran et al. 1995; Filippo and Lambowitz 2002; Blocker et al. 2005). Thus, these group II intron-encoded proteins are evolutionarily related to retroviruses and retroelements known to have this same form of mobility within the genome.

Group II intron maturases

It was discovered that some mitochondrial group II introns contain ORFs encoding their own splicing factors termed 'maturases' (Saldanha et al. 1993). Maturases are thought be required as translated protein for *in vivo* splicing of some group II introns (Mohr, Perlman, and Lambowitz 1993). Although the maturases of yeast and *Lactococcus* only process the intron in which they are encoded (Matsuura, Noah, and Lambowitz 2001; Cui et al. 2004; Rambo and Doudna 2004), at least two maturases, CRS2 and MatK, can splice several different introns (Liere and Link 1995; Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999; Osteimer et al. 2003). Both of these maturases are thought to function in the chloroplast (Liere and Link 1995; Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999; Osteimer et al. 2003).

Sequence analysis of 34 intron-encoded ORFs identified three domains that are generally maintained in most maturases: a reverse transcriptase (RT) domain, domain X, and a zinc finger-like region (Mohr, Perlman, and Lambowitz 1993). The RT domain is thought to be an ancestral character, remnant from the origin of these introns as non-LTR retrotransposons (Mohr, Perlman, and Lambowitz 1993; Moran et al. 1995). The RT domain is active in certain maturases (Moran et al. 1995; Matsuura et al. 1997; Saldanha et al. 1999; Wank et al. 1999). The zinc finger-like domain comprises the core of the DNA endonuclease activity of these maturases (Moran et al., 1995), while the maturase activity is retained in domain X (Mohr et al., 1993). All three regions of the maturase enzyme are thought to act in concert to achieve group II intron mobility (Saldanha et al. 1999; Singh et al. 2002; Rambo and Doudna 2004). However, only the RT domain and domain X are required for the splicing activity (Cui et al. 2004; Rambo and Doudna 2004). A general mechanism for this mobility/maturase activity is through maturase domain binding to the group II intron followed by folding the intron to form a lariat structure by bringing

the attacking adenine to the 5' end of the intron (Mohr, Perlman, and Lambowitz 1993; Saldanha et al. 1999; Kelchner 2002). This results in splicing the intron lariat structure out of the precursor RNA. The maturase then remains bound to the excised RNA to form a ribonucleoprotein particle (RNP) (Saldanha et al. 1999). Next, the DNA endonuclease domain creates a double-strand break at the target insertion site (Saldanha et al. 1999). Once the break is formed, the reverse transcriptase domain is activated to integrate the excised group II intron into a new site by DNA-primed reverse transcription (Saldanha et al. 1999). Although the RT and DNA endonuclease activity have been well studied in these introns, the maturase activity is less well understood.

Group II intron maturases studied thus far include primarily the *Lactococcus* LtrA maturase protein (Matsuura et al. 1997; Saldanha et al. 1999; Wank et al. 1999; Singh et al. 2002; Noah and Lambowitz 2003), yeast mitochondrial maturases (Moran et al. 1994) and a few nuclear-encoded maturases (Jenkins, Khulhanek, and Barkan 1997; Mohr and Lambowitz 2003). A mechanism of splicing has been defined for the LtrA maturase (Matsuura, Noah, and Lambowitz 2001; Singh et al. 2002; Rambo and Doudna 2004), and preliminary research has indicated aspects of nuclear-encoded maturase function (Jenkins, Khulhanek, and Barkan 1997; Osteimer et al. 2003; Ostheimer et al. 2005). However, studies on mitochondrial maturases have not defined a mechanism of group II intron processing. The mechanism of bacterial maturase LtrA is the most defined, and shown to be influenced by magnesium concentration (Matsuura et al. 1997; Noah and Lambowitz 2003). LtrA binds to a high affinity region on the group II intron referred to as DIVa (Matsuura, Noah, and Lambowitz 2001; Singh et al. 2002). This region is also the site of the ORF for the maturase in the intron (Matsuura, Noah, and Lambowitz 2001; Singh et al. 2002; Rambo and Doudna 2004). Once bound, the protein interacts with other

conserved regions in the intron to form the final lariat structure for excision (Matsuura, Noah, and Lambowitz 2001; Singh et al. 2002).

The nuclear-encoded maturase CRS2 is transported to the chloroplast where it processes nine of the ten chloroplast-encoded group IIB introns (Osteimer et al. 2003). CRS2 forms a complex with CAF1 and CAF2 for binding and processing group IIB introns (Osteimer et al. 2003). However, no other details of the splicing mechanism have been defined. CRS1 is a nuclear-encoded chloroplast maturase that acts only on the group IIA intron of *atpF* in the chloroplast (Till et al. 2001). However, the group IIA intron of *atpF* also requires an additional, yet to be identified, factor from the chloroplast for complete excision (Jenkins, Khulhanek, and Barkan 1997). Since the protein product of *matK* is the only putative group II intron maturase encoded in the choroplast genome (Neuhaus and Link 1987), I hypothesize that the additional chloroplast-encoded factor for intron excision in *atpF* is MatK.

MatK

Amino acid sequence analysis of the *mat*K gene from plants has revealed a conserved domain X region (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). Due to the homology of this domain domain X of group II intron maturases of the mitochondria, it has been hypothesized that *mat*K encodes a maturase that functions to splice group II introns (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). In order to analyze this possible function of MatK, Liere and Link (1995) initiated a set of *in vitro* binding assays using bacterially expressed MatK from *Sinapis alba* (Linn.). Since mitochondrial maturases are known to aid in splicing of the intron in which they reside (Matsuura, Noah, and Lambowitz 2001; Cui et al. 2004; Rambo and Doudna 2004), this same test was done with the *in vitro* expressed MatK and the full-length *trn*K

ORF-containing gene region. They reported the formation of a *trnK* RNA-*trnK* ORF protein complex. This demonstrated that MatK binds to the group II intron in which it resides in order to splice itself out (Liere and Link 1995). Further assays involving other group II intron-containing chloroplast genes indicated that the bacterially-expressed MatK had specificity to *trnK* and *trnG* precursor transcripts but not to at least one other chloroplast group II intron containing gene (Liere and Link, 1995). These assays supported the idea that MatK functions as a group II intron maturase but did not demonstrate the actual splicing activity of this maturase.

Null mutants have often been used to determine the exact function and necessity of a protein *in vivo*. No null mutants exist for *mat*K; however, the barley chloroplast mutant, referred to as white barley *albostrians,* lacks functional plastid ribosome activity (Vogel et al. 1997; Vogel, Borner, and Hess 1999). This mutant would then be incapable of translation of all chloroplast protein. Therefore, if all chloroplast introns were spliced using a nuclear-encoded maturase only, this barley mutant should still retain normal group II intron processing. Studies by Vogel et al. (1997, 1999) indicated that although chloroplast group IIB and other non-group IIA introns were still efficiently spliced out in the *albostrians* mutant, group IIA introns were not. Thus, the essential splicing factor for group IIA intron removal was lacking. Further analysis by Vogel et al. (1997) demonstrated the presence of the *trn*K/*matK* precursor RNA in the *albostrians* mutant but a lack of mature *trnK*. The absence of mature *trnK* transcript suggested a lack of the splicing factor needed to excise the group IIA intron, which also includes the *matK* ORF, from the precursor to form mature *trnK* (Vogel et al. 1997). Since the expressed *mat*K protein product is thought to be a splicing maturase that aids in the removal of group II introns including the one in which it resides (Matsuura, Noah, and Lambowitz 2001; Cui et al. 2004; Rambo and Doudna 2004), it was hypothesized that the inability of this barley mutant to

efficiently translate MatK protein prevented group IIA intron removal in the plastid organelle. Vogel et al. (1999) using a MatK-specific antibody, was able to support their hypothesis by demonstrating an absence of MatK protein product in the barley mutant. Their studies concluded that it was the lack of the translated MatK protein product in these ribosome-deficient barley mutants that prevented group IIA intron splicing (Vogel et al. 1997; Vogel, Borner, and Hess 1999). The barley mutant studies provide further evidence that *mat*K is expressed and functions in the chloroplast genome.

As indicated earlier, group II intron maturases generally have three domains, a RT domain, domain X, and a DNA endonuclease domain (Mohr, Perlman, and Lambowitz 1993). *matK* lacks the DNA endonuclease domain and contains only remnants of the RT domain (Mohr, Perlman, and Lambowitz 1993). In addition to this unusual characteristic for a maturase, MatK has been found to bind to group II introns other than the one in which it is encoded (Liere and Link 1995). This would suggest that MatK functions to process group II introns in a novel mechanism where it either does not require the RT domain or only requires the elements of the RT domain that remains in MatK. Further, the matK gene was found in the residual chloroplast genome of the parasitic beechdrop *Epifagus virginiana* (Ems et al. 1995). More than 60 percent of *E. virginiana* chloroplast genes, mostly photosynthetic genes, have been lost over evolutionary time (Ems et al. 1995), presumably due to lack of necessity. Surprisingly, the *trn*K gene has been eliminated from this vestigal genome, but the *matK* gene remains (Ems et al. 1995). This implies that the *matK* gene serves some essential function that has not been replaced by a nuclear-encoded gene. Transcript analysis from *E. virginiana* did indicate proper group II intron splicing (Ems et al. 1995), thus it is possible that the splicing maturase protein, MatK, is retained and expressed in this plant.

matK in plant systematics

Evolutionary studies in plants utilize several methodologies in order to obtain the most clearly defined robust phylogenetic trees. Molecular sequence data has revolutionized evolutionary studies and enhanced the resolution of phylogenetic trees immensely. Genes used in plant systematics fall along a spectrum of rate of substitution with the two extremes representing fast or slow-evolving genes. Which one to use is usually determined by the level of phylogenetic analysis one wishes to pursue. Slow-evolving genes, such as *rbcL* and *atpB*, have high sequence conservation among plant groups. This high sequence conservation allows for good resolution above the family level, but low resolution below this level (Hilu and Liang 1997; Goldman et al. 2001). Fast-evolving genes, such as *matK,* provide enough characters for evolutionary analysis at the family level and below (Hilu and Liang 1997; Goldman et al. 2001). The *matK* gene is considered to be fast-evolving due to the fact that it has a high rate of substitutions and more variable sites compared to other genes (Olmstead and Palmer 1994; Johnson and Soltis 1995; Hilu and Liang 1997; Soltis and Soltis 1998). The *matK* ORF is not homogenous in rate of nucleotide substitution but instead contains regions of variable rates of substitution (Johnson and Soltis 1995; Hilu and Liang 1997). One of the conserved regions in *mat*K is the putative functional domain X (Hilu and Liang 1997).

In phylogenetic analysis, phylogenetically informative characters are those characters that are variable but are not the product of homoplasy (parallel evolution) and not so variable that alignment between specific taxonomic levels cannot be accomplished. *mat*K provides many informative characters in regions that do not have excessive variability nor excessive conserved sequence and can be aligned to determine evolutionary relationships from the species to the

divisional or even higher taxonomic levels (Johnson and Soltis 1995; Hilu and Liang 1997; Hilu, Alice, and Liang 1999; Hilu et al. 2003). *mat*K has been useful for determining phylogenetic histories for several plant taxa including the Saxifragaceae (Johnson and Soltis 1995), Orchidaceae (Kores et al. 2000; Whitten, Williams, and Chase 2000; Goldman et al. 2001; Kores et al. 2001), the asterids (Bremer et al. 2002), as well as across all angiosperms taxa (Hilu et al. 2003). Phylogenetic studies using *mat*K have produced more robust trees than had previously been determined using multiple genes (Hilu et al. 2003).

Despite this extensive use of the *mat*K gene for phylogenetic analysis, some still dispute its expression and functionality in the chloroplast genome, stating that *mat*K is just a pseudogene (Kores et al. 2000; Whitten, Williams, and Chase 2000; Goldman et al. 2001; Kores et al. 2001). Nonetheless, the researchers stating this fact have also utilized *mat*K for some of their phylogenetic studies. The rationale behind their denotion of *matK* as a pseudogene include that stop codons were found within the *mat*K ORF, indels occur that create frame-shift mutations, and that there was an equal level of substitution for all three codon positions (Kores et al. 2000; Whitten, Williams, and Chase 2000; Kores et al. 2001). Pseudogenes can fall into two categories: genes that are not transcribed and genes that are transcribed but contain premature stop codons and produce truncated, non-functional proteins (Mighell et al. 2000; Balakirev and Ayala 2003). The stop codons found within the *mat*K ORF of members of the Orchidaceae may place *matK* in the second category of pseudogene that produces a truncated protein. However, this result would depend on the reading frame translated.

Contrary to some of the findings of the *matK* gene sequence from the orchids, sequence analysis from nine representative species across the plant kingdom demonstrated that the indels within the *mat*K gene occurred in multiples of three, conserving the *mat*K reading frame (Ems et

al. 1995). Additionally, frame-shift mutations found in the 3' region of *mat*K of the Poaceae, which could also alter or destroy the reading frame, appear to be limited to the very 3' region of this gene not affecting the functionality of domain X (Hilu and Alice 1999). Thus, the ORF of *mat*K appears to be intact and maintained in these plant species (Ems et al. 1995; Hilu and Alice 1999). Further, the presence of the *matK* gene without *trnK* retained in the residual chloroplast genome of *Epifagus* (Ems et al. 1995) and the finding of a protein product for MatK in extracts from barley (Vogel, Borner, and Hess 1999) support that this gene is translated into an essential functional protein product in the chloroplast genome.

Project objectives

Why focus on *matK*? The importance of investigating this gene is three-fold: (1) to investigate the expression of the only putative chloroplast-encoded group II intron maturase, (2) to explore the molecular evolution of protein structure for a fast-evolving gene, and (3) to support its utility in plant systematics. The study presented here addressed these aspects of *matK* by utilizing molecular techniques to examine the RNA and protein product of *matK,* as well as a bioinformatics approach to analyze protein chemical and genetic structure.

Post-transcriptional mechanisms in the chloroplast are poorly understood at present. Examination of *matK* expression at both the RNA and protein level provides insight into one of the possible regulatory mechanisms of post-transcriptional processing in this organelle. In addition, since the genetic structure of *matK* as a potential maturase is different from that of other group II intron maturases (ie. *matK* lacks a complete RT domain and a DNA endonuclease domain) (Mohr, Perlman, and Lambowitz 1993), it is of great interest to elucidate if the remaining remnants of the RT domain and domain X are enough to maintain function of this

enzyme. Furthermore, the fact that MatK may splice group II introns other than the one in which it resides, a characteristic not found in the LtrA maturase of bacteria (Matsuura, Noah, and Lambowitz 2001), suggests that MatK functions through a novel splicing mechanism.

Although several crystal structures are available for a number of enzymes, no crystal structure exists for any group II intron maturase. Only recently a homology model has been generated for the *Lactococcus lactis* LtrA maturase based off the RT domain in this enzyme (Blocker et al. 2005). However, the lack of a complete RT domain in MatK precludes enough sequence similarity to deduce a homology model of this putative chloroplast maturase. Thus, the structure of MatK is unknown. Analyzing the chemical structure of MatK offers a new potential for not only elucidating features of structural and functional importance in this unusual group II intron maturase, but also for understanding protein evolution of a fast-evolving gene. MatK evolves at a fast pace, resulting in substitution rates suggestive of a pseudogene (Whitten, Williams, and Chase 2000). Identification of a MatK protein product suggests that the substitution observed at the nucleotide and amino acid level is not reflected as a shift in final chemical structure, providing a model for the buffering of fast-evolutionary change on protein structure and function.

Although a transcript for *matK* has been identified from a few species across land plants (Vogel et al. 1997; Kugita et al. 2003; Nakamura et al. 2003; Wolf, Rowe, and Hasebe 2004) and a protein product of the expected size observed from barley plant extracts (Vogel, Borner, and Hess 1999), this evidence alone has not sufficed to nullify the idea that *matK* may be a pseudogene in some plant species (Whitten, Williams, and Chase 2000; Kugita et al. 2003). If *matK* is a pseudogene, its utility in plant systematics would be greatly impacted. Pseudogenes tend to evolve at a faster rate than their functional homologs (Ophir et al. 1999) suggesting a loss

of functional constraint, resulting in long branch attraction when used in phylogenetic analysis (Felsenstein 1985) and unreliable evolutionary trees. As noted earlier, several phylogenies from below the family level (Johnson and Soltis 1995; Hilu, Alice, and Liang 1999; Whitten, Williams, and Chase 2000; Salazar et al. 2003) through all angiosperms (Hilu et al. 2003) have been produced using *matK* sequence data. This fact highlights the critical importance of elucidating whether *matK* is or is not a pseudogene. In order to resolve this issue, we examined the RNA and protein product expressed from this gene using both molecular and computational analysis. Therefore, we were able to address all aspects that define a pseudogene, (1) genes that are no longer transcribed, (2) genes that form truncated, non-functional protein, and (3) genes that lack evolutionary constraint.

Chapter 2

matK TRANSCRIPTION IN LAND PLANTS: IMPLI CATIONS OF EXPRESSION ON PUTATIVE MATURASE FUNCTION

ABSTRACT

The strong phylogenetic signal from *matK* has rendered it an invaluable molecule in plant systematic and evolutionary studies. Molecular information from this single gene has produced phylogenies as robust as those derived from several other genes combined. As yet, little is known concerning *matK* expression and function. The *matK* gene is located within the group II intron of *trnK* (tRNAlys, UUU) in most land plants, and has been proposed as the only chloroplast-encoded group II intron maturase. Potential substrates for this maturase include introns found within genes of the translation machinery of the chloroplast and at least one photosynthesis related gene. Using RT-PCR and Northern blot analyses, we examined transcription of *matK* across land plants to determine features of expression for this important gene. Our results revealed two predominant *matK* RNA transcripts in plants. These transcripts are greatly decreased in *Oryza sativa* (rice) plants grown in the dark as well as four weeks post-germination. Therefore, the level of these RNA transcripts is regulated by plant developmental stage and etiolation. Furthermore, we report the first evidence of a transcript for *matK* separate from the precursor for *trnK*. We also identify a protein product for MatK from plant extracts. This work provides insight into the transcription and protein expression of the only putative group II intron maturase encoded in the chloroplast and supports the utility of this gene in plant systematics.

Key Words: land plants; *matK*; maturase; orchids; transcript; chloroplast genome

INTRODUCTION

matK has recently emerged as an invaluable gene in plant systematics and evolution. Sequence information from *matK* has generated phylogenies as robust as those constructed from two to several other genes combined and using data sets that amounted up to 15,000 nucleotides (Hilu et al. 2003). The phylogenetic signal from *matK* has been used to resolve relationships from the species level to across broad groups of land plants (Johnson and Soltis 1994; Johnson and Soltis 1995; Hilu and Liang 1997; Hayashi and Kawano 2000; Hilu et al. 2003; Cameron 2005) and green algae (Sanders, Karol, and McCourt 2003). The mode and tempo of *matK* evolution is distinct from other chloroplast genes. The rate of nucleotide substitution in *matK* is three times higher than that of *rbcL* (Johnson and Soltis 1994; Olmstead and Palmer 1994), denoting it as a fast- or rapidly-evolving gene. This high rate of nucleotide substitution is accompanied by relatively high rates of amino acid substitution (Olmstead and Palmer 1994; Hilu and Liang 1997; Whitten, Williams, and Chase 2000), as well as by the presence of indels across its open reading frame (Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003). These features, along with the detection of frame shift mutations, have raised questions concerning functional constraint on this gene (Kores et al. 2000; Whitten, Williams, and Chase 2000; Hidalgo et al. 2004). Although direct proof of MatK function is lacking, evidence from DNA sequence and RNA secondary structure analyses demonstrated evolutionary constraint in *matK* (Young and dePamphilis 2000). In addition, du Jardin et al. (1994) and Vogel et al. (1999) have demonstrated that a protein product is expressed from this gene in plant extracts from potato and barley, respectively.

Sequence analysis of *matK* identified a conserved domain, domain X, with homology to mitochondrial group II intron maturases (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and

Link 1987; Mohr, Perlman, and Lambowitz 1993). Consequently, *matK* is the only putative group II intron maturase found in the chloroplast genome of land plants and some algae (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). The presence of 15 genes in the chloroplast genome of higher plants that contain a total of 16 group II introns (Kohchi et al. 1988; Ems et al. 1995; Maier et al. 1995) suggests an essential requirement for a group II intron maturase in this organelle. These group II introns are nested within genes, such as *trnK* and *rpl2* (Ems et al. 1995; Vogel, Borner, and Hess 1999), whose tRNA or protein products are required for normal chloroplast function. Studies by Vogel et al. (1997, 1999) and Hess et al. (1994) indicated that, although a nuclear maturase may process some chloroplast group II introns, there are at least six genes, *trnK*, *atpF, trnI*, *rps12*, *rpl2*, and *trnA,* which require a maturase expressed from the chloroplast for intron excision.

Studies at the RNA and protein level are needed to examine expression and function of *matK*. Previous studies have noted a *matK* transcript in five plant species (Kanno and Hirai 1993; Vogel et al. 1997; Kugita et al. 2003; Nakamura et al. 2003; Wolf, Rowe, and Hasebe 2004), but only one examined the details of this transcript (Vogel et al. 1997). In the present study, we increased the number of species across land plants in which a transcript for *matK* has been identified, determined the approximate size of *matK* transcripts from the monocot *Oryza sativa* L. (rice, Poaceae), and the eudicot *Solanum tuberosum* L. (potato, Solanaceae), and assessed *matK* RNA levels during plant development of these two species. To ensure results of *matK* transcription during plant development were not exclusive to one group of angiosperm, both a monocot and a eudicot model species were examined. In addition, change in RNA levels for *matK* and *trnK* was examined in response to etiolation treatment in *O. sativa*. Further, we demonstrated a distinct separation between regulation of RNA levels for the mature *trnK*

transcript and that of *matK*. This evidence supports that a separate transcript for *matK* from an unspliced *trnK* precursor exists and that this *matK* transcript could be translated into a protein product. To support this hypothesis, we have shown that MatK protein is present in *Oryza* crude protein extract. Investigating these aspects of *matK* is important for understanding the molecular expression of the only putative chloroplast-encoded maturase, assessing its potential function in chloroplast machinery, and further promoting its utility in plant systematics and molecular evolution.

MATERIALS AND METHODS

Plant material

Sixteen species representing the four major land plant lineages, bryophytes, monilophytes (ferns and fern allies, Pryer et al., 2004), gymnosperms, and angiosperms, were used (Table 2.1). Plants were grown from seed stock at the Virginia Tech Biology greenhouse, collected from the field, or obtained as potted plants. Both young and adult leaf material were collected for each plant, placed in Ziploc freezer bags, and stored at –80 ^{o}C. Herbarium vouchers are deposited in the herbarium at Virginia Tech, Blacksburg, VA, USA (VPI) and Dresden, Germany (DR) (Table 2.1).

RNA isolation

All solutions used for RNA isolation were treated with 0.1% DEPC followed by autoclaving to remove residual DEPC and RNases prior to use. Total RNA was isolated by grinding tissue under liquid nitrogen followed by phenol/chloroform/LiCl extraction and ethanol precipitation according to Altenbach and Howell (1981). Pelleted RNA was dissolved in water and

TABLE 2.1. Plant species used in this study along with their respective lineage, specimen origin and voucher information. Abbreviations in parentheses denote herbaria where voucher specimens are filed. Abbreviation for herbarium followed Holmgren et al., 1990.

Plant Species	Taxa Lineage	Location/Source Collection	Voucher
Bartramia pomiformis Hedw.	bryophytes	Pembroke, VA	D. Quandt (DR)
Atrichum undulatum Hedw.	bryophytes	Pembroke, VA	D. Quandt (DR)
Phaeoceros laevis L. (Prosk).	bryophytes	Mandanici I, Sicily	C. Neinhuis (DR)
Adiantum hispidulum Swartz	monilophytes	Biology greenhouse, Virginia Tech., Blacksburg, VA	M. Barthet (VPI)
Pinus strobes L.	gymnosperms	Blowing Rock, NC	M. Barthet (VPI)
Ginkgo biloba L.	gymnosperms	Virginia Tech, Blacksburg, VA	M. Barthet (VPI)
Ephedra viridis Coville	gymnosperms	Richters Herbs, Ontario, Canada	M. Barthet (VPI)
Ephedra nevadensis Wats.	gymnosperms	Richters Herbs, Ontario, Canada	M. Barthet (VPI)
Gnetum gnemon L.	gymnosperms	University of Connecticut	M. Barthet (VPI)
Nymphea oderata Ait.	angiosperms	Burleson County, TX	K. Neihaus (VPI)
Sagittaria latifolia Willd.	angiosperms	Claytor Lake, VA	M. Barthet (VPI)
Oryza sativa L.	angiosperms	Valley Seed Service, CA	M. Barthet (VPI)
Zea mays L.	angiosperms	Wetsel Seed Co., VA	K. H. Hilu (VPI)
Spathoglottis plicata Blume	angiosperms	Gardino Nursery Corp., FL	M. Barthet (VPI)
Solanum tuberosum L.	angiosperms	Blacksburg Feed and Seed, VA	M. Barthet (VPI)
Arabidopsis thaliana L. (Heynh.)	angiosperms	Eric Beers Virginia Tech., Blacksburg, VA	M. Barthet (VPI)

TABLE 2.2. Primers used for the amplification of *matK, trnK, rbcL*, or *matR* along with primer sequences, direction of amplification, minimum Tm for annealing, and taxa for which they were designed.

Primer	Sequence	[b]Primer direction	Tm (^{o}C)	Taxa	Gene	References
matKIF-BS	TTCGTCGACGCGTACAAGATGCTTC	F		bryophytes/ mosses	*matK*	this study
MatK1750R	GTTAAAGTGGTCCAAGCTAA	R	45	bryophytes/ mosses	*matK*	this study
1024matKFhorn	TTCGTAGACGAGTCTAAGATGCTTC	F		hornworts	*matK*	this study
1700matKRliver	TACAAAKTGGA6GTCCTAA	R	46	horworts	*matK*	this study
matK1024Fadian	TCCGACGACAAATTAAAGATGTTTC	F		*Adiantum*	*matK*	this study
matK1750Radian	GTAGAAGTAACCCAGGCCGA	R	48	*Adiantum*	*matK*	this study
GYmatKF	CTGGAAGTTCCGTTCT	F		gymnosperms	*matK*	this study
GYmatKR	CAACRATCGTGAATGAGA	R	48	gymnosperms	*matK*	this study
GnmatKF	GGATGCGTATCAAAAGTTC	F		Gnetophyta	*matK*	this study
GnmatKR	TCTGTAGTAACTCCACC	R	48	Gnetophyta	*matK*	this study
BSAGMATKF	AGGMTATTTAGAAATAGA	F		basal angiosperms	*matK*	this study
BSAGMATKR	CKCAATAAATGCAAAGA	R	45	basal angiosperms	*matK*	this study

AlismamatKF	CGAATGTATCAACAGAAT	F		Alismataceae	*matK*	this study
AlismamatKR	GAGGATTGTTTACGAAG	R	48	Alismataceae	*matK*	this study
[a]W	TACCCTATCCTATCCAT	F		Poaceae	*matK*	Hilu et al., 1999
[a]9R	TACGAGCTAAAGTTCTAGC	R	46	Poaceae	*matK*	Hilu et al., 1999
OcmatKF	CTCACTTGCTCATKATC	F		Orchidaceae	*matK*	this study
OcmatKR	CTTTGATCCAGCATTGA	R	45	Oricidaceae	*matK*	this study
[a]corematK1	ATGTATCAACAGAATCRT	F		core eudicots	*matK*	this study
[a]corematK2	ATGCAAAGAAGARGCATC	R	45	core eudicots	*matK*	this study
ArabmatKF	TTTGGGGCATACCAGTC	F		*Arabidopsis*	*matK*	this study
ArabmatKR	ATTCATCAGAAGCGGCG	R	46	*Arabidopsis*	*matK*	this study
[a]trnK3exF	CAAGCACGATTTGGGGA	F		*Oryza sativa*	*trnK*	this study
[a]trnK3exR	GTTTCCATATGGGTTGC	R	50	*Oryza sativa*	*trnK*	this study
[a]5extrnKF	CCTTTTGGTATCTGAGTG	F				
[a]trnK5exR	AGTACTCTACCATGAG	R	48	*Oryza sativa*	*trnK*	this study
3914	GGGGTTGCTAACTCAACGG	F	60	*Oryza sativa*	*trnK*	Johnson & Soltis, 1994
rbcLOsF	CCATATCGAGTAGACCCTGT	F		*Oryza sativa*	*rbcL*	this study
rbcLOsR	TCCACCGCGTAGACACTCAT	R	46	*Oryza sativa*	*rbcL*	this study
matRF	CATCGACCGACATCGATTC	F		*Oryza sativa*	*matR*	this study
matRrevdX	CCTGATAACTAGTATCGCC	R	46	*Oryza sativa*	*matR*	this study

[a]Primers used for probe synthesis.

[b] F= forward, R = reverse

immediately quantified using a Beckman DU 520 UV spectrophotometer (Beckman-Coulter, Fullerton, CA, USA) with fixed absorption spectra of 260 nm and 280 nm, and stored at –80 ^{o}C.

DNase treatment

Prior to RT-PCR, all RNA samples were amplified by PCR using taxon specific primer pairs (Table 2.2) with RNase included in the PCR reaction mix in order to check for DNA contamination. If DNA contamination was found, DNase treatment was performed prior to RT-PCR as follows: 1 µl of RNAse inhibitor (40 u/µl, Invitrogen, Carlsbad, CA, USA) was added along with 6.7 µl of 5X Reverse Transcriptase Buffer (Invitrogen, Carlsbad, CA, USA), 4.3 µl DNase (1 u/µl, Promega, Madison, WI, USA), and up to 10 µg of total RNA for a 20 µl reaction. Samples were then incubated at 37 ^{o}C for 30 minutes. One microliter of DNase stop solution (Promega, Madison, WI, USA) was then added, followed by incubation at 65 ^{o}C for 10 minutes to denature DNase. Ten microliters of this reaction was then used directly for Reverse Transcription (RT)-PCR.

Primer design

To avoid random amplification of other maturases containing the conserved domain X, primers were specifically designed to anneal to the 5' region of matK away from that domain (Table 2.2). The specificity of these primers was predicted by Blast search in GenBank. Some primers were species- or family-specific; however, degenerate primers were made where possible to span large groups of land plants. For example, corematK1 and corematK2 primers are able to amplify *matK* from most core eudicot families while still maintaining specificity to matK.

RT-PCR

Ten micrograms of DNA-free RNA was utilized in all RT-PCR reactions. The RT reaction was performed according to Shirley and Hwang (1995) with the following modifications: Superscript II (200 u/µl, Invitrogen, Carlsbad, CA, USA) was used instead of Moloney Murine Leukemia Virus reverse-transcriptase and an oligo $dT_{(15)}$ primer was used for the reverse primer. First strand cDNA was amplified using the following PCR conditions: 1) starting cycle of 95 ^{o}C for 3 min., 48 ^{o}C for 3 min., and 72 ^{o}C for 3 min., 2) main cycle program of 95 ^{o}C for 30 sec., 48 ^{o}C for 1.30 min., and 72 ^{o}C for 3 min., with main cycles repeated 50 times, and 3) 72 ^{o}C for 20 min. to complete end extension. The annealing temperature varied depending on the primers used (Table 2.2). Once cDNA was successfully amplified, a negative control containing RNA not reverse transcribed (no-RT) and RNAse was co-amplified with cDNA confirming that all PCR products from RT-PCR were the result of cDNA amplification not genomic DNA. All PCR products were separated on 1.5 % agarose gels.

3' RACE

Total RNA was isolated according to Altenbach and Howell (1981) from *O. sativa* tissue stored at –80 ^{o}C. As with RT-PCR, this RNA was then DNase treated as described above. 3' RACE was performed using the TaKaRa 3'-Full RACE Core Set (TAKARA BIO Inc, Otsu, Shiga, Japan) according to the manufactures' directions with the exception of using trnK3exR as the reverse primer instead of oligo dT in the RT reaction for mature *trnK* cDNA. The primer combinations W/trnK3exR, W/9R, and 3914/trnK3exR were used in a PCR reaction with cDNA generated from 3' RACE to amplify transcripts of *matK* with the 3' exon of *trnK*, *matK* alone,

and the mature transcript of *trnK*, respectively. A no-RT control was included in the synthesis of cDNA using 3' RACE.

Probe synthesis

matK probes for *O. sativa* and *S. tuberosum* Northern blot assays were constructed to include mostly the 5' N-terminus region and to avoid the conserved maturase domain X, which can be found in other group II intron maturases (Mohr, Perlman, and Lambowitz 1993). The rice *matK* probe consisted of 876 base pairs (bp) that included 250 bp of domain X, and was amplified using primers 9R and W (Table 2.2). The potato probe consisted of a 333 bp product of the *matK* N-terminal region and was amplified using the primers corematK1 and corematK2 (Table 2.2). Two *trnK* probes for rice were designed to exclusively hybridize to the 5' or 3' exon of *trnK*. The primers 5extrnKF and trnK5exR (Table 2.2) were used to amplify a sequence segment from 132 base pairs upstream to the end of the 5' exon of *trnK* to generate the 5' *trnK* exon probe. The 3' exon probe was generated using primers trnK3exF and trnK3exR (Table 2.2). These primers amplified a sequence section from 71 bp upstream of the *trnK* 3' exon to the end of that exon.

Probes were generated from *O. sativa* or *S. tuberosum* genomic DNA isolated using CTAB/chloroform/isoamyl alcohol extraction and isopropanol precipitation (Doyle and Doyle 1990). The annealing temperature (Tm) for probe amplification was 48 ^{o}C for the *O. sativa matK* probe, 45 ^{o}C for the *S. tuberosum matK* probe, and 48 ^{o}C and 50 ^{o}C for the *O. sativa trnK* 5' and 3' exon probes, respectively. Digoxygenin (Dig)-labeled probe synthesis was performed according to supplier instructions (Roche, Indianapolis, IN, USA).

Northern blot

Total RNA was separated on a 1% formaldehyde gel (Gerard and Miller 1986). A RNA size standard (Promega, Madison, WI, USA) was included on the gel for size estimation of transcripts. Prior to use, the gel box, support, and combs for the formaldehyde gel were treated with 3% SDS to remove contaminating nucleases. The RNA was then transferred to a nylon membrane and hybridized with the Dig-labeled *matK* probes described above at both low (50 ^{o}C) and high stringency (65 ^{o}C). Northern blotted RNA was also hybridized with the Dig-labeled *trnK* 5' and 3' exon probes at high stringency. Since RNA editing of *matK* has been observed previously (Vogel et al. 1997; Wolf, Rowe, and Hasebe 2004), a low stringency hybridization was performed to ensure binding of the DNA probes to possible RNA edited transcripts of *matK*, which may vary slightly in sequence. All Northern blot transfers were performed according to Church and Gilbert (1984) with the exclusion of the denaturing and neutralizing steps, which result in RNA degradation. A chemiluminescent signal was detected on the blots using anti-Dig antibody conjugated with alkaline phosphatase (Roche, Indianapolis, IN, USA) and the chemiluminescent substrate CDP-Star (Roche, Indianapolis, IN, USA) according to the manufactures' instructions.

Etiolation

Oryza sativa seed stock stored at 4 ^{o}C was planted in vermiculite at the Virginia Tech Biology greenhouse and grown under uniform light and water conditions or at 24 ^{o}C in the dark for two weeks. Plants grown in the dark were uniformly watered every four days in the dark. After two weeks, tissue from rice plants grown in the dark was harvested in a dark room. Following collection of tissue in the dark, remaining dark-grown *O. sativa* plants were placed in a Percival

growth chamber and exposed to light for up to 24 hours. Tissue was collected from light exposed plants at 4 and 24 hours after light exposure. Tissue was also harvested from *O. sativa* plants grown in the greenhouse under uniform light and water conditions as a control. All tissue was placed in Ziploc freezer bags, frozen immediately in liquid nitrogen, and stored at –80 ^{o}C. RNA was isolated by grinding under liquid nitrogen followed by phenol/chloroform extraction and ethanol precipitation (Altenbach and Howell 1981) and separated on 1% formaldehyde gels (Gerard and Miller 1986). Following gel electrophoresis, RNA was transferred to uncharged nylon (Church and Gilbert 1984), hybridized with *matK* and *trnK*-Dig-labeled probes, and detected by chemiluminescence using CDP-Star (Roche, Indianapolis, IN, USA) following the manufacturers' instructions.

Time-point analysis

Oryza sativa and *S. tuberosum* plants were grown from seed stock stored at 4 ^{o}C and from tubers, respectively, in the Virginia Tech Biology greenhouse under uniform light and water conditions. Plant leaves were collected at four time-points (2, 4, 6, and 8-weeks post-germination), placed in Ziploc freezer bags, and stored at –80 ^{o}C. RNA was isolated according to Altenbach and Howell (1981) from the leaf tissue and resolved on 1% formaldehyde gels (Gerard and Miller 1986). The RNA was then transferred to a nylon membrane as described above, hybridized with the *matK*-Dig-labeled probes, and detected by chemiluminescence using CDP-Star (Roche, Indianapolis, IN, USA) according to the manufacturers' instructions.

Probe specificity

Total plant RNA that may contain domain X maturases other than *matK* was used in the Northern blot analyses. The 876 bp and 333 bp Dig-labeled *matK* PCR probes described above from *O. sativa* and *S. tuberosum* were tested for target specificity against Southern blot membranes cross-linked by ultraviolet light with PCR product for *matK*, *rbcL*, and *mat-r*. The *mat-r* gene is a mitochondrial group II intron maturase that also contains domain X (Farré and Araya 1999).

Southern blotting

The two chloroplast and one mitochondrial genes *matK, rbcL*, and *matR*, respectively, were amplified from *O. sativa* genomic DNA isolated according to Doyle and Doyle (1990) using primer combinations W/9R (*matK*), rbcLOsF/rbcLOsR (*rbc*L), and matRF/matRrevdX (*mat-r*) (Table 2.2). Amplification of *mat-r* included 186 bp of domain X. Amplified products were separated on a 1.5% agarose gel and transferred to nylon membrane using the Southern blot method (Church and Gilbert 1984). The *O. sativa* and *S. tuberosum matK* probes were hybridized to nylon membrane (Church and Gilbert 1984) at 50 and 65 ^{o}C followed by Dig-detection and developing using CDP-Star (Roche, Indianapolis, IN, USA) as the chemiluminescent substrate.

Sequencing

Amplified products from RT-PCR, 3' RACE, *rbcL*, *mat-r*, and the probes for *matK* and the 5' and 3' exons of *trnK*, were gel-extracted using the Qiagen Qiaquick gel extraction kit (Qiagen, Valancia, CA, USA) and sequenced to confirm their gene identity. Cycle sequencing was

performed using 2.5 μl Applied Biosystems (ABI) BigDye Terminator sequencing dye (ABI, Framingham, MA, USA) and a final concentration of 1.25 μM of primer in a 15 μl reaction. The amount of DNA template used for sequencing ranged from 2.5 to 25 ng of Qiagen gel extracted PCR product. Cycle sequencing was performed following the ABI cycle sequencing protocol (Framingham, MA, USA) with the exception of varying the annealing temperature depending on the primer (Table 2.2). Products of the sequencing reaction were separated through capillary gel electrophoresis at the Virginia Bioinformatics Core Laboratory Facility (Blacksburg, VA, USA).

Protein isolation and Western blot analysis

Crude protein was isolated from two-week old *O. sativa* plants grown from seed stock at the Virginia Tech Biology greenhouse grown under uniform light and water conditions. Tissue was ground under liquid nitrogen, boiled in 500 μl 1 X Laemmli SDS sample buffer (62.5 mM Tris, pH 6.8, 2% SDS, 10% glycerol, and 5% β-mercaptoethanol) for 15 minutes, then centrifuged at 15, 000 x g two times at room temperature. The supernatant was collected and placed in a new tube after each centrifugation. Protein extract was then stored at –20 ^{o}C. Concentration of protein was determined by the BioRad Protein Microassay (Bio-Rad, Hercules, CA, USA) using BSA as the standard. Seventy five micrograms of crude protein extract was separated by SDS-PAGE (7.5% acrylamide) along with a protein size standard (New England Biolabs, NEB, Ipswitch, MA) then transferred to 0.22 μm nitrocellulose. Protein transfer was checked by Ponceau S staining. Membranes were blocked for 1 hour in 5% Carnation non-fat dry milk diluted in phosphate buffered saline plus Tween 20 (PBS-T) at room temperature followed by incubation with anti-MatK (1:50) diluted in PBS-T overnight at 4 ^{o}C on a nutating mixer. The anti-MatK antibody was produced in a rabbit against a 15 amino acid peptide sequence of *O.*

sativa MatK (Barthet and Hilu, unpublished data). Following overnight incubation with the primary antibody, membranes were rinsed, and then washed with PBS-T for 1X 15minutes, and 2X 5 minutes, followed by incubation with horse-radish peroxidase (HRP)-conjugated anti-rabbit (1:2000, Cell Signaling Technology, Danvers, MA, USA) for 1 hour at room temperature. Membranes were washed with PBS-T in the same manner as previously noted and chemiluminescent signal detected using ECL peroxidase/luminol system (Amersham Biosciences, Piscataway, NJ, USA) or West Pico chemiluminescent detection system (Pierce Biotechnology, Inc., Rockford, IL, USA).

RESULTS

Transcript size of matK and trnK

Hybridization of the *matK* probes to total RNA from *O. sativa* and *S. tuberosum* identified two predominant transcripts of 2.6 kilobases (kb) and 2.9 kb (Figure 2.1A). Three smaller, less prominent, transcripts of approximately 1.8, 1.5, and 0.9 kb were also observed from the Northern blot for both rice and potato (Figure 2.1A). These experiments were reproduced in two biological replicates and several technical replicates, each time with the same results.

Southern blot hybridization of the *matK* probes at high stringency (65 ^{o}C) to amplified *matK*, *rbcL*, and *mat-r* from *O. sativa* genomic DNA resulted in a strong band exclusively for *matK* with no cross-reactivity to *rbcL* or *mat-r* (Figure 2.1B). Only very weak cross-reactivity to *mat-r* was observed when the same hybridization was repeated at a low stringency temperature of 50 ^{o}C (data not shown). The Southern blot experiments were repeated twice with newly amplified products each time.

The 5' *trnK* probe, which was designed to bind exclusively to this exon of *trnK,* hybridized to transcripts of 2.9 kb and 2.6 kb, and four smaller transcripts of 1.8 kb, 1.5 kb, 0.9 kb, and <0.5 kb (Figure 2.2B). The 3' *trnK* probe, a probe designed to hybridize only to the 3' exon of *trnK* (Figure 2.2A), bound to transcripts of approximately 5.7 kb, 2.9 kb, and 2.6 kb, as well as three smaller transcripts of 1.8 kb, 1.5 kb, and 0.9 kb (Figure 2.2B). The Northern blot experiments using the *trnK* exon probes were repeated in several biological and technical replicates, each time producing similar results.

RT-PCR using 3' RACE amplified a 876 bp product for *matK,* a 1377 bp product for *matK*-3' exon of *trnK,* and a 120 bp product for mature *trnK* (Figure 2.3). The 3' RACE products were sequenced and confirmed as *matK*, *matK* with the 3' exon of *trnK*, and mature *trnK* transcript, respectively. A control sample lacking reverse transcriptase for the 3'RACE reaction using primers 9R and W did not result in detectable product. The 3' RACE experiment was repeated once and the same results obtained.

In summary, an independent transcript for *matK* from the *trnK* precursor could not be identified based on size difference alone. The Southern blot results demonstrate the specificity of the *matK* probe to only this gene and its transcribed RNA. The Northern blot data is further supported by 3' RACE results which confirm that a transcript containing both *matK* and the 3' exon of *trnK* does exist as well as a mature transcript of *trnK*.

matK/trnK RNA levels with etiolation

The mature transcript for *trnK* was found in equal levels in *O. sativa* plants grown in light and dark (Figure 2.4A and B). However, the 2.6 and 2.9 kb transcripts found with both the 5' *trnK* probe and *matK* probe appeared at different intensities in light and dark treatments (Figure

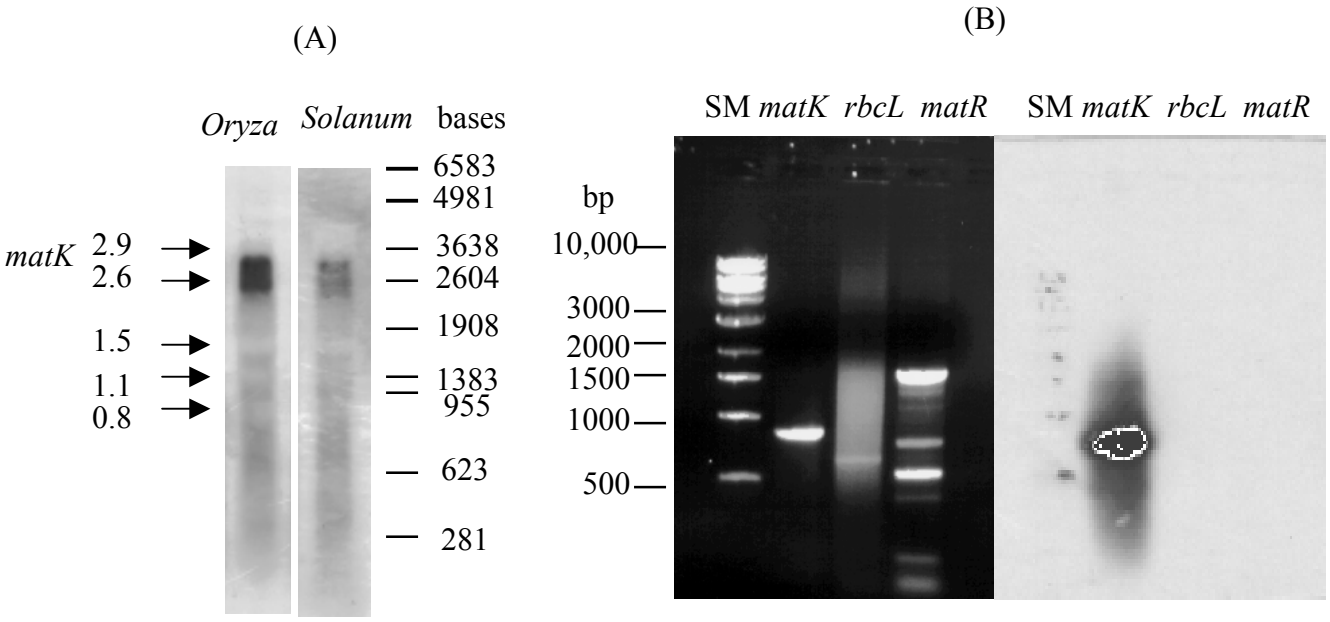


Figure 2.1. *matK* transcript size and probe specificity

(A) Size of *matK* transcripts from rice and potato. 20 μg of total RNA isolated from rice (left) or potato (right) tissue was loaded in each lane of a 1% formaldehyde gel. RNA was subsequently blotted to nylon membrane and probed with a *matK* specific probe. RNA size markers (Promega) are noted on the right. Arrows indicate *matK* transcripts. (B) Test of the specificity of the *matK* probe to *rbcL* and *matR*. PCR amplified products for two chloroplast genes, *matK* and *rbc*L, and one mitochondrial maturase gene, *mat*R, from rice were resolved on (right) 1.5 % agarose gel followed by (left) Southern blotting to a nylon membrane and probing with a *matK* genomic DNA probe. Lanes from left to right represent size marker (1 kb ladder from Promega); *matK* amplified product; *rbc*L PCR product; and *mat*R PCR product.

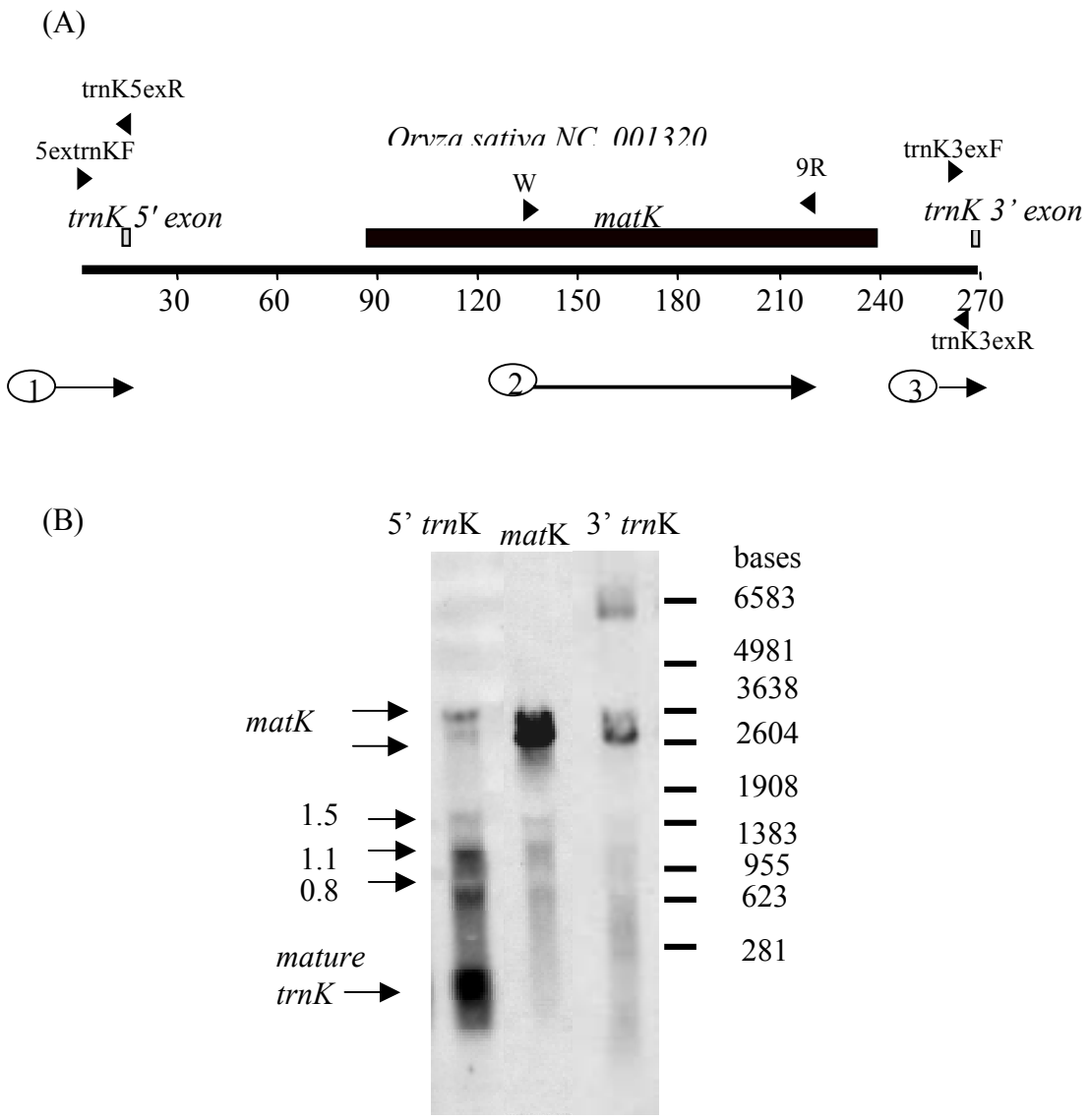


Figure 2.2. Comparison of *trnK* and *matK* transcript size. (A) Annotation of the *trnK/matK* DNA region with primers used to construct the *trnK* 5' and 3' exon probes noted with ▶. Probes are shown as a numerical value with an arrow. (B) Northern blot comparing the size of *matK* transcripts to transcripts containing the *trnK* 5' or 3' exon. 20 μg of total RNA isolated from rice was loaded in duplicated lanes on a 1% formaldehyde gel. RNA was subsequently blotted to nylon membrane and probed with the (1) *trnK* 5' exon probe, (2) *matK* probe, or the (3) *trnK* 3' exon probe. Transcripts for the *matK/trnK* gene unit are noted to the left. RNA size markers (Promega) are noted on the right.

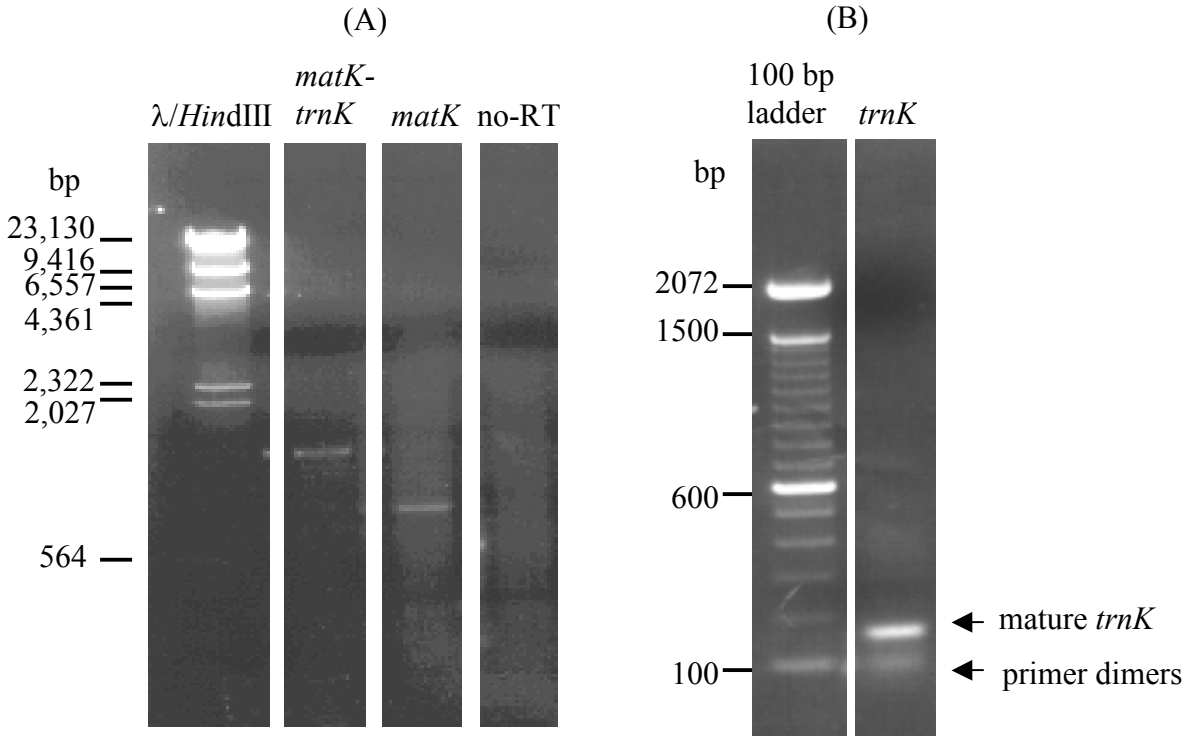


Figure 2.3. PCR products from 3'RACE performed on *O. sativa* RNA using primers for *matK* and *trnK*. (A) 1% agarose gel: lanes from left to right represent *Lambda*/*Hin*dIII size markers; RT-PCR product from using primers W and trnK3exR; RT-PCR product of *matK* only from using primers W and 9R; and no-RT control. (B) 1.7 % agarose gel: lanes from left to right represent 100 bp ladder (Gibco); and RT-PCR products using primers 3914 and trnK3exR. The mature transcript for *trnK* and primer dimmers are indicated with arrows to the right.

2.4A). The 2.6 and 2.9 kb transcripts were almost non-apparent in dark grown plants. Subsequent exposure of dark grown plants to light increased RNA levels of the 2.6 and 2.9 kb transcripts. However, even after 24 hours of light exposure, RNA levels did not accumulate as high as observed with greenhouse controls (Figure 2.4B). Biological experiments were repeated once and Northern blots repeated one to two times within each biological replicate with similar results.

Time-point analysis

Oryza sativa and *S. tuberosum* tissue was used to ascertain relative levels of *matK* RNA at four time points. The predominant *matK* transcripts of 2.9 kb and 2.6 kb were identified at two, four, six, and eight weeks post-germination for both species. In *O. sativa*, the highest relative level of RNA was observed at the eight-week time point, and a decrease in RNA was apparent at the four-week time point (Figure 2.4C). For *O. sativa*, the two- and six-week time points had similar level of *matK* RNA (Figure 2.4C). RNA levels of *matK* in potato decreased at the four-week time point. However, the highest level of *matK* RNA in *S. tuberosum* was detected at six weeks post-germination instead of eight weeks as in *O. sativa* (data not shown). Biological replicates were performed once with *O. sativa*. Within each of these biological experiments, Northern blot assays were repeated one to two times. Northern blot assays with potato RNA were repeated twice. Although loading controls are subject to interpretation, in all replicates, biological and technical, similar results were obtained. Changing stringency conditions did not alter results. Thus, although it appears that *matK* RNA is present at all time-points examined, levels of this RNA vary during developmental stages in a species-dependent manner.

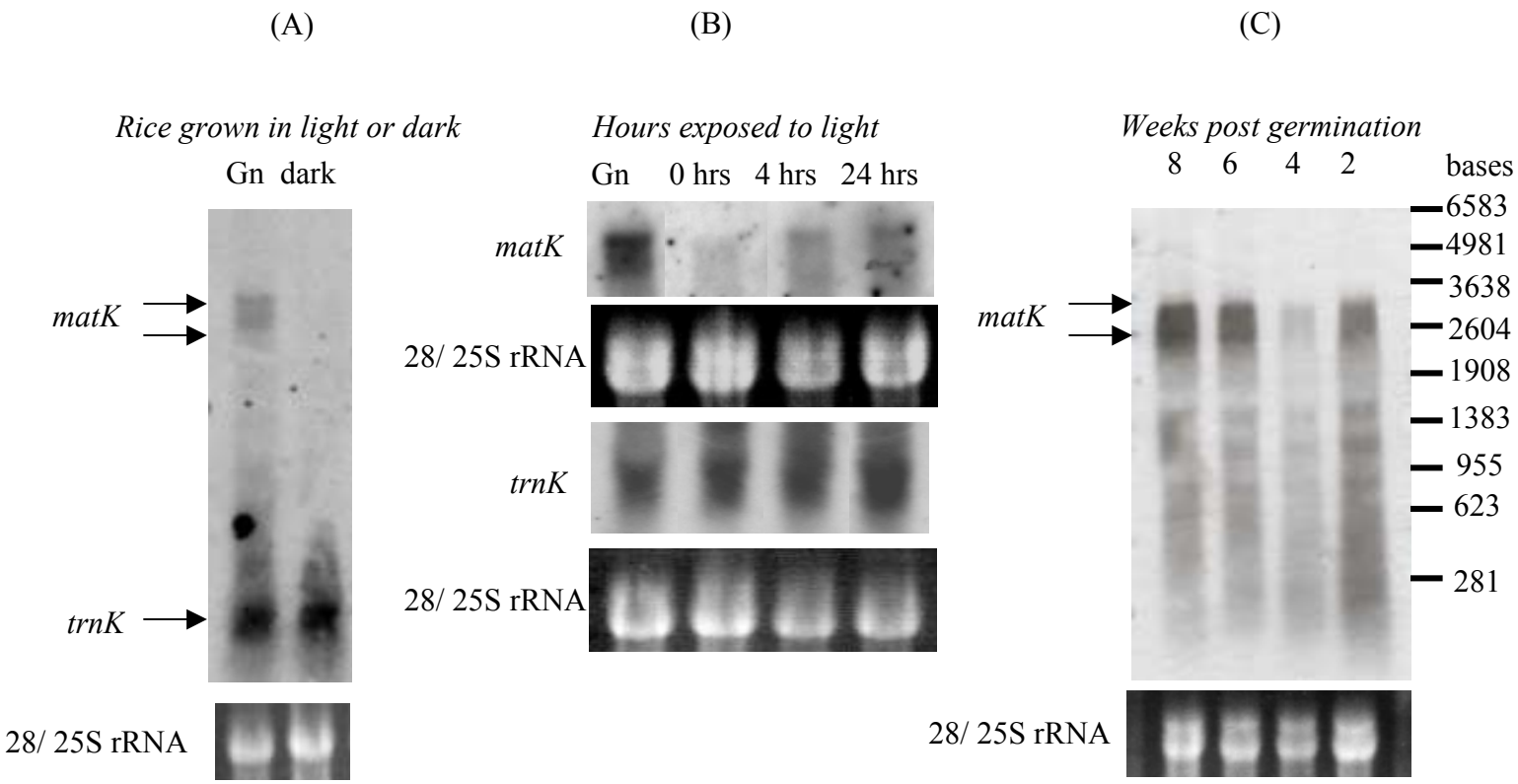


Figure 2.4. The influence of light and development on RNA levels of *matK* and the mature transcript of *trnK* in *O. sativa*. *matK* and the mature transcript of *trnK* from (A) light and dark treatment and (B) extension of etiolation with light exposure for 4 and 24 hours. (C) *matK* RNA levels at 2-4 weeks post-germination in *Oryza sativa*. 20 µg of total RNA isolated from tissue collected at each time point was loaded in each lane of a 1% formaldehyde gel. RNA was subsequently blotted to nylon membrane and probed with a *matK* and *trnK* specific probe. Prior to blotting, the gel was stained with ethidium bromide and the quantity of 28/ 25S rRNA examined to assess uniformity in loading of each well. Arrows indicate predominant transcripts for *matK* and the mature transcript for *trnK*. The 28/25S *rRNA* loading control is shown below in blot picture. RNA size markers (Promega) are noted on the right of (C). Gn = greenhouse control; dark = rice plants grown in the dark.

TABLE 2.3. Species for which an RT-PCR product for *matK* transcription was observed, their respective phylogenetic position, and size of RT-PCR product.

Major taxonomic group	Lineage	Family	Species	Size of product (in bp)
bryophytes		Bartramiaceae	*Bartramia pomiformis*	690
		Polytrichaceae	*Atrichum undulatum*	690
		Anthocerotaceae	*Phaeoceros laevis*	665
monilophytes		Adiantaceae	*Adiantum hispidulum*	450
angiosperms	basal ([1]ANITA)	Nymphaeaceae	*Nymphea oderata*	522
	monocots	Alismataceae	*Sagitaria latifolia*	503
		Poaceae	*Oryza sativa*	876
		Poaceae	*Zea mays*	876
		Orchidaceae	*Spathoglottis plicata*	449
	core eudicots	Solanaceae	*Solanum tuberosum*	333
		Brassicaceae	*Arabidopsis thaliana*	621

[1] Phylogenetic group consisting of *Amborellaceae*, Nymphaeaceae, Illiciaceae, Trimeniaceae, and *Austrobaileyales* (Hilu et al, 2003).

matK transcription across land plants

RT-PCR on RNA produced product from eleven plant species that spanned three major plant lineages, bryophytes, monoliphytes, and angiosperms (Table 2.3). Only a single predominant PCR band was produced in each RT-PCR reaction. The PCR products ranged from 333 bp to 876 bp depending on the particular primer pair utilized during the reaction (Table 2.2). RT-PCR from two representative species of mosses *Bartramia pomiformis* Hedw. (Figure 2.5) and *Atrichum undulatum* Hedw (data not shown), resulted in a single PCR band of 690 bp for each species. These species represent two distant bryophyte families, Bartramiaceae and Polytrichaceae, respectively. RT-PCR of the monilophyte *Adiantum hispidulum* Swartz produced a PCR product of 450 bp (Figure 2.5). RT-PCR of RNA produced *matK* PCR product (Table 2.2) from species representing major angiosperm lineages from early diverging Nympheaceae to core eudicot lineages (Qiu et al. 1999; Hilu et al. 2003). The size of the PCR product for flowering plants ranged from 333 bp to 876 bp (Table 2.3, Figure 2.5) depending on the primer pair used (Table 2.2). Therefore, the *matK* gene is transcribed in a variety of land plants from the early diverging mosses, to ferns, to the late diverging lineages of angiosperms.

All RT-PCR products resulting from this study were sequenced and confirmed as *matK* except that from *A. hispidulum*, where low amplification prevented adequate sequencing. Negative controls, RNA that lacked reverse transcription (no-RT controls), for all RT-PCR reactions lacked PCR product, confirming that the RNA used for RT-PCR was devoid of genomic DNA contamination and that RT-PCR products were the result of cDNA amplification (Figure 2.5).

In spite of several attempts with different protocols, an adequate amount of DNA-free RNA for RT-PCR could not be isolated from any of our gymnosperm samples. Difficulty of

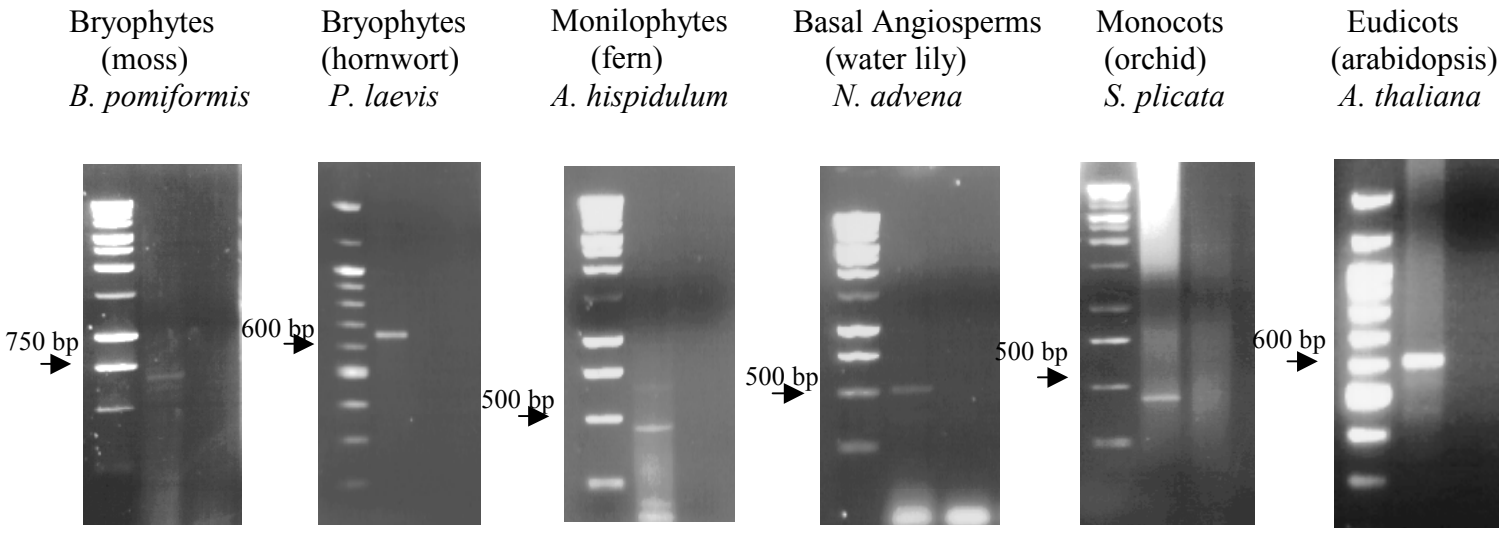


Figure 2.5. RT-PCR products of *matK* cDNA from representative species of land plants. Lanes from left to right represent: DNA size marker (a 250 bp or 100 bp ladder); RT-PCR product; and no-RT negative control. Key ladder sizes for PCR products are indicated; PCR product size is indicated below the band.

RNA isolation may be due to a high concentration of polyphenolics in the gymnosperms species sampled in this study. It is known that the high concentration of polyphenolics in gymnosperms may cause complications with RNA extraction in these plants (Kiefer, Heller, and Ernst 2000).

MatK Protein

An immunoreactive protein of ~55 kDa was detected using anti-MatK antibody against *O. sativa* protein extract on Western blots (Figure 2.6). A band of this size was not detected when using pre-immune serum against the same extract (data not shown). In addition, three less intense bands of much lower molecular mass, ~40, ~30, and ~25 kDa, were also detected with the anti-MatK antibody (Figure 2.6). Protein bands of ~40 and 25 kDa were detected on immunoblots incubated with pre-immune serum (data not shown). The same results were obtained using two biological replicates and two to three immunoblot assays for each replicate.

DISCUSSION

Transcript size of matK and trnK 5' and 3' exon analysis

Two prominent transcripts of approximately 2.9 and 2.6 kb were observed for *matK* in both *O. sativa* and *S. tuberosum* (Figure 2.1A). Similar Northern blot results were obtained by Vogel et al. (1997) in *Hordeum vulgare* L. (barley, Poaceae) using a probe specific to the 5' exon of *trnK*. However, a transcription map of *O. sativa* constructed by Kanno and Hirai (1993) noted only a single transcript for the *trnK* gene region. The *matK* transcripts from *O. sativa* identified in this study occasionally were observed to have a dark region between the two bands, making it difficult to distinguish between the two transcripts. If the RNA was not well separated on the

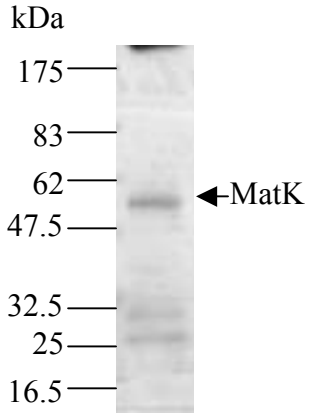


Figure 2.6. Western blot detection of MatK protein in *O. sativa* extracts. Seventy five micrograms of crude protein extract from rice was separated by 7.5% SDS-PAGE, transferred to 0.22 um nitrocellulose, then blocked in 5% Carnation non-fat dry milk. Following blocking, the membrane was incubated with a 1:50 dilution of anti-MatK in PBS-T overnight. Secondary incubation used a 1:2000 dilution of HRP-conjugated anti-rabbit in 5% Carnation non-fat dry milk. Protein was detected by chemiluminescence. The full-length MatK protein is indicated by an arrow. Protein standards (New England Biolabs) are noted on the left.

formaldehyde gel, it is possible that the close proximity of the two bands would render them appear as one large transcript band. This could have been the reason for Kanno and Hirai (1993) to describe only a single RNA species for this gene region.

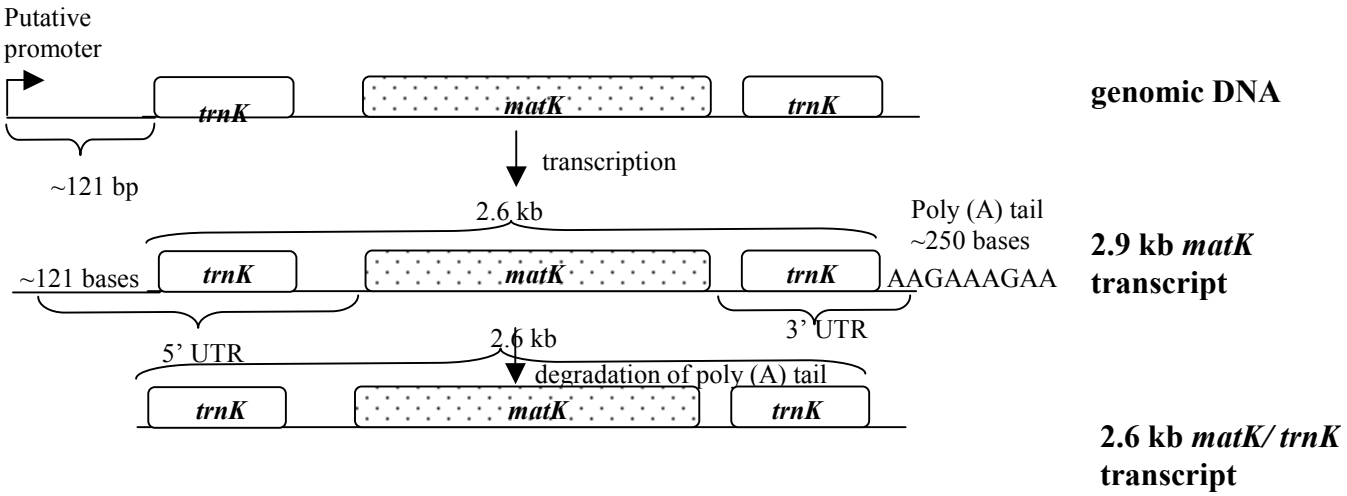
Identification of more than one transcript for *matK* could possibly be attributed to cross hybridization of our probe to another maturase. We tested this possibility by hybridizing the *matK* probes at both high and low stringency to a Southern blot containing amplified products of *matK*, *rbcL*, and *matR*. The latter gene is a mitochondrial group II intron maturase that also contains domain X (Farré and Araya 1999). The presence of only one strong hybridization signal on the membrane for PCR amplified *matK* DNA and the lack of a signal for the other two genes verified the specificity of the *matK* probe to *matK* transcripts only (Figure 2.1B). This result supports our findings that *matK/trnK* transcription produces two predominant RNA bands.

Polyadenylated chloroplast transcripts have been observed on blots as doublet bands with an intervening shaded area between them (Walter, Kilian, and Kudla 2002). Thus, it is very likely that the *matK* transcripts we observed indicate stages of polyadenylation processing. It is known that some mRNAs from the chloroplast undergo several steps of maturation, such as polyadenylation (Kudla, Hayes, and Gruissem 1996; Lisitsky, Klaff, and Schuster 1996; Schuster, Lisitsky, and Klaff 1999; Komine et al. 2000; Komine et al. 2002), cis- and trans-splicing (Kohchi et al. 1988; Ems et al. 1995; Rivier, Goldschmidt-Clermont, and Rochaix 2001; Vogel and Börner 2002), and processing of 5' and 3' ends (Stern and Gruissem 1989; Schuster and Gruissem 1991; Vogel and Hess 2001). It is possible that the *matK* transcript undergoes some or all of these processing events. Further experimentation is necessary to determine if a polyadenylated (poly (A)) tail is attached to the 2.9 kb or 2.6 kb *matK* transcripts.

Two different promoters have been suggested for the *trnK/matK* gene region. Vogel et al. (1997) suggested a promoter located 350 bp upstream of the *trnK* 5' exon, whereas Neuhaus and Link (1987) identified a promoter region only 121 bp upstream of this exon. Since probes for the 5' and 3' exons of *trnK* hybridized to two transcripts of the same size recognized by the *matK* probe (Figure 2.2B), it is possible that *matK* is transcribed as part of a *trnK* unspliced precursor transcript or as a gene unit that includes the *trnK* exons as 5' and 3' UTR segments. We hypothesize that *trnK* and *matK* are transcribed independent of each other, possibly by two separate promoters. The length of the *trnK*/*matK* gene region in *O. sativa* (GenBank accession: NC_001320) is ~2.6 kb (Figure 2.7). It is likely that a transcript meant to produce MatK protein is transcribed using the 121 bp upstream promoter described by Neuhaus and Link (1987). This transcript would still include the *trnK* exons but as part of the 5' and 3' UTR of the *matK* transcript (Figure 2.7A). The addition of the short 121 bp region between this proposed promoter and the 5' exon of *trnK* to the 2.6 kb *trnK/matK* region would result in a transcript of approximately 2.7 kb and would correspond well with the ~2.6 kb transcript our observed from our Northern blot data. The addition of a poly (A) to this 2.6 kb transcript would result in a second larger transcript. Thus, we hypothesize that the larger transcript observed for the *matK* gene of 2.9 kb is the precursor transcript that includes a poly (A) tail. This tail is then degraded, resulting in the mature *matK* 2.6 kb transcript (Figure 2.7A).

Vogel et al. (1997) suggested from primer extension analysis that the promoter for *matK* was located around 350 bp upstream of the first *trn*K exon. Boyer and Mullet (1988) noted that this upstream region contains a putative –35 promoter site for transcription initiation. If the promoter described by Boyer and Mullett (1988) was used to transcribe *trnK*, this would produce a 2.9 kb transcript, as we observed from the Northern blots. If the *trnK* unspliced precursor then

(A)



(B)

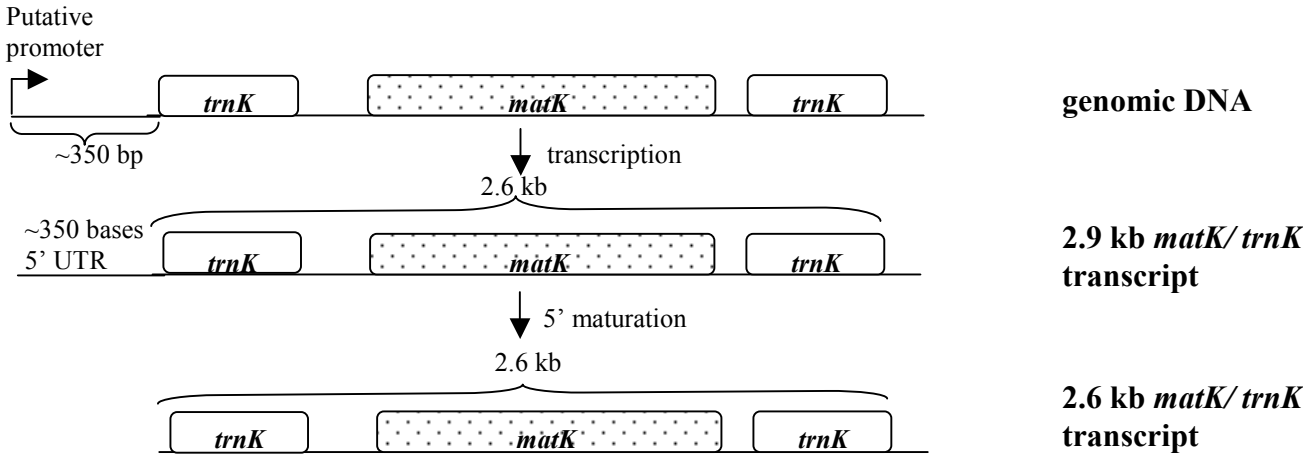


Figure 2.7. Model of transcription for *matK* and *trnK* genes. (A) Scenario for transcription for *matK* transcription that would results in a 2.9 kb polyadenylated transcript and a 2.6 kb mature transcript. (B) Model of *trnK* transcription using the promoter described by Boyer and Mullet (1988) resulting in the 2.9 kb and 2.6 kb *matK/trnK* transcripts. Site of putative promoter is noted along with sizes of the 5' and 3' UTRs and poly (A) tail. White boxes indicate 5' and 3' *trnK* exons while spotted box represents the *matK* open reading frame.

underwent 5' maturation, this would produce the smaller 2.6 kb transcript (Figure 2.7B). This scenario for *matK/trnK* transcription does not include a large poly (A) tail and would terminate shortly after the 3' exon of *trnK*. A 2.8 kb *trnK* transcript has been observed previously in total RNA from *Sinapsis alba* L. (Brassicaceae) (Neuhaus and Link 1987; Nickelsen and Link 1991). This transcript has been shown to terminate shortly after the 3' exon (Neuhaus and Link 1987; Nickelsen and Link 1991). Our proposed model of *trnK* transcription using the promoter suggested by Vogel et al. (1997) and Boyer and Mullett (1988) corresponds to the termination of the transcript described for *trnK* RNA from *S. alba*. These proposed *trnK* precursor transcripts would still include the group II intron and, thus, *matK* (Figure 2.7B). Therefore, the unspliced *trnK* precursor transcripts would be identified using the *matK* probe generated in this study. A RT-PCR product was found using 3' RACE on rice RNA that included both the *matK* coding region (amplified using primer W) and the 3' exon of *trnK* (amplified using trnK3exR) (Figure 2.3), confirming our Northern blot data. However, since both promoter scenarios for *trnK* and *matK* would result in transcripts of similar sizes and containing the *trnK* exons as well as *matK*, it is impossible to distinguish between this transcript possibilities using Northern blot data. Based on the size of the transcripts identified for *matK* and the characteristics of the transcript bands, this gene most likely uses the promoter identified by Neuhaus and Link (1987) and includes a poly (A) tail, while *trnK* transcript requires the promoter described by Vogel et al. (1997). The use of two separate promoters would enhance control of gene expression for each of these two genes, *trnK* and *matK*, in the chloroplast.

The 3' exon of *trnK* also hybridized to a transcript of 5.7 kb which was not observed using either the 5' exon or *matK* probes. RNAse protection assays identified a transcript that

extended beyond the *trnK* 3' exon and into the *psbA* gene region (Nickelsen and Link 1991). This longer transcript includes at least part of the *psbA* region and is considered a read-through transcript for the *psbA* gene (Nickelsen and Link 1991). The *psbA* gene is located 220 bp downstream of *trnK* and is transcribed in the same direction as *trnK* (GenBank accession: NC_001320). The larger *trnK* band of 5.7 kb that we observed by Northern blot analysis is most likely this *trnK/psbA* read-through transcription product and includes part, if not all, of *psbA*.

A mature *trnK* transcript, expected to be around 60 bases in length, was identified by Northern blot using the probe for the *trnK* 5'exon (Figure 2.2B). RT-PCR from 3' RACE experiments using the same total RNA from rice used in the Northern blot assays produced an approximately 120 bp band that was confirmed by sequencing as the mature transcript of *trnK* (Figure 2.3). Thus, although a spliced product for the group II intron of *trnK* that contains *matK* only was never observed on Northern blots, this intron is spliced to produce a mature *trnK* transcript.

In addition to the predominant transcripts for *trnK/matK* (2.9 kb and 2.6 kb), three much weaker bands of 1.8, 1.5, and 0.9 kb with homology to the *matK* and *trnK* gene probes were observed by Northern blot (Figure 2.1A and Figure 2.2B). Since these RNA bands were found using the *matK* and *trnK* 5' and 3' exon probes, it is likely that they represent splicing intermediates of the *trnK*/*matK* transcript.

In conclusion, the 2.6 and 2.9 kb transcripts observed using the various probes for *trnK* and *matK* could be either *matK* transcripts with a 5' UTR and 3' UTR that included the *trnK* exons or unspliced *trnK* precursor transcripts. We found no evidence for a separate transcript for *matK* using these probes based on a size difference. However, as stated above, it is still possible that a separate transcript for *matK* does exist but cannot be differentiated from a *trnK* precursor

based on size alone. Characterization of RNA levels for *matK* and the mature transcript of *trnK* in response to light has distinguished these two transcripts. These results are discussed below.

Etiolation of matK/trnK transcripts

A separate transcript for *matK* from the potential *trnK* precursor transcripts of 2.6 and 2.9 kb could not be found based on size using probes for the 5' *trnK* exon, 3' *trnK* exon, and *matK*. However, a distinct difference in the level of RNA for the 2.6 and 2.9 kb transcripts from that of the mature *trnK* transcript was found when rice plants were kept in the dark for two weeks verses grown in light (Figure 2.4A). Although the 2.6 and 2.9 kb transcripts are regulated by light, this is not evident for the mature *trnK* transcript. This evidence strongly supports that the 2.6 and 2.9 kb transcripts are separate entities from the mature *trnK* transcript and may be indicative of a *matK* transcript that could continue to translation and not a *trnK* precursor. If the 2.6 and 2.9 kb transcripts were unspliced precursor transcripts of *trnK*, they should accumulate at equal levels as the mature transcript or have a corresponding pattern of expression. Instead, it appears that the mature transcript for *trnK* has constitutive RNA levels in light and dark while the levels of the larger 2.6 and 2.9 kb transcripts are influenced by light (Figure 2.4B). Based on this evidence, an unspliced *trnK* precursor does not appear to accumulate in the plant, but instead is rapidly converted to the mature ~60 base *tRNA*-lysine. Since our signals for the mature *trnK* transcript are near saturation, the upper 2.6 and 2.9 kb transcripts may indicate the premature transcript for this tRNA. However, the very evident difference in the amount of 2.6 and 2.9 kb transcript between light and dark-grown rice plants does not seem to correlate with the slight difference in mature *trnK* transcript in these treatments (Figures 2.4A and B). Therefore, we

propose that the 2.6 and 2.9 kb transcripts indicate RNA levels of the *matK* transcript, which happens to include the *trnK* exons as part of a 5' and 3' UTR, but are not indicative of transcription for the mature *tRNA*-lysine. The 2.6 and 2.9 kb transcripts, therefore, could represent the *matK* transcript proposed in our model that included a 5' and 3' UTR and poly (A) tail, which undergoes processing to produce transcripts of these sizes (Figure 2.7A). Nakamura et al. (2003) observed a five-fold increase in *matK* transcription with light in a tobacco chloroplast microarray study. Their observation of light affecting *matK* RNA levels further supports the etiolation response observed in our study for the 2.6 and 2.9 kb transcripts.

Several chloroplast genes have been shown to have light-induced transcription (Klein and Mullet 1990). As might be expected, many genes that have light-induced transcription are involved in photosynthesis (Klein and Mullet 1990). Although MatK was not directly implicated, a chloroplast-encoded maturase has been suggested to be one of at least two maturases required to process the group II intron of *atpF* (Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999). The protein product of this gene encodes the B subunit of the protein channel CF_0 of chloroplast ATP synthase (Herrmann et al. 1993; Kostrzewa and Zetsche 1993). The activity of ATP synthase is directly related to the formation of a proton gradient during the light dependent reactions of photosynthesis and, thus, can be influenced by photoperiod. The expression of other subunits of ATP synthase have been shown to be up-regulated by light (Jiao, Hilaire, and Guikema 2004; Mackenzie, Johnson, and Campbell 2005). However, in the case of *atpF,* it may be that instead of the expression of this gene being directly up-regulated by light, the expression of at least one of the two maturases (MatK?) needed to splice the group II intron within *atpF* mRNA is induced. This form of regulation where the splicing of an intron regulates the translation of a protein has been observed previously for the

psbA gene of *Chlamydomonas reinhardtii* (Lee and Herrin 2003). *psbA* encodes an essential component of the photosystem II reaction center (Lee and Herrin 2003) and, therefore, is associated directly with photosynthesis in the chloroplast. Light was shown to increase the splicing activity of group I introns in the *psbA* gene of this algae (Lee and Herrin, 2003). We may be observing a similar form of protein regulation with *matK* in which the level of *matK* RNA expression in the plant may directly regulate the amount of MatK protein produced. The quantity of MatK translated may then determine the amount of a properly spliced protein (AtpF) required for photosynthesis.

matK RNA levels during plant development

Although RNA for the *matK* gene was observed at all time points examined, it does not appear to be a constitutive event with equal levels of RNA expressed throughout plant development (Figure 2.4C). Replicates of this experiment produced the same results every time verifying the decrease in *matK* RNA levels observed at 4-weeks post-germination are a product of biological regulation for the amount transcript at this time point. The down-regulation of *matK* mRNA observed in this study may be directly associated with the inhibition of translation for one of its proposed mRNA substrates. Similar to the regulation proposed for AtpF by light induction of MatK, it is possible that the mRNA for proteins involved in plant development are also substrates for MatK splicing. There are several genes observed to be involved in the regulation of plant development (Jan et al. 2005; Wang et al. 2006). So far, substrates for the proposed maturase activity of MatK include transcripts primarily from genes involved in translation, such as *trnK*, *trnA, trnI* (Vogel, Borner, and Hess 1999), and *rpl2* (Hess et al. 1994; Ems et al. 1995). Further investigation would be required to identify which substrate, these or one yet unidentified,

are correlated with plant development and would be down-regulated at four-weeks post germination in rice.

matK transcription across land plants

A *matK* transcript has only been previously identified in the two grasses *O. sativa* and *Hordeum vulgare* (Kanno and Hirai 1993; Vogel et al. 1997), *Nicotiana tabacum* (Nakamura et al. 2003), *Anthoceros formosae* (Kugita et al. 2003), and very recently, *Adiantum capillus-veneris* (Wolf, Rowe, and Hasebe 2004). Only one study investigated *matK* transcription in detail (Vogel et al. 1997). In the present study, we expanded the overall land plant diversity examined, particularly in angiosperms, to further confirm that *matK* is a transcribed gene. We identified a *matK* transcript from a total of eleven plant species (Table 2.3) spanning the land plant phylogenetic tree from bryophytes to angiosperms.

Angiosperms are the largest group of land plants. In order to gain adequate representation of *matK* transcription in this group, we chose representative species from early diverging Nymphaeaceae to the monocots, the asterids (Solanaceae), and rosids (Brassicaceae) (Hilu et al. 2003). Within monocots, representatives included species from early diverging Alismataceae, to members of Poaceae and Orchidaceae (Hilu et al. 2003; Davis et al. 2004). Finding a *matK* transcript expressed in the orchid *Spathoglottis plicata* Blume (Orchidaceae) was of particular interest in this study since the *matK* gene was first noted to be absent in this species (Goldman et al. 2001) and then later reported to be present but as a possible pseudogene (Freudenstein et al. 2004). Our results demonstrate that *matK* is transcribed in this orchid. Further, sequencing of the cDNA amplified product was easily translated into protein sequence that lacked premature stop codons (data not shown).

Monilophytes (ferns and fern allies) have acquired several rearrangements in their chloroplast genome (Wolf et al. 2003). Among these, leptosporangiate ferns are of special interest as they have a large rearrangement in the chloroplast genome that appears to have occurred in the *trnK* intron (Wolf et al. 2003). This rearrangement has resulted in the loss of *trnK*, but retained *matK* (Wolf et al. 2003). Consequently, *matK* RNA obtained from the leptosporangiate fern *Adiantum* could not possibly be the result of an unspliced *trn*K precursor transcript. We detected a *matK* transcript of ~450 bp in *Adiantum hispidulum* (Table 2.3). Although we were unable to sequence the RT- PCR amplified product for this species in our study, Wolf, Rowe, and Hasebe (2004) sequenced a *matK* transcript from another maidenhair species, *A. capillus-veneris*. These findings imply that *matK* is transcribed in *Adiantum*. The identification of a *matK* transcript from *A. capillus-veneris*, and most likely *A. hispidulum,* strengthen the model we are presenting of *matK* as an independent gene with its own transcription regulation and not merely a part of the *trnK* group II intron to be spliced out.

MatK protein

The separation we observed between transcripts for *matK* and that of *trnK* by etiolation suggest that *matK* would be translated into a protein product in the plant. An antibody generated against a 15 amino acid peptide segment from the N-terminal region of *O. sativa* MatK was used in immunoblot analyses of *O. sativa* crude protein extract. A protein of ~65 kDa is expected for MatK based on protein sequence for this gene in rice (GenBank accession: NP_039361). However, protein alignment of MatK sequences from other plant species identified the actual start codon of MatK 31 amino acids further downstream from this sequence. This results in an expected protein of ~61 kDa. Using our MatK antibody, we detected a strong immunoreactive

band of ~55 kDa from Western blots of crude protein extract from *O. sativa* (Figure 2.6), corresponding closely to the predicted size for MatK in this grass. However, using this antibody, we also detected three smaller bands of 40, 30, and 25 kDa (Figure 2.6). Pre-bleed tests identified the ~40 kDa and the ~25 kDa bands as background products from pre-immune serum, but the ~30 kDa band specifically reacted to the MatK antibody (Barthet and Hilu, unpublished). This small protein band could be the result of proteolysis of the full-length MatK polypeptide.

Three studies have previously noted a MatK protein from plant extracts (du Jardin et al. 1994; Liere and Link 1995; Vogel, Borner, and Hess 1999). However, two of these studies identified a band from Western blots that was not of the expected size but substantially smaller than predicted by amino acid sequence (du Jardin et al. 1994; Liere and Link 1995). du Jardin et al. (1994) explain the difference in size by suggesting a hydrophobic structure to MatK that was not denatured by SDS causing faster-migration of the polypeptide than would otherwise have occurred. However, the size discrepancy casts doubt on whether the protein identified from these studies is truly MatK. Vogel et al. (1999), on the other hand, did find a protein band close to the expected size in barley extracts using a MatK antibody against a 274 amino acid residue N-terminal portion of MatK. Our results more closely resemble those from Vogel et al. (1999) and identify a protein near the size predicted from amino acid sequence for MatK. This study and Vogel et al. (1999) confirm the presence of a protein product for MatK and support that an independent transcript for this gene must exist for this protein to be produced.

Could matK be a pseudogene?

Pseudogenes are characterized by a rapid rate of evolution compared to functional homologs (Ophir et al. 1999), a high rate of nonsynonymous mutation (Torrents et al. 2003; Coin and

Durbin 2004), frame-shift mutations that result in premature stop codons, lack of introns and 5' promoter elements, 3' polyadenylation tracts, and direct repeats flanking the coding sequence (Rogers 1985; Vanin 1985; Mighell et al. 2000; Balakirev and Ayala 2003; Coin and Durbin 2004). The *matK* gene has been noted to be a fast-evolving gene compared to other chloroplast genes (Johnson and Soltis 1994; Olmstead and Palmer 1994). It has also been noted to lack an intron (Neuhaus and Link 1987; Hilu and Liang 1997), may contain frame-shifts (Kores et al. 2000; Goldman et al. 2001), and have premature stop codons (Kores et al. 2000; Kores et al. 2001; Kugita et al. 2003; Hidalgo et al. 2004). In addition to these features, the *matK* gene has also been found to have a relatively equal level of substitution at all three codon positions, resulting in a high level of nonsynomous amino acid substitution (Hilu and Liang 1997; Whitten, Williams, and Chase 2000). These features of *matK* tend to suggest that this gene could have evolved into a pseudogene in some species. Since pseudogenes do evolve at a faster pace than their functional homologs (Ophir et al. 1999), systematic analysis using a pseudogene could result in long-branch attraction (Felsenstein 1978) and invalid phylogenies (Bailey et al. 2003).

In contrast to the above characteristics for *matK* that imply a pseudogene possibility, several studies have shown that indels in *matK* usually occur in multiples of three conserving the reading frame (Hilu and Liang 1997; Soltis and Soltis 1998; Hilu, Alice, and Liang 1999; Whitten, Williams, and Chase 2000). Although nonsynomous amino acid substitutions occur in the MatK open reading frame, the rate of nonsynomous substitution is lower than the synonomous substitution (Young and dePamphilis 2000). Tests of evolutionary constraint on the *matK* nucleotide and putative amino acid sequence, which included examining rate of substitution at each codon position and RNA secondary structure, indicated that this gene is under evolutionary constraint (Young and dePamphilis 2000).

In addition to this evidence, *matK* is retained in the holoparasitic plant *Epifagus virginiana* L. (Orobanchaceae) even though 60% of chloroplast genes in this plant have been lost, including *trnK* (Ems et al. 1995). This suggests a functional importance for MatK. *matK* is the only gene that possesses domain X, the putative maturase domain, in the plastid genome of most land plants (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). Although MatK would not function as the group II intron maturase for *trnK* in *Epifagus,* other group II intron containing genes are still expressed in the residual plastid genome of this plant and would require a maturase for splicing (Ems et al. 1995).

Pseudogenes are described as genes that either lack transcription or functional protein product (Mighell et al. 2000). In our study, we have clearly demonstrated that a *matK* transcript can be found in a wide range of land plants and is expressed separate from *trnK*. Thus, this gene is expressed at the RNA level (Figures. 2.2B and 2.5). In addition to this, *matK* RNA levels are not constitutative, but appear to be regulated light and developmental stage in rice (Figure 2.4). This suggests a functional relationship between MatK expression and activity on substrates. A pseudogene would not be under functional constraint and, therefore, would not be expected to have its transcription regulated as if the protein product was to perform a function. RNA editing has also been found in transcripts for *matK* from *H. vulgare* (Vogel et al. 1997) and *A. capillus-veneris* (Wolf, Rowe, and Hasebe 2004). The RNA editing found in *matK* in both cases resulted in a change in amino acid and, in barley, the restoration of a conserved sequence motif (Vogel et al. 1997). This kind of RNA editing would not be a feature of a non-functional protein. In addition to the evidence from transcription analysis that support *matK* functionality, we identified a protein product of the expected size for MatK from protein extracts of *O. sativa* (Figure 2.6). Thus, not only is *matK* transcribed, but also translated to make a product putatively

involved in development and possibly photosynthesis in the plant. These results strongly suggest that *matK* is not a pseudogene but an expressed gene of the chloroplast. Further, the variability in *matK* RNA levels reflects the potential function of MatK in regulating the translation of proteins involved in plant developmental control and photosynthesis at the post-transcription level. Based on the data presented here, we conclude that *matK* has an essential function in the chloroplast. Results from this study support the utility of *matK* in plant systematics. Future studies will directly address the putative function of MatK as a group II intron maturase to further clarify the role of this important gene in the chloroplast.

ACKNOWLEDGMENTS

The authors would like to thank NSF Deep Time, Sigma Xi, Virginia Academy of Science, and Virginia Tech for their support of this research. Special thanks to Dietmar Quandt for help in providing plant material and primer design, Christoph Neinhaus for help in plant collections, and Sabrina Majumder for her assistance in Northern blots experiments.

Chapter 3

MODE AND TEMPO OF PLASTID *matK* GENE EVOLUTION: AN ASSESSMENT OF FUNCTIONAL CONSTRAINT

ABSTRACT

The *matK* gene evolves at a rate three times higher than *rbcL* and *atpB* at the nucleotide level and six times higher at the amino acid level. The high substitution rate in this gene, as well as the presence of indels, has suggested relaxed evolutionary constraint on *matK* and instability in protein structure. Constraint, relating to proteins, can be divided into two categories: structural and functional constraint. Function, like structure, of the protein can be affected by mutation in amino acid sequence, but is also greatly influenced by maintance of overall charge and polarity in functionally important regions of the protein. Substitution of amino acids of similar chemical characteristics in their side chains could act as silent mutations and not affect protein function. The *matK* gene is proposed to encode an essential group II intron maturase of the chloroplast. If functional, evolutionary constraint may exist in MatK to maintain chemical properties of amino acid side chains conserving function of the protein. We have examined putative amino acid sequences for MatK from across green plants to identify regions of amino acid sequence and side chain chemical conservation. In addition, we have compared variability in the composition of amino acid side chains found within MatK to the pseudogene InfA, the slow-evolving protein RbcL, and the mitochondrial maturase Mat-r. Analysis of side chain chemical variability has demonstrated evolutionary constraint on MatK, identified three regions of functional or structural importance, suggested hydrophobic tendencies in this protein similar to other maturases, and supported MatK as a functional protein of the chloroplast.

INTRODUCTION

Among the genes used in plant molecular systematics and evolution, the plastid maturase gene *matK* is unique in its rate of substitution at the nucleotide and amino acid levels. Nucleotide variation in *matK* per site is three times higher than the widely used *rbcL* (Soltis and Soltis 1998). In addition, when comparing *matK* sequences from tobacco and rice, the amino acid substitution rate for *matK* was found to be six fold higher than that of *rbcL* (Olmstead and Palmer 1994), and consequently, it is even higher than that of *psaA* and *psbB* (Olmstead and Palmer 1994; Sanderson 2002). This raises the question of how a gene with these high substitution rates maintains stable protein structure? In this study, we examined chemical variation as it pertains to changes in amino acid categories (acid, base, uncharged at pH = 7, and nonpolar) within the *matK* reading frame to determine elements of protein structure, assessed the evolutionary pattern of variation over phylogenetic distance, and compared patterns of amino acid chemical variation and composition in *matK* from across land plants to the slowly evolving gene *rbcL,* the pseudogene *infA,* and the mitochondrial maturase *mat-r*.

The *matK* gene is located on the large single copy region of the chloroplast genome, nested between the 5' and 3' exons of *trnK*, tRNA-lysine, within a group II intron. The gene is approximately 1500 bp in length, corresponding to 500 amino acids. Previous examination at the nucleotide and amino acid levels indicated that *matK* is not homogenous across its ORF but exhibits varying rates of substitution (Hilu and Liang 1997). One highly conserved region of 448 bp is located in the 3' terminus of *matK* (Hilu and Liang 1997). Sequence analysis indicated that this region displayed homology to domain X of mitochondrial group II intron maturases (Sugita et al. 1985; Neuhaus and Link 1987). Based on rate of nucleotide substitution (Hilu and Liang

1997; Soltis and Soltis 1998) and RNA secondary structure (Young and dePamphilis 2000), it was suggested that functional constraint exists in domain X.

matK is the only gene found in the chloroplast genome of higher plants that contains this putative maturase domain (Neuhaus and Link 1987). There are 16 group II introns nested within 15 chloroplast genes (Kohchi et al. 1988; Ems et al. 1995; Maier et al. 1995), which would require a maturase for intron splicing and proper protein translation. Studies of the white barley mutant *albostrians* demonstrated that although some group II introns were processed by an imported nuclear maturase, there were at least five plastid genes with group II introns that would require a chloroplast maturase for splicing (Vogel, Borner, and Hess 1999). Their Western blot analysis indicated that a protein is produced from the *matK* gene. This evidence suggested a potential functional role for MatK as a group II intron maturase in the chloroplast.

How does MatK accommodate the high nonsynonymous mutation rate and remain functional? A high rate of nonsynonymous substitution in a gene may lead to instability in protein structure and loss of function. However, chemically conserved amino acid replacement, the most prominent form of amino acid replacement in functionally or structurally important regions of protein-coding genes, may act as silent mutations, minimizing the impact on protein structure and/or function (Clarke 1970; Graur 1985; Graur and Li 1988; Wolfe and dePamphilis 1998). Should this mode of substitution be demonstrated for *matK* , it would point to purifying selection on MatK protein structure and function. Although the high rate of nonsynonymous mutations in *matK* would suggest neutral selection (Hilu and Liang 1997; Soltis and Soltis 1998; Whitten, Williams, and Chase 2000; Hilu et al. 2003), purifying selection was previously observed in *matK* at the level of RNA structure (Young and dePamphilis 2000), but has not been demonstrated at the amino acid level. Considering the putative functional importance of MatK

in the chloroplast, we hypothesized that the nonsynomous amino acid substitutions found in MatK are under evolutionary constraint with regard to conserving the category of amino acid side chains (acid, base, uncharged pH = 7, nonpolar), maintaining structure and function of this protein. Thus, variation in MatK amino acid category composition, refered to hereafter as side chain composition, should deviate significantly from a pseudogene and more similar to genes known to be functionally constrained.

In addition to the high rate of substitution, *matK* also displays varying number and size of insertions and deletions (Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003). Most indels identified in *matK* from several plant taxa have been found in multiples of three, conserving the reading frame (Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003). Apparent frame shift indels found in some members of the Orchidaceae (orchid family) have led some to suggest that *matK* might be a pseudogene in this family (Kores et al. 2000; Whitten, Williams, and Chase 2000; Goldman et al. 2001; Kores et al. 2001; Salazar et al. 2003). We have examined *matK* sequences in these species for the potential presence of alternate out-of-frame start codons that may correct the reading frame and eliminate premature stop codons in the Orchidaceae.

Information from this study will provide insight into how mode and tempo of nucleotide and amino acid substitution can impact protein structure in this rapidly evolving gene. This information will also be important for *matK* utility in systematic and evolutionary studies since pseudogenes evolve at different rates than their functional counterparts (Ophir et al. 1999; Mighell et al. 2000) resulting in alignment problems (Bailey et al. 2003) and phylogenies with long-branch attraction (Felsenstein 1978; Bailey et al. 2003).

MATERIALS AND METHODS

Sources of MatK sequences, alignment, and phylogenetic analysis

matK amino acid sequences from 122 species representing 60 families were used in the various analyses of this study (Supplemental Tables 1, Appendix A). This large data collection was split into smaller sets for different analyses as will be discussed below. *matK* sequences for all analyses, except two from the genus *Spathoglottis* (Orchidaceae), were downloaded from GenBank. *Spathoglottis* sequences were generated from amplified DNA in this study and will be discussed in the methods section directly related to the Orchidaceae.

Amino acid sequences of 31 of the 122 species noted above were aligned using MacVector© and Accelrys DsGene© (data set C, Supplemental Table 1, Appendix A) and used for determining separation between the N-terminal region and domain X of MatK as well as division of MatK into the seven sectors describe below. These taxa included one green algae, two bryophytes, two monilophytes, four gymnosperms, and 23 angiosperms. MatK amino acid sequences were either downloaded directly from the protein database of GenBank or translated using MacVector© if a protein sequence was unavailable (e.g. for *Huperzia*). To ensure that amino acid alignments were accurate, phylogenetic trees were constructed in Mega (version 3.1) (Kumar, Tamura, and Nei 2004) using Maximum Parsimony and Neighbor-Joining analysis and in PAUP* (version 4.0b6) (Swofford 2001) using 50% majority-rule with strict consensus in heuristic search step-wise addition for comparison. A consensus tree was compiled from all possible tree combinations. Bootstrap values for Mega trees were calculated using 1000 replicates. In both PAUP* and Mega, gaps were treated as missing data and all aligned positions were given equal weight.

MatK Amino acid and side chain composition

A data set of 52 species from 45 families (Supplemental Table 1, Appendix A) spanning green plants were evaluated for MatK amino acid composition using SAPS (Brendel et al. 1992) in Biology Workbench® SDSC. This data set was termed "set A" for simplicity (Supplemental Table 1, Appendix A). Forty-one of these species represent various angiosperm lineages, while the remaining species represent algae (1), bryophytes (3), monilophytes (3), and gymnosperms (4). Two of these gymnosperms, *Gnetum gnemon* and *Welwitschia mirabilis,* are members of the Gnetales, a group known to have unique and conflicting morphological and molecular characters (Donoghue and Doyle 2000; Magallon and Sanderson 2002). A separate analysis of 16 species (one green algae, three bryophytes, three monilophytes, eight gymnosperms including six gnetophyta, and one angiosperm, data set B, Supplemental Table 1, Appendix A) was performed to examine characteristics of side chain composition in this plant group in detail. Amino acids were classified into the following four categories based on their side chains: basic, acid, polar uncharged at pH = 7, and nonpolar (Alberts et al. 1994), and the composition of each amino acid category was evaluated as a percentage of total protein. This composition is referred to as the side chain composition. Since standard deviation (SD) represents variation from a mean, variation in side chain composition of MatK across plant groups was determined by calculating the SD for each of the four amino acid categories among all taxa. In addition, the SD and average percent of each amino acid in total protein was determined for all species.

Since previous study of *matK* sequences demonstrated that nucleotide and amino acid substitution rates were not homogenous throughout the reading frame (Hilu and Liang 1997), we divided the putative MatK reading frame of 31 green plant species (data set C, Supplemental Table 1, Appendix A) into two domains: N-terminal and domain X, as determined by Pfam

(Bateman et al. 2002). MatK was further subdivided into sectors to ensure that regions unidentified by Pfam but may still hold functional or structural importance by conserved side chain composition were not disregarded. Since the MatK reading frame contains several indels that could affect the position of regions of sequence composition, our division of MatK into sectors followed conserved regions determined by MeMe (Bailey and Elkan 1994). This approach resulted in seven sectors of almost equal size. Each sector contained an average of 72 amino acids plus or minus 3 with the exception of the carboxy terminus sector, which contained an average of 80 amino acids. The relative equality between segments allowed for accurate statistical measurement. Amino acid and side chain composition of the N-terminal domain, domain X, and the seven sectors, was analyzed using Biology Workbench and SAPS (Brendel et al. 1992). The SD for each amino acid catagory was determined to compare the extent of variability in side chain composition within these divisions of MatK. An average of the SD for all four amino acid categories was then used as the measure of variation for each sector or domain. Likewise, the average SD of the percent of each amino acid comprising a sector for all taxa examined was calculated in Excel. To determine the impact of indels on amino acid and side chain composition in the MatK ORF, putative MatK amino acid sequences from 13 angiosperm and one gymnosperm species (data set D, Supplemental Table 1, Appendix A) that contained only small (2 amino acid) indels were examined for composition. The SD in side chain composition and variation in percent of each amino acid was determined for these 14 species and compared to that of data sets A and B.

Deviation in side chain composition at various genetic distances

We tested if the peaks and valleys in *matK* nucleotide substitution observed by Hilu and Liang (1997) were also reflected in side chain composition. Ten species of *Oryza* (rice) were used to examine intrageneric differences in composition (data set E, Supplemental Table 1, Appendix A). Since sequences available for *matK* were limited for some genera and plant families, three species of each group were used in analysis above the intrageneric level for balanced statistical calculation. Three species from each of the three grass genera, *Oryza, Hordeum* (barley), and *Sporobolus* (prairie dropseed) (data set E, Supplemental Table 1, Appendix A) were used to assess intergeneric variation. To examine variation at the ordinal level, three genera from each of three families, Poaceae, Joinvileaceae, and Restionaceae (data set E, Supplemental Table 1, Appendix A) of the order Poales were compared. Variation among monocots was evaluated using the three orders Alismatatales, Poales, and Asparagales, that span the monocot phylogenetic tree (Hilu et al. 2003). Variation in side chain composition was then analyzed across 40 angiosperm species from 39 families (data set E, Supplemental Table 1, Appendix A), and 52 green plant species from 52 families (data set A, Supplemental Table 1, Appendix A). Since members of a genus cannot be equally compared to members of an order due to magnitude of genetic distance, the coefficient of variation was used for a more accurate depiction of change in variation from shallow to deep level phylogeny. To ensure that our analysis was not skewed by our choice in model and to not limit the analysis to strictly monocots, the same experiment was conducted using the eudicot family Brassicaceae and *Arabidopsis* as the model (data set E, Supplemental Table 1, Appendix A).

Transmembrane and secondary structure prediction

Transmembrane regions in MatK for 22 species of data set A spanning green plants were predicted using TMAP and TMHMM computer programs through Biology Workbench. In addition, the maturases Mat-r (GenBank accession: AE47664), Cob I1 (GenBank accession: X54421*),* Cox1 I2 (GenBank accessions: NC_005256 and CAC28096), and LtrA (GenBank accession: P0A3U0) were also examined for putative transmembrane segments with these programs. Secondary structure for MatK amino acid sequence from *Oryza sativa* (GenBank: NP_039361) was predicted using JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/submit.html). The raw protein sequence for *O. sativa* was entered in the appropriate field and default settings maintained for the prediction. This method of structure prediction has been demonstrated to be effective for group II intron maturases (Blocker et al. 2005).

Analysis of the Orchidaceae

Because of previous implications that *matK* is a pseudogene in several members of the Orichidaceae (Kores et al. 2000; Goldman et al. 2001; Kores et al. 2001; Salazar et al. 2003), we examined this particular plant family in detail. A separate nucleotide and amino acid data set of 50 species representing 29 seed plant families were used for analysis of the Orchidaceae (Supplemental Table 2, Appendix A). *matK* sequences for all Orchidaceae, with the exception of two sequences for species in the *Spathoglottis* genus were downloaded from GenBank (Supplemental Table 2, Appendix A). Genomic DNA from the two orchids, *Spathoglottis plicata* and *Spathoglottis gracilis,* was isolated according to Doyle and Doyle (1990). The primers SpRend (5' TTATGTAATCCGTGTAAAAT 3') and 5UTRSp (5' CATTTCATAACACAAGAA 3') were designed in this study to amplify the entire *matK* gene

from both orchid species. The PCR protocol used for this amplification was: 1) starting cycle of 95 ^{o}C for 3 min., 45 ^{o}C for 3 min., and 72 ^{o}C for 3 min., 2) main cycle program of 95 ^{o}C for 30 sec., 45 ^{o}C for 1.30 min., and 72 ^{o}C for 3 min., with main cycles repeated 50 times, and 3) 72 ^{o}C for 20 min. to complete end extension. Amplified products were separated on 0.8% agarose gels and purified using the Qiagen Qiaquick PCR extraction kit (Qiagen, Valancia, CA, USA). Cycle sequencing was accomplished using 2.5 μl ABI BigDye Terminator sequencing dye (ABI, Framingham, MA, USA) and a final concentration of 1.25 μM of primer in a 15 μl reaction performed following the ABI cycle sequencing protocol (Framingham, MA, USA). Sequencing was accomplished by the external primers SpRend and 5UTRSp, and the three internal primers SpRseq169 (5' CTTGTGCCCCCAAAAGCCA 3'), OcmatKF (5' CTCACTTGCTCATKATC 3'), and SpwalkF (5' TACCCCATCCCGTCCAT 3') designed in this study. Products of the sequencing reaction were separated through capillary gel electrophoresis at the Virginia Bioinformatics Core Laboratory Facility (Blacksburg, VA, USA). Sequencing was performed for both strands, and forward and reverse sequences were aligned in QuickAlign (Müller and Müller 2003) to generate the full-length *matK* sequence for the two orchid species. An amino acid alignment was constructed using MatK protein sequences in MacVector© and Accelrys DsGene©. For those orchids noted to contain *matK* as a pseudogene, the nucleotide sequence was imported into Accelrys DsGene© and translated into protein sequence for the alignment.

Analysis of side chain composition among four proteins

The high degree of nucleotide and amino acid sequence variation in *matK* led some to suggest that *matK* may be a pseudogene (Whitten, Williams, and Chase, 2000; Kores et al. 2001). To address this point, variation in side chain composition in the putative MatK protein was

compared to that in the pseudogene InfA (Millen et al. 2001), the slow-evolving functional protein RbcL (Wolfe, W.-H.Li, and Sharp 1987; Chase et al. 1993; Kellogg and Juliana 1997), and the mitochondrial maturase Mat-r (Farré and Araya 1999). Comparisons among MatK, InfA, and RbcL used 22 species from 16 families (Supplemental Table 3, Appendix A). The number of species used was constrained by availability of sequences from GenBank of species that have all three genes/proteins. The chloroplast *infA* gene has been horizontally transferred to the nucleus in several plants, but, a residual pseudogene copy has been retained in several chloroplast genomes (Millen et al. 2001). Nucleotide sequences for i*nfA* pseudogenes were translated in MaCVector© into amino acid sequence containing premature stop codons. Thirty-two species from 31 families across land plants were used to compare variation in side chain composition between MatK and Mat-r (Supplemental Table 4, Appendix A). Again, the number of species used was limited to sequences available from Genbank for both proteins. Only complete proteins sequences from the same species for each protein were used to avoid deviation in side chain composition caused by using missing data. All sequences for each of the four genes/proteins compared, MatK, RbcL, InfA, and Mat-r, were downloaded from GenBank. Side chain composition was determined as stated previously using SAPS (Brendel et al. 1992) and SD, calculated in Excel, for each amino acid catagory was used as a measure of variation in side chain composition. A Student's T-test from Excel using a two-tailed distribution and assuming equal variance was used to determine statistical significance of variation in side chain composition between two proteins of different genes.

RESULTS

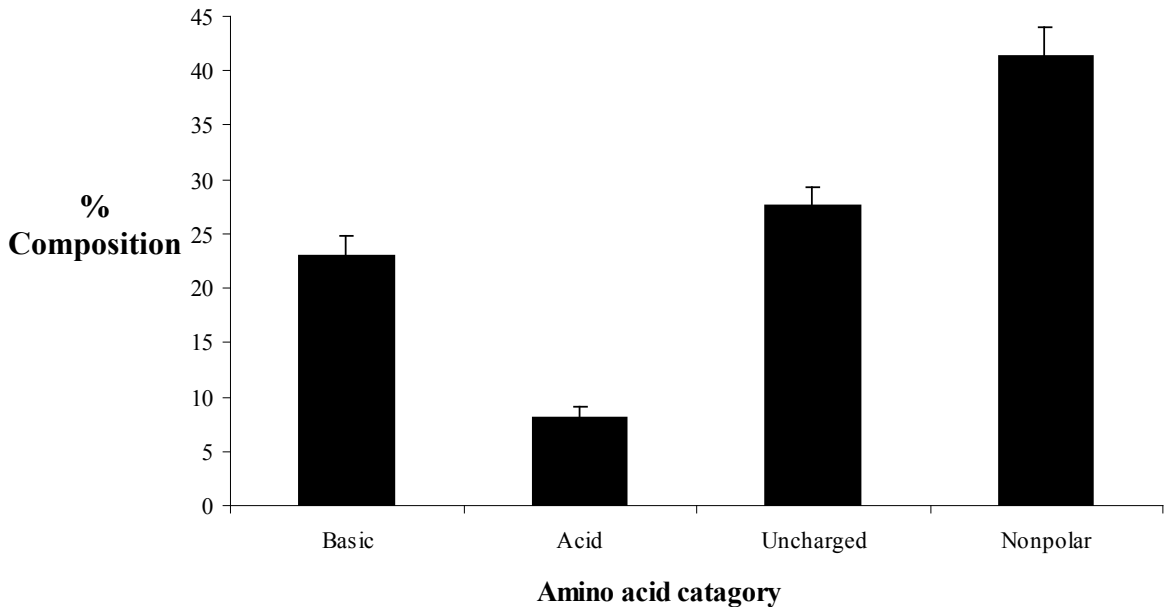
Overall amino acid and side chain composition

The amount of each amino acid category in the putative MatK reading frame of 52 green plant species revealed an average of 17.2% (SD = 1.24) basic amino acids, 8.1% (SD = 0.89) acidic amino acids, 27.6% (SD = 1.73) uncharged (pH = 7) amino acids, and 47% (SD = 1.63) hydrophobic amino acids (Figure 3.1A). Side chain composition was fairly uniform across all groups with the exception of Gnetophyta (Figure 3.1B). Excluding Gnetophyta, the average variability in uncharged (pH =7) and nonpolar amino acids in MatK decreased to 1.2. An analysis using 16 species (data set B, Supplemental Table 1, Appendix A), showed a statistically significant 16% increase in nonpolar (hydrophobic) amino acids (Student's t-test, p = 0.0001) and a corresponding decrease in uncharged (pH =7) amino acids in the gnetophyta compared to the other green plant species.

Amino acid composition of MatK showed leucine and serine to comprise the highest percent (12.6 and 10.3, respectively), while cysteine, methionine, and tryptophan consisted of the lowest percent (1.5, 1.5, and 1.6, correspondingly, Table 3.1) of any amino acid in total protein. In terms of variability, lysine was the most variable in percent amino acid, while tryptophan was the least variable (Table 3.1). Glycine, the smallest amino acid, comprised only 3.3% of total amino acids in MatK and had a standard deviation of 0.7 (Table 3.1).

TMHMM and TMAP (Persson and Argos 1994) transmembrane prediction programs in Biology Workbench® SDSC identified 1-6 putative transmembrane domains in MatK protein from most taxa except for the fern Psilotum nudum, which did not contain any putative transmembrane segments (data not shown). For example, four transmembrane domains were

(A)


% Composition
45
40
35
30
25
20
15
10
5
0
Basic
Acid
Uncharged
Nonpolar
Amino acid catagory

(B)

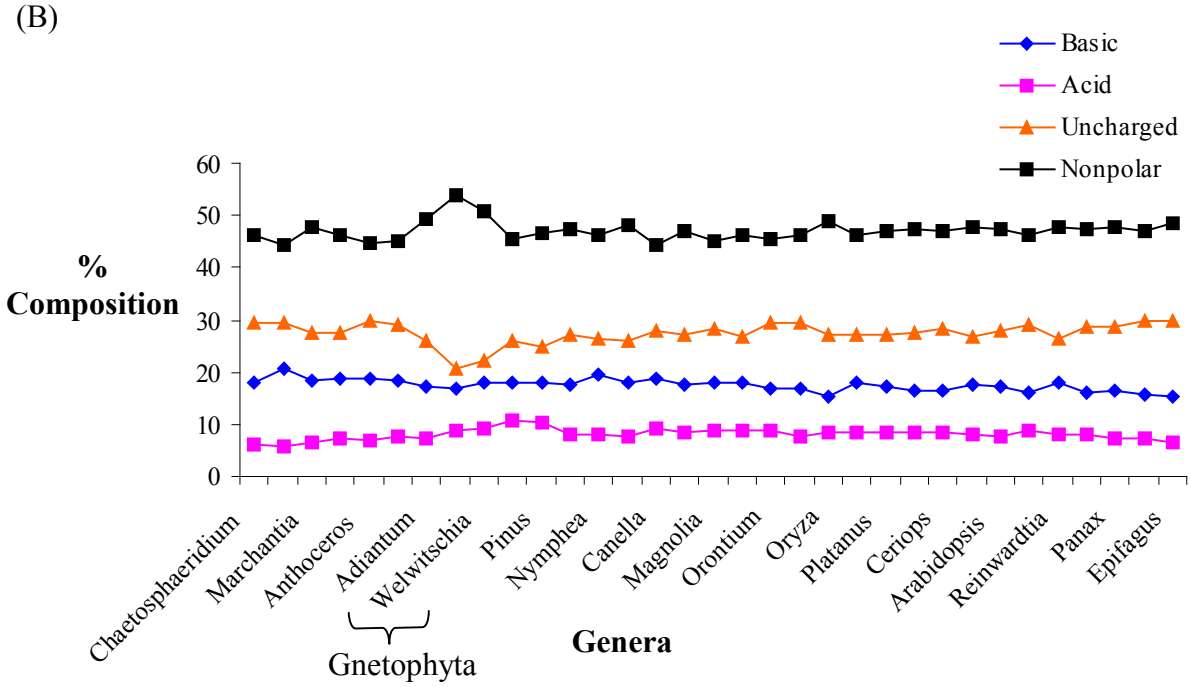

Basic
Acid
Uncharged
Nonpolar
% Composition
60
50
40
30
20
10
0
Chaetosphaeridium
Marchantia
Anthoceros
Adiantum
Welwitschia
Pinus
Nymphea
Canella
Magnolia
Orontium
Oryza
Platanus
Ceriops
Arabidopsis
Reinwardtia
Panax
Epifagus
Gnetophyta
Genera

Figure 3.1.—Side chain composition of putative MatK protein from 52 species (data set A). Amino acid categories were defined according to Alberts et al. (1994) and consist of the following amino acids: basic K, R, H; acidic D, E; uncharged at pH = 7 N, Q, S, T, Y; and nonpolar A, G, V, L, I, P, F, M, W, C. Error bars represent standard deviation. (A) Average percent of each amino acid catagory in total protein. (B) Side chain composition of MatK across 52 green plant genera. The genus *Welwitschia* is shown as one of two genera of the gnetophytes used in this analysis. Plant taxa are arranged in phylogenetic order from left to right.

TABLE 3.1.

Amino acid composition of MatK from 52 green plant species (data set A).

Amino acid	Average % in total protein	Standard Deviation
A Ala	2.55	0.63
C Cys	1.45	0.48
D Asp	3.35	0.68
E Glu	4.77	0.65
Fphe	8.16	1.64
G Gly	3.3	0.7
H His	4.14	1.17
I Iso	8.36	1.19
K Lys	6.87	1.81
L Leu	12.57	1.02
M Met	1.525	0.48
N Asn	5.29	1.04
P Pro	3.23	0.67
Q Gln	3.57	0.68
R Arg	6.2	1
S Ser	10.28	1.24
T Thr	3.03	0.67
V Val	4.34	0.84
W Trp	1.6	0.35
Y Tyr	5.43	0.88

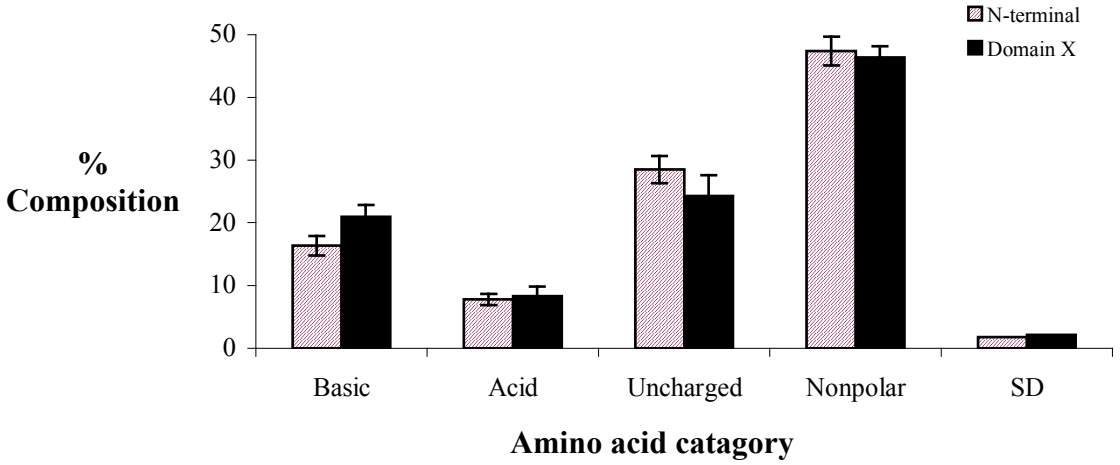
predicted in the MatK ORF of the green alga *Chaetosphaeridium*, five to six in members of Gnetophyta, and at least one to four in other land plants. TMAP analysis of protein coding sequences in this study predicted transmembrane domains in the ORF of four other group II intron maturases: Mat-r, Cob I1, Cox1 I2, and LtrA.

Side chain composition across MatK ORF

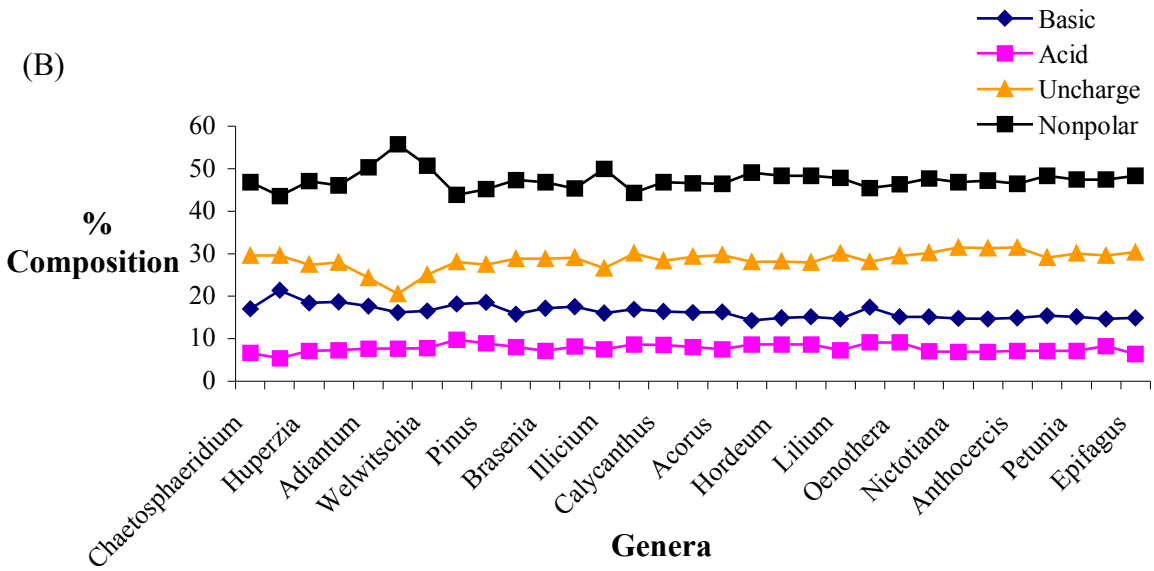
Analysis of 31 green plant taxa (data set C, Supplemental Table 1, Appendix A) demonstrated that domain X contained approximately 28% more basic amino acids and 15% less uncharged amino acids than the N-terminal region (Figure 3.2A). However, the two regions, N-terminal and domain X, displayed very similar deviation in basic composition (SD = 1.6 to 1.8, respectively). Both nonpolar and acidic side chain composition was very similar between the two regions (Figure 3.2A). The average variation for all four amino acid categories between the N-terminal region and domain X was not statistically significant (Student's t-test, $p = 0.36$). Side chain composition remained fairly stable in the N-terminal region across plant phylogeny, with the exception of the Gnetophyta (Figure 3.2B). In contrast, the percent of uncharged and basic amino acid in domain X fluctuated greatly among all genera (Figure 3.2C).

Considering the side chain composition of the seven sectors, sectors 3 and 5 were highly hydrophobic, while sector 4 contained lower hydrophobicity and more polarity (Figure 3.3A). Sectors 1 and 7, which correspond to the beginning and very end of MatK, appeared as "hot

(A)


N-terminal
Domain X
% Composition
50
40
30
20
10
0
Basic
Acid
Uncharged
Nonpolar
SD
Amino acid catagory

(B)


Basic
Acid
Uncharge
Nonpolar
% Composition
60
50
40
30
20
10
0
Chaetosphaeridium
Huperzia
Adiantum
Welwitschia
Pinus
Brasenia
Illicium
Calycanthus
Acorus
Hordeum
Lilium
Oenothera
Nictotiana
Anthocercis
Petunia
Epifagus
Genera

(C)

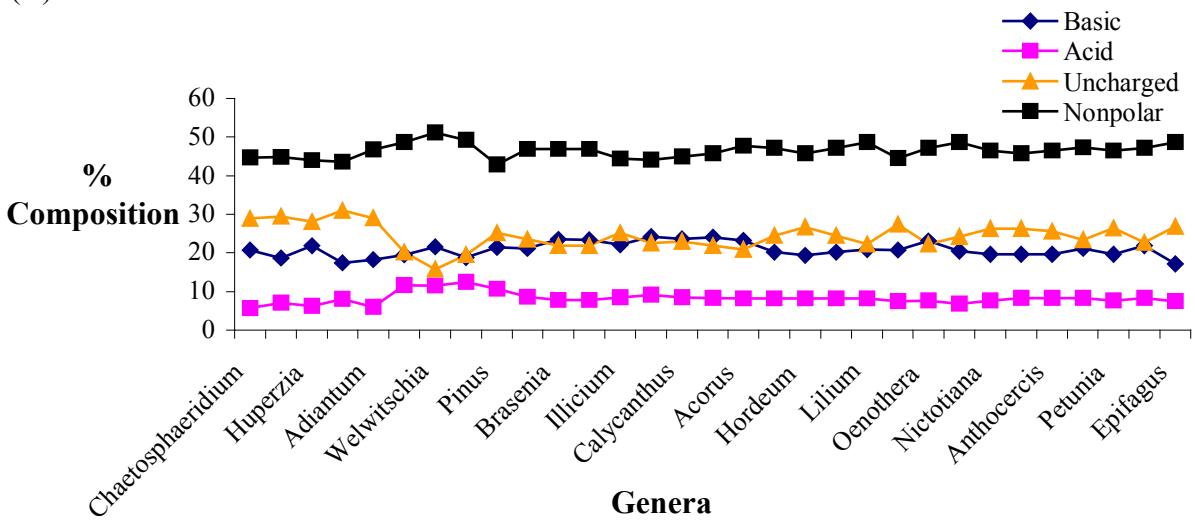
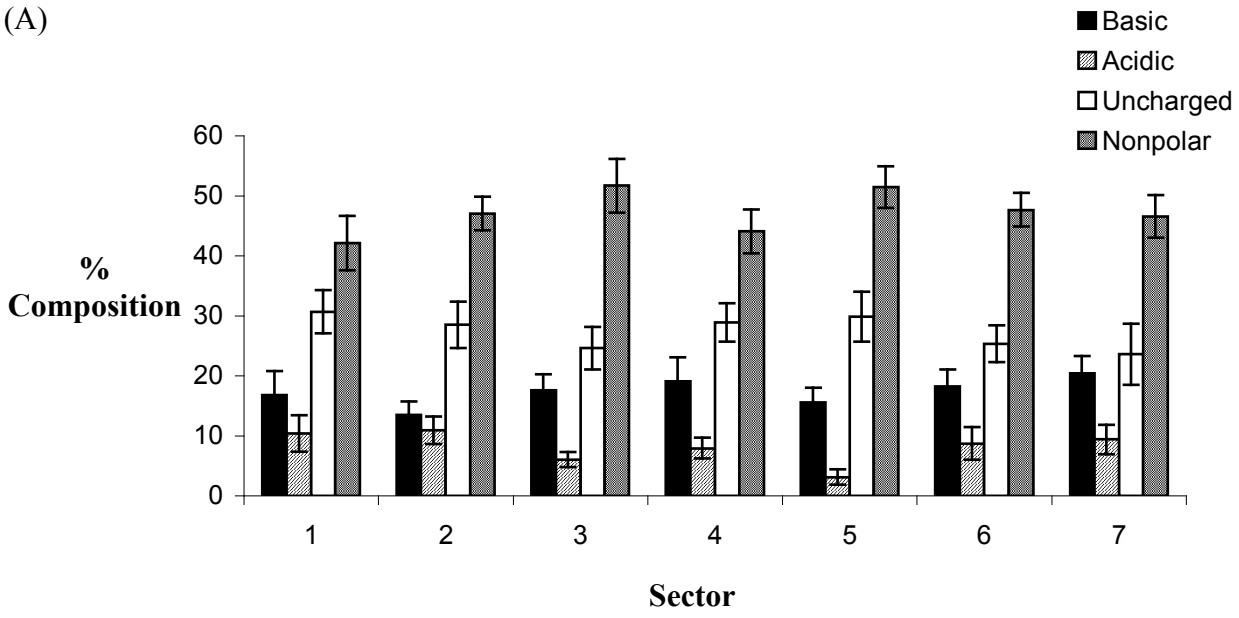
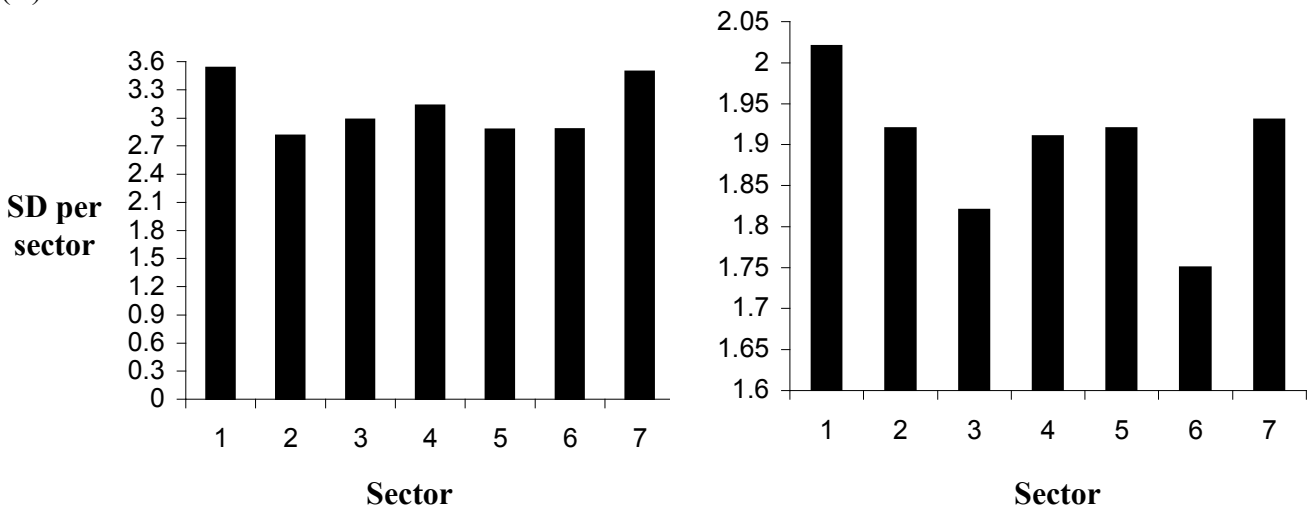

Basic
Acid
Uncharged
Nonpolar
% Composition
60
50
40
30
20
10
0
Chaetosphaeridium
Huperzia
Adiantum
Welwitschia
Pinus
Brasenia
Illicium
Calycanthus
Acorus
Hordeum
Lilium
Oenothera
Nictotiana
Anthocercis
Petunia
Epifagus
Genera

Figure 3.2.--Comparison of side chain composition in MatK between the N-terminal region and domain X for 31 taxa of green plants (data set B). (A) Average side chain composition and standard deviation of these two domains. Error bars represent standard deviation. (B) Evolution of the N-terminal region, or (C) domain X across green plants. Plant genera are arranged in phylogenetic order from left to right. Amino acid categories are defined in Figure 3.1.

(A)


Basic
Acidic
Uncharged
Nonpolar
% Composition
60
50
40
30
20
10
0
1
2
3
4
5
6
7
Sector

(B)


SD per sector
3.6
3.3
3
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0
1
2
3
4
5
6
7
Sector
2.05
2
1.95
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1
2
3
4
5
6
7
Sector

(C)

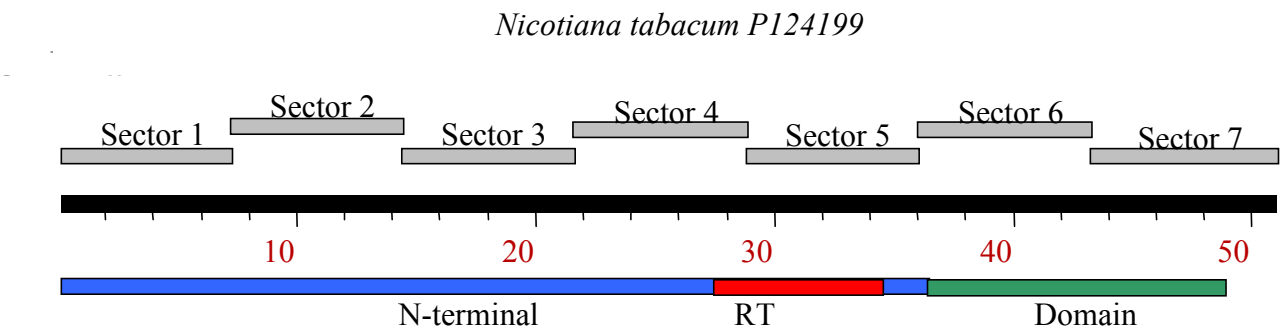

Nicotiana tabacum P124199
Sector 1
Sector 2
Sector 3
Sector 4
Sector 5
Sector 6
Sector 7
10
20
30
40
50
N-terminal
RT
Domain

Figure 3.3.—Side chain and amino acid composition and variation across the MatK reading frame of 31 species (data set C). (A) Average side chain composition for each sector of MatK. Error bars represent standard deviation. (B) Standard deviation in (left) side chain and (right) amino acid composition for MatK sectors. Amino acid categories are defined in Figure 3.1. (C) Arrangement of sectors and domains along the MatK reading frame of *N. tabacum*. Sectors delineated at the ends by consensus sequence determined with Meme (Bailey and Elkan 1994) and by eye. The position of the N-terminal region and domain X were determined by domain analysis in Blastp. The location of the RT domain was based of consensus sequence determined by Mohr et al. (1993).

spots" of high deviation in side chain composition and amino acid makeup (Figure 3.3B). Sectors 2, 5 and 6 had the lowest variation in side chain composition (Figure 3.3B), and were almost equal in their standard deviation (2.81, 2.87, 2.88, respectively). Sector 6 displayed the lowest deviation in amino acid composition of any sector (Figure 3.3B).

Sector 2 showed homology, though low (Pfam; E-value < 2), to bacterial proteins of the DUF877 family. Pfam (E-value <2) also identified sector 5 as having homology to three proteins: a glycoprotein hormone, the DUF636 protein family, and a ZZ type zinc finger protein. However, Pfam (E-value <2) did not identify the homology of sector 5 to the reverse transcriptase (RT) domain reported for other maturases (Mohr, Perlman, and Lambowitz 1993), although it was detected here using amino acid sequence analysis (Figure 3.3C). Sector 6 had high homology to domain X (Figure 3.3C).

To determine the impact of indels on the assessment of side chain composition in the MatK ORF, a data set of 14 taxa (data set D, Supplemental Table 1, Appendix A) that had minimal amount of indels was analyzed. The pattern of side chain composition and variation for the whole putative protein, the N-terminal and domain X, and the seven sectors in this reduced data set was not notably different from data sets A and C (Supplemental Table 1, Appendix A) that included indels. However, the SD in side chain composition was lower for these 14 taxa in all three assessments compared to those of than the larger data sets A and C (comparisons for the seven sectors is shown in Table 3.2). The lower SD observed for data set D when analyzing variation in side chain composition for the whole protein and for the N-terminal region and domain X was not statistically significant from the SD of the larger data sets (data set A and C, respectively). Though, when comparing the SD of each sector in MatK for this smaller data set of 14 taxa to the SD from data set C, the reduced variation observed with data set D was

TABLE 3.2.

Comparison of SD in side chain composition of

MatK between taxa with and without indels

(data sets C and D).

Sector	Standard deviation for 31 taxa	Standard deviation for 14 taxa lacking indels
1	3.53	2.95
2	2.81	2.35
3	2.98	2.35
4	3.13	2.69
5	2.87	2.27
6	2.88	2.59
7	3.49	2.87

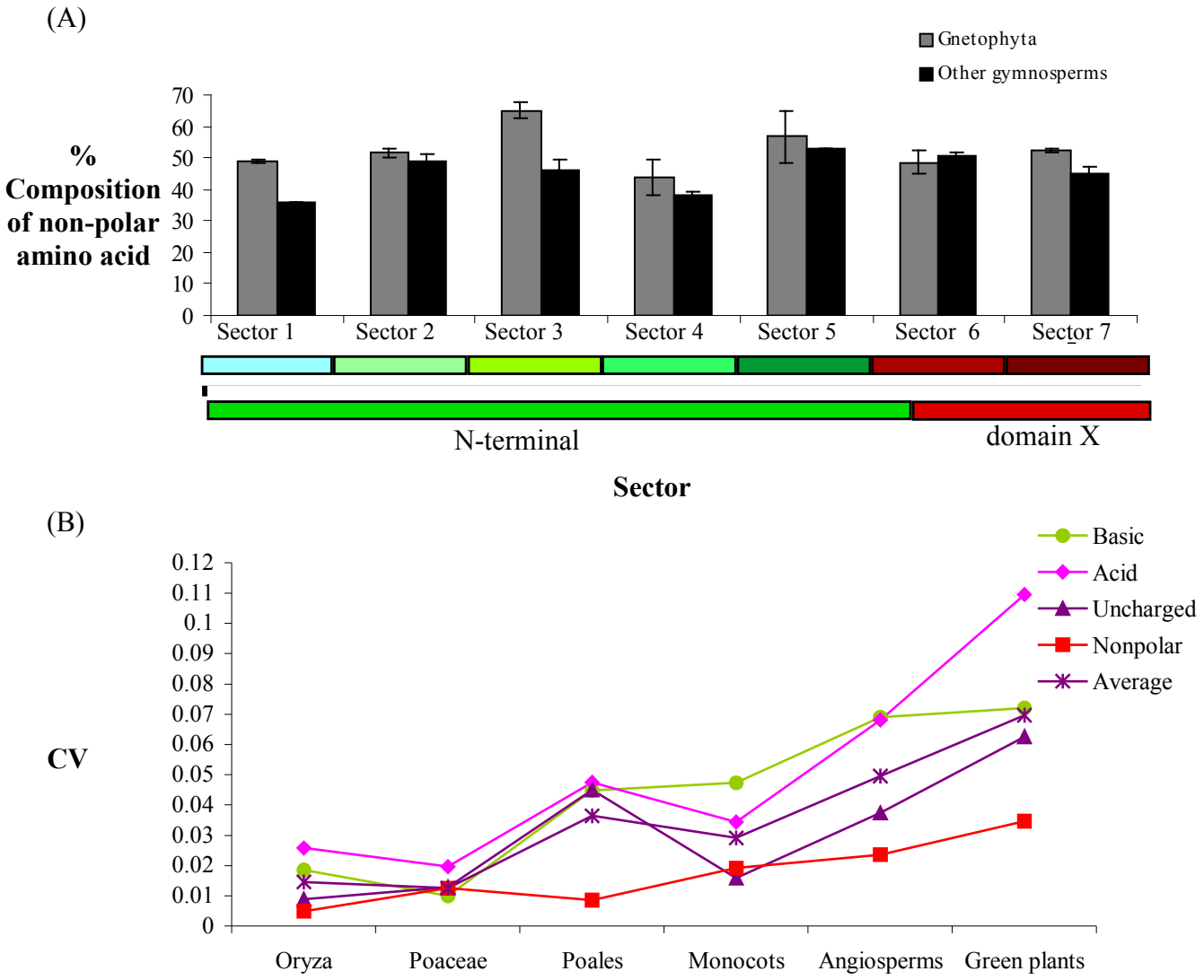


Figure 3.4.--Analysis of side chain composition and variability in gymnosperms and across phylogenetic distance in green plants. (A) Comparison of the percent of nonpolar amino acids in the seven sectors of MatK between different groups of gymnosperms. Gnetophyta are represented by *Gnetum* and *Welwitshia* while "other gymnosperms" are represented by *Ginkgo* and *Pinus*. Error bars represent standard deviation. (B) Coefficient of variation of side chain composition of MatK at different genetic distances from the intrageneric level using the genus *Oryza* as the model to across green plants. Assessment of deviation in side chain composition focused on monocots, specifically the Poaceae.

statistically significant for three of the seven sectors, specifically sectors 2, 3, and 7. Further, the pattern of variation in amino acid composition was different for these 14 taxa than observed with data set C with the exception of sector 6 which displayed low in variability for both analyses (data not shown).

In Gnetophyta, subdivision of the MatK reading frame into the two regions, N-terminal and domain X, indicated that the increase in hydrophobicity observed in the plant group was localized primarily to the N-terminal region of MatK (Figure 3.4A). When further subdividing this N-terminal region into sectors, sector 3 was identified to contain the highest proportion of nonpolar amino acids of any sector in MatK (65% hydrophobic) for the Gnetophyta (Figure 3.4A). The high hydrophobicity of sector 3 was not observed in other gymnosperms (Figure 3.4A). A putative transmembrane domain was predicted using TMAP in this sector exclusively for members of this plant group (data not shown).

Variation in side chain composition across green plants

Deviation in side chain composition in MatK reading frame was evident at various taxonomic higherarchial levels. Comparison of side chain composition among ten species of *Oryza* showed a coefficient of variation (CV) of 0.014. The side chain composition did not deviate much (CV = 0.013) at the intergeneric level when *Oryza* was compared with two other Poaceae genera (*Hordeum*, *Sporobolus*; the three belong to different subfamilies). However, the CV value increased by nearly three fold (0.036) when three families of the order Poales (Poaceae, Joinvilleaceae, Restionaceae) were contrasted. Among the monocot orders Alismatales, Poales, and Asparagales, comparison of the side chain composition demonstrated a CV of 0.029. The coefficient of variation in side chain composition in angiosperms using forty species from 39 families was 0.049, whereas for green plants (52 species, 45 families) it was 0.07.

Looking at changes in percent variance of each of the four amino acid categories, a slightly different pattern was observed. When comparing at the phylogenetic level across three different orders of monocots, variation in basic amino acids slightly increased, while variation in acidic and uncharged (pH =7) amino acids decreased (Figure 3.4B). However, above this level (deeper in phylogeny), variation in basic amino acids leveled off whereas the other three amino acid categories displayed an increase in variation, particularly the acidic group (Figure 3.4B). A significant change in pattern of variation in side chain composition occurred with nonpolar amino acids. In contrast to the other amino acid categories, variation in this amino acid catagory decreased among families of the same order compared to variation within the same family (CV = 0.0126, Poaceae, to CV = 0.0085, Poales, Figure 3.4B) but continually increased at deeper phylogenetic levels (Figure 3.4B).

In a similar analysis using the eudicot *Arabidopsis* and the Brassicaceae as the model (instead of *Oryza* and the Poaceae) a similar pattern of variation was detected at the different phylogenetic levels with one exception (data not shown). Unlike what was observed using the monocots, variation in acidic composition in the eudicots dramatically decreased at the interfamilial level (CV = 0.006) compared with variation in acidic amino acid composition at the intergeneric level (CV = 0.04). All of the other amino acid categories had a pattern of variation at each phylogenetic level similar to that of the monocots (data not shown).

JPRED Secondary Structure

Secondary structure prediction of the MatK reading frame from *Oryza sativa* (GenBank: NP_039361) using JPRED identified 19 α-helical regions and 12 β-strands of substantial length

(ie. more than three amino acids in a row) (Figure 3.5). The placement of seven of these helices along the MatK reading frame was found to be similar to positions noted for α-helices in the LtrA group II intron maturase (Blocker et al. 2005) by visual assessment. The first helix found at the beginning of MatK contains a stretch of 7 amino acid residues (Figure 3.5 box a) verses the 8 residues of LtrA (Blocker et al. 2005). The two helices in region 3a of LtrA contain 11 residues each (Blocker et al. 2005), while 13 residues are found in each for MatK (Figure 3.5, box b). Three α- helices in domain X of MatK (Figure 3.5) coincide to helices αH, αI, and αJ in domain X of LtrA but with slightly different lengths.

Side chain composition in MatK verses InfA, RbcL, and Mat-r

Deviation in composition among proteins was determined by calculating the standard deviation for the percent of each the four amino acid categories for entire ORFs of 22 species representing 16 families from across green plants (Table 3.3). The pseudogene *infA* displayed a much greater degree of deviation from its mean side chain composition than either MatK or RbcL (Figure 3.6A). Variation in side chain composition between InfA and MatK was statistically significant (Student's t-test, $p = 0.02$), in contrast with deviation between MatK and RbcL, which was not significant (Student's t-test, $p = 0.57$). Variation in side chain composition was not significant (Student's t-test, $p=0.052$) when MatK was compared to the mitochondrial maturase Mat-r for 34 taxa (Table 3.4). Relative composition in total protein of each of the four amino acid categories followed the same pattern for all four proteins with nonpolar amino acids comprising the highest percentage and acidic amino acids the lowest percentage of protein (Figure 3.6).

Oryza sativa

Figure 3.5.—Secondary structure of MatK determined by JPRED. Alpha helices are represented by "H" and β-strands by "E". Box "a" surrounds two α-helices common between MatK and LtrA in the N-terminal region (Blocker et al. 2005). Box "b" encloses two α-helices, the second of which is part of the "3a" insertion of LtrA (Blocker et al. 2005). Amino acids composing domain X as indicated by Blastp analysis are underlined and the conserved α-helices, αH, αI, αJ of LtrA noted. The "ti" insertion found in LtrA is also indicated.

TABLE 3. 3

SD in side chain composition of three proteins

when compared across 22 taxa.

Amino acid catagory/ Protein	MatK	RbcL	InfA
Nonpolar	1.34	0.7	4.8
Uncharged at pH =7	1.5	2.03	4.7
Basic	1.15	0.24	2.6
Acid	1.56	1.49	3.6
Total deviation	1.3875	1.115	3.925

TABLE 3.4.

SD in side chain composition of two maturases

when compared across 34 taxa.

Amino acid catagory/ Protein	MatK	Mat-r
Nonpolar	1.83	0.64
Uncharge at pH =7	2.90	1.56
Basic	1.69	0.72
Acid	0.63	0.42
Total deviation	1.77	0.91

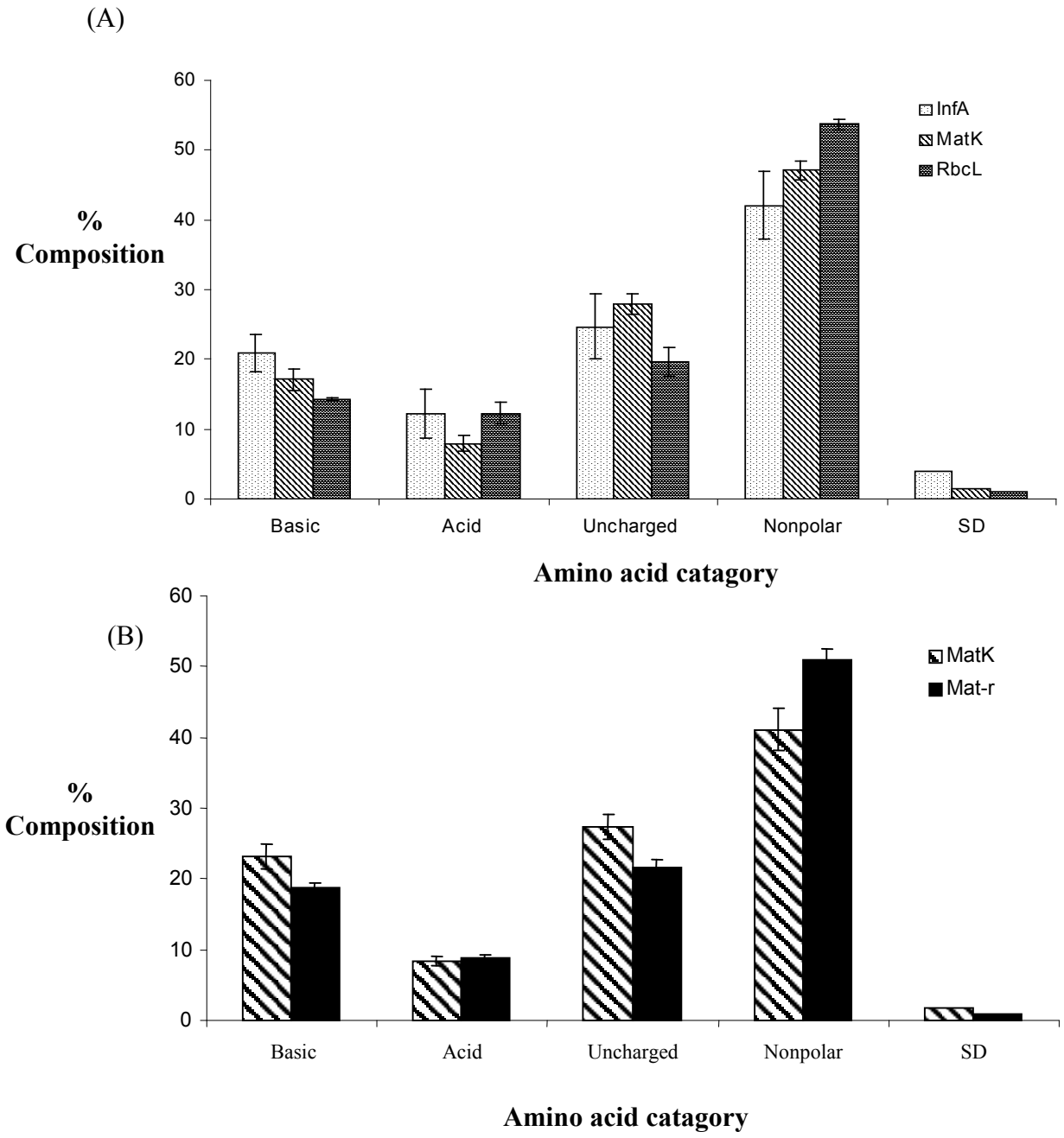


Figure 3.6.--Comparison of side chain composition and SD between putative proteins. (A) Comparison of the average side chain composition and variation among three chloroplast proteins, InfA, MatK and RbcL, for 22 species. (B) Comparison of the average side chain composition and variation between the putative chloroplast maturase MatK and the mitochondrial maturase Mat-r for 34 taxa. Error bars represent standard deviation.

Genetic buffers in MatK

A nucleotide alignment of *matK* ORF was generated for 50 species representing 29 seed plant families. Eleven orchid species from ten different genera were included in the alignment, of which eight genera, nine species, are described as containing *matK* as a pseudogene (Salazar et al. 2003; Gravendeel et al. 2004) or possible pseudogene in the case of *Spathoglottis* (Freudenstein et al. 2004). The alignment identified a consensus sequence, ATG*GAAGAA,* at the beginning of MatK, with ATG as the start codon followed by conserved sequence (noted in italics, Figure 3.7A). Translation from this start codon resulted in a full-length reading frame for all taxa with the exception of the nine orchids mentioned above. Translation from the consensus start of MatK for these nine orchid species resulted in premature stop codons in the MatK reading frame (Figure 3.7B). However, analysis of the upstream nucleotide sequence of MatK for these orchids identified an alternate start codon (Figure 3.8A). This alternate start codon is one base pair out of frame with the consensus start codon for other seed plants (Figure 3.8A), including the two orchid species *Cynorkis* and *Neuwiedia*, in which *matK* was not described as a pseudogene. It is to be noted that this alternate start codon was not identified in the other seed plants examined here, including *Cynorkis* and *Neuwiedia*. Translation from this alternate start codon produced a full-length MatK reading frame with comparable conserved amino acid sequences as translation from the ATG*GAAGAA* position for other seed plants (Figure 3.8B). The alignment of *matK* sequences for the eleven orchid species also revealed a one base-pair insertion 28 base-pairs downstream of the ATG*GAAGAA* start position exclusively in *Cynorkis* and *Neuwiedia* (Figure 3.8A).

The orchid genus *Spathoglottis* has been specifically noted to lack *matK* (Goldman et al. 2001) or possibly have it as a pseudogene (Freudenstein et al. 2004). *matK* sequences generated

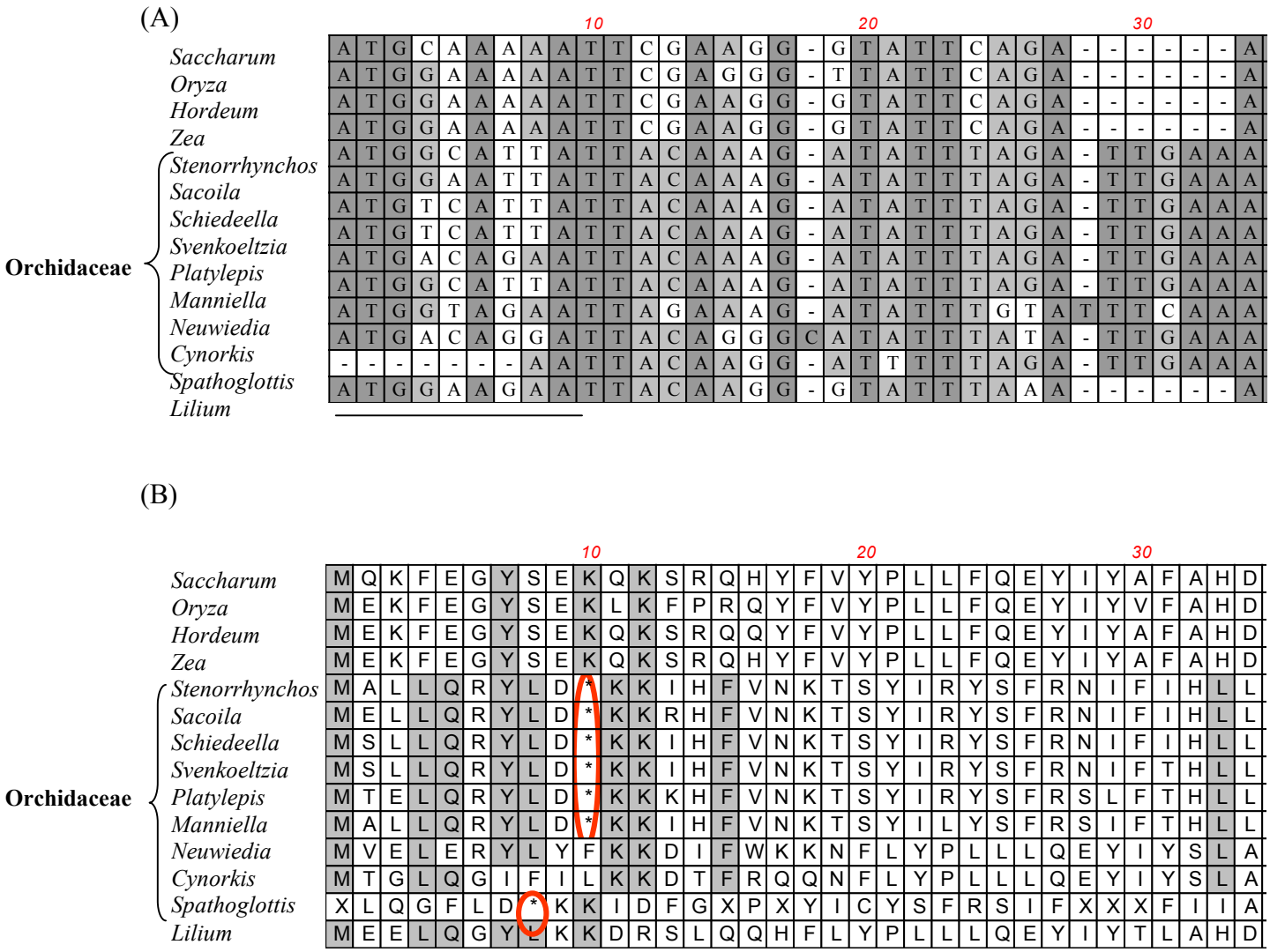


Figure 3.7---Alignment of 15 species of monocots from the families Poaceae, Orchidaceae, and Liliaceae generated with DS Gene©. (A) Nucleotide alignment starting with the consensus start codon for *matK*. Underlined is the ATG*GAAGAA* consensus starting position used by most angiosperm genera. (B) Protein alignment from translation of the nucleotide sequences above using the consensus start codon. Premature stop codons in the Orchidaceae are highlighted in red. Color coding of nucleotides and amino acids are as follows: dark gray = 100% identity; light grey = consensus match; and white = mismatch. Brackets indicate members of the Orchidaceae. Members of the orchid family, Orchidaceae, are bracketed.

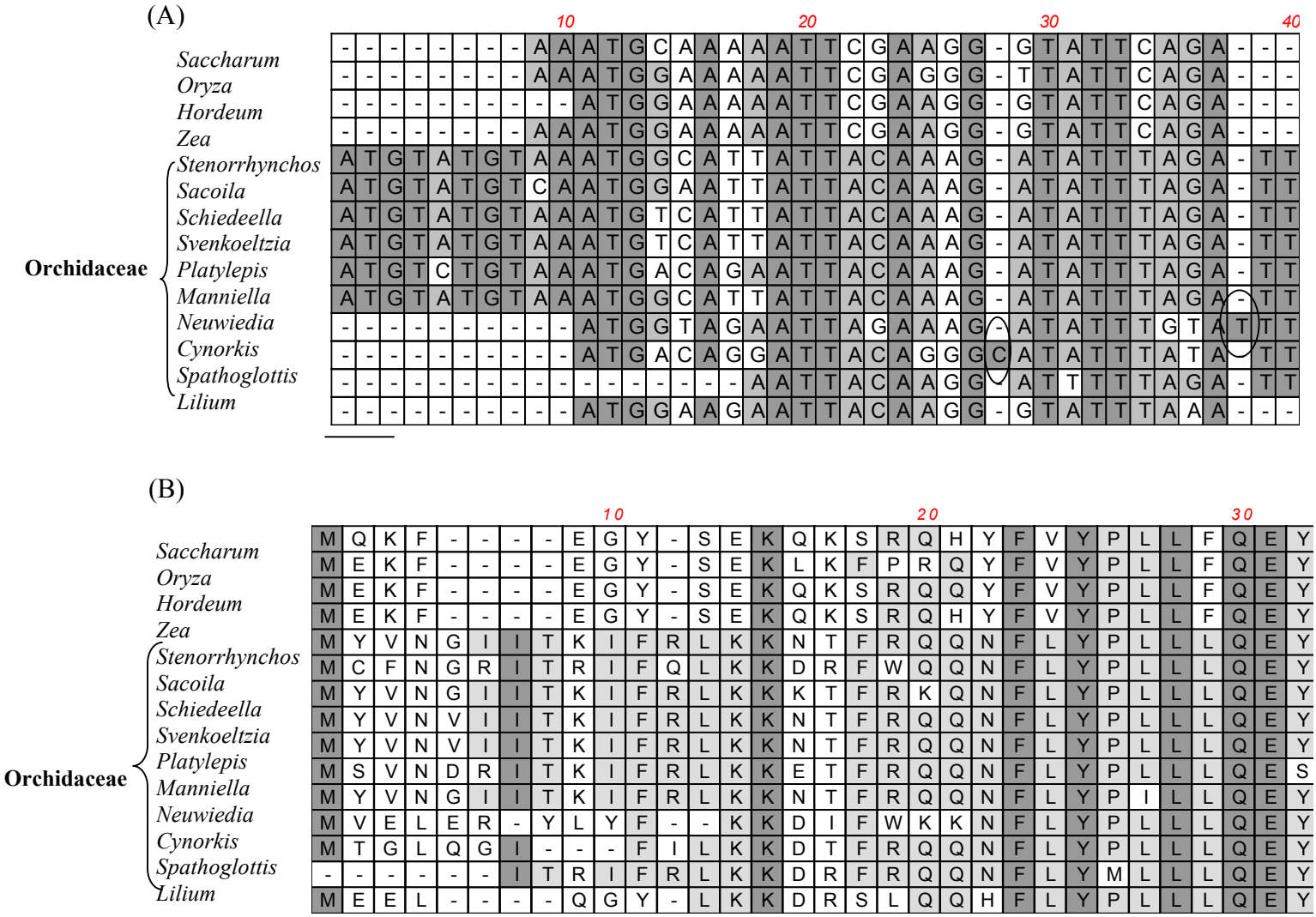


Figure 3.8---Alignment of 15 species of monocots from the families Poaceae, Orchidaceae, and Liliaceae generated with DS Gene©. (A) Nucleotide alignment starting with the alternate start codon for *matK* found in the Orchidaceae. Circled are one-base pair insertions in members of the Orchidaceae that lack the alternate start codon. (B) Protein alignment from translation of the above nucleotide sequences using the alternate start codon. Color coding of nucleotides and amino acids are as follows: dark gray = 100% identity; light grey = consensus match; and white = mismatch. Brackets indicate members of the Orchidaceae. Underlined is the alternate start codon.

in this study for two species of *Spathoglottis* (*S. gracilis* and *S. plicata*) were translated into a full-length reading frame. Translation of *matK* from these two species required the use of a position eight base pairs downstream from the consensus ATG*GAAGAA* (Figure 3.8A). The sequences of both *S. plicata* and *S. gracilis* produced in this study lacked an ATG start codon but instead began the reading frame with isoleucine (AAT) as the first amino acid (the translated sequence of *S. plicata* is shown in Figure 3.8B). This isoleucine starting position is one base pair out of frame with the consensus ATG*GAAGAA,* a situation similar to the alternate start codon.

DISCUSSION

The high rate of nucleotide substitution, the elevated rate of nonsynonymous substitution and its subsequent amino acid changes, and the prevalence of insertion and deletions of various lengths known in *matK* (Olmstead and Palmer 1994; Soltis and Soltis 1998; Hilu et al. 2003) are expected to influence protein structure and function. Structural or amino acid information can be used to address the impact of these sequence features. Since a crystal structure for MatK does not exist, we have focused on analyzing the variability in composition of the four amino acid categories, acid, base, uncharged at pH = 7, and nonpolar, to determine conserved elements that potentially relate to structure and/or function. The *matK* gene has a high rate of nonsynonymous amino acid substitution. However, it has been proposed that for protein coding genes, amino acid substitutions tend to be restrained to specific amino acid categories in which one amino acid is exchanged for another of similar chemical characteristics (Clarke 1970; Graur 1985; Garcia-Maroto et al. 1991; Wolfe and dePamphilis 1998; Xia and Li 1998). Is it likely that this

tendency is followed even in a fast-evolving gene like *matK* that exhibits six fold the amino acid substitution of *rbcL* and *atpB* (Olmstead and Palmer 1994)?

Side chain composition and variation in MatK

Since structure and function are determined by proportion of amino acids and their chemical characteristics (Clarke 1970; Graur 1985; Garcia-Maroto et al. 1991; Wolfe and dePamphilis 1998; Xia and Li 1998; Blocker et al. 2005), we expect that amino acid substitution in MatK to be under specific constraints that should not deviate to a significant degree, maintaining structure and function of this protein. Our analysis of MatK side chain composition for 53 green plant species demonstrated that nonpolar amino acids constitute a major component (47%) of MatK (Figure 3.1A). The remaining groups are divided into 27.6% uncharged (pH =7), 17.2% basic, and 8.1% acidic amino acids. Thus, MatK is approximately 50% polar and 50% hydrophobic, with low overall negative charge. This suggests high hydrophobicity with the potential for membrane association for this protein. The high percent of leucine, a nonpolar amino acid, in MatK further supports the hydrophobicity (Table 3.1).

Hydrophobic regions of a protein tend to either be packed in the internal protein tertiary structure as part of α-helices (Engelman et al. 1980; Blocker et al. 2005) to avoid exposure to the polar environment of the cell, or embedded in a membrane in the form of a transmembrane domain (Engelman et al. 1980; Engelman, Steitz, and Goldman 1986; Engelman et al. 2003; Cuthbertson, Doyle, and Sansom 2005). Further examination of hydrophobic regions in MatK by TMAP (Persson and Argos 1994) and TMHMM transmembrane prediction programs indicated that, although the number of predicted transmembrane domains varied in different taxa, there was at least one putative transmembrane domain found within most predicted MatK protein sequences. This suggests that MatK could be integral membrane protein of the chloroplast. It is

unclear at this point how this localization is related to MatK function. TMAP analysis of the mitochondrial group II intron maturases Mat-r (Wahleithner, MacFarlane, and Wolstenholme 1990; Thomson et al. 1994; Farré and Araya 1999), Cox1-I2 (Mohr, Perlman, and Lambowitz 1993), and Cob I1 (Mohr, Perlman, and Lambowitz 1993) also predicted transmembrane domains in the putative amino acid sequences for these enymes. Although LtrA is considered to be related to the "mitochondrial" lineage of mobile group II intron maturase (Blocker et al. 2005), transmembrane domains were not predicted to be present in the LtrA maturase protein. Instead, the hydrophobic regions of LtrA tend to be packed into alpha helices similar in structural position to those found in HIV-1 RT (Blocker et al. 2005). Since TMAP and TMHMM transmembrane prediction programs calculated that transmembrane domains were present in three mitochondrial group II intron maturase ORFs, it appears that, if these predictions are correct, transmembrane segments are an attribute of mitochondrial group II intron maturases. However, since this characteristic was not predicted to be present in the bacterial maturase LtrA, transmembrane segments in group II intron maturases may not have arisen until the eukaryotic mitochondrial group II intron maturase structure diverged from that of the bacteria LtrA maturase. Domain X, the putative functional domain of MatK (Hilu and Liang 1997), has close homology to domain X of eukaryotic mitochondrial group II intron maturases (Sugita, Shinozaki, and Sugiura 1985; Neuhaus and Link 1987). Thus, the prediction of transmembrane segments and the homology of domain X in MatK to mitochondrial group II intron maturases may reflect a possible mitochondrial origin for this gene.

When deviation in the proportion of each amino acid catagory is considered across green plants, prominent differences were evident. The acidic group, which comprises the lowest proportion, displayed the least deviation (SD =0.89). In contrast, uncharged (pH=7) and

nonpolar amino acid groups have almost twice the standard deviation (1.72 and 1.63, respectively) and the basic amino acids have SD = 1.23. Part of the variability observed within the uncharged (pH=7) and nonpolar amino acids is attributed to pronounced deviation in only one amino acid (lysine for the former and phenalanine for the latter; Table 3.1). The constraints on acidic amino acids are a rather significant finding. The operation of such constraints imply purifying selection acting on the acidic amino acid group (Graur 1985), an overall intolerance of change in negative charge to this protein, and functional constraint (Cheng et al. 2005) on MatK.

Grooves lined with positively charged amino acids are aspects of functional regions in the LtrA group II intron maturase of *Lactococcus lactis* (Blocker et al. 2005). Specifically, a positively charged amino acid track forms the template-primer binding region of LtrA (Blocker et al. 2005). This template-primer binding region spans the majority of the ORF for LtrA and is homologous to the template-primer binding region of HIV-1 reverse transcriptase (Blocker et al. 2005). The high content of basic amino acids found in MatK relative to the low percentage of acidic amino acids (17.2% verses 8.1%, respectively), as well as the lack of variation that implies high constraint in these two specific amino acid categories, suggests that MatK also shares this template-primer track characteristic. Further, at least part of this template-primer binding region was previously shown to lack strongly conserved sequence among group II intron maturases but, nonetheless, be functionally important and identifiable by a conserved cluster of basic amino acids followed by an α-helix (Filippo and Lambowitz 2002). These results from San Filippo and Lambowitz (2002) support our implication of functional constraint in MatK based on conserved side chain composition.

Although the nonpolar amino acid group was the most variable among the four amino acid categories, the hydrophobic amino acids cysteine, methionine, and tryptophan were the most

conserved in MatK (Table 3.1). Two of these amino acids, cysteine and tryptophan, are considered highly immutable through evolution (Graur 1985; Graur and Li 1988), a pattern expected in functional but not pseudogenes. Their low mutability may be attributed to their importance in tertiary protein structure. Replacement of cysteine residues may prevent the formation of disulfide bridges, a characteristic often found in chloroplast proteins (Anderson and Manabe 1979; Gopalan et al. 2004), needed for structure and/or function (Rajarathnam et al. 1999), while substitution for tryptophan, a bulky aromatic amino acid, may distort overall structure (Guo et al. 2004). Therefore, in this respect, MatK follows this expected pattern of amino acid variability, and the low variation in these two critical amino acids implies a high structural and/or functional constraint.

Graur (1985) proposed that the rate of evolution of a protein could be correlated to the amount of certain amino acids in that protein. He further stated that a high content of cysteine, glycine, and tyrosine reflect a slow evolutionary rate in genes. The proportion of these three amino acids in MatK is relatively low (Table 3.1), which reflects the high rate of evolution of *matK*. Glycine is the smallest amino acid and the least mutable as substitution of this amino acid, even to the next smallest amino acid, would increase the size in that position 10 times, greatly distorting tertiary structure (Graur 1985). Although, the standard deviation for glycine (0.7) in MatK was substantially higher than that of cysteine (0.48) and tryptophan (0.35) (Table 3.1), this variation is much lower than the glycine standard deviation (1.43) in the conserved protein RbcL (Kellogg and Juliana 1997; Wolfe and dePamphilis 1998) when comparing the same species. It is possible that the relatively high rate of variation in glycine compared to other amino acids in MatK is compensated for by its presence in low proportion in MatK. Thus,

evolution of MatK protein may proceed at a faster rate without adverse affects on protein structure.

Variation across MatK

Identification of functionally or structurally important regions of a protein coding sequence is generally accomplished by examining areas of high conservation as an indication of evolutionary constraint (Cheng et al. 2005). However, distinguishing between structurally verses functionally important sites is more difficult. Computational and structural studies have indicted that functionally important sites may tend to rely less on actual amino acid sequence, but more on conservative amino acid replacement (Zhu and Karlin 1996; Aloy et al. 2001; Blocker et al. 2005). Based on these findings, conservation in amino acid sequence implies more of a structural constraint whereas conservation of side chain composition suggests more of a functional constraint. Hilu and Liang (1997) and Hilu et al. (2003) indicated that domain X had a lower number of variable sites in nucleotide sequence compared to the N-terminal region and lacked indels, suggesting a structural and possibly a functional importance for domain X. Despite the difference in rate of nucleotide mutation, the two regions displayed remarkably similar variation in side chain composition (Figure 3.2A). Therefore, our results demonstrate that these substitutions are being constrained into specific amino acid categories, maintaining function of MatK. Young and dePamphilis (2000) also found that the N-teminal region and domain X were not significantly different from each other when examining substitution rates of codon position in DNA.

Since the N-terminal domain and domain X of MatK are very large (450 and 155 amino acids, respectively), the MatK reading frame was divided into seven sectors of approximately

equal size to gain a better understanding of variation in side chain composition. Using this division, the N-terminal region was comprised of sectors 1-5, while domain X comprised sectors 6 and 7 (Figure 3.3C). Sectors 3 and 5 contain high level of hydrophobic amino acids compared to the rest of MatK (Figure 3.3A), implying possible hydrophobic pockets in tertiary structure for this protein. These same hydrophobic regions were identified in the 14 taxa data set that lacked indels. The fact that these hydrophobic regions are maintained in MatK regardless of insertion or deletion of amino acid sequence signifies their importance for the overall structure/function of this protein.

Sectors 2 and 5 of the N-terminal region have highly conserved side chain composition (Figure 3.3B). The conserved side chain composition observed for sectors 2 and 5 was not correlated with high conservation in amino acid composition (Figure 3.3B). Following the previous studies concerning separation of functional veses structural constraint on amino acid evolution (Zhu and Karlin 1996; Aloy et al. 2001; Blocker et al. 2005), this pattern implies a more functional and less structural constraint on these sectors. Although Pfam predicted sector 2 to be homologous to DUF877 bacterial protein family, this protein family is of unknown function and thus we cannot conclude at this time about a potential function for this region.

Surprisingly, Pfam did not predict any homology for sector 5 in spite of its low variability in side chain composition. Our sequence homology analysis showed that sector 5 overlaps remnants of the RT domain characteristic of other maturases (Mohr, Perlman, and Lambowitz 1993). Since only remnants of the RT domain is found in MatK (Mohr, Perlman, and Lambowitz 1993), it is surprisingly that the side chain composition of this sector remains highly conserved with low variation (SD = 2.87, Table 3.2). Cui et al. (2004) indicated that the LtrA group II intron maturase of *Lactococcus lactis* requires both domain X and the RT domain for

proper function. Thus, the conserved nature of amino acid replacement in the residual sequence of the RT domain found in MatK may reflect the core of the RT domain needed to accompany domain X for proper maturase function.

As expected based on the lack of indels in domain X (Hilu and Liang 1997) and its putative role in maturase function (Mohr, Perlman, and Lambowitz 1993; Moran et al. 1994), sector 6, which encompasses the core of domain X (Figure 3.3C), displayed very low deviation in both amino acid and side chain composition (Figure 3.3B). This conservation indicates both a structural and functional importance to this domain.

Unlike sector 6, sector 7 displayed high variation in side chain and amino acid composition. Sector 7 corresponds to the last approximately 80 amino acids of *matK*. Although this still encompasses part of domain X, about half of this region lies outside the core of domain X defined by Mohr, Perlman, and Lambowitz (1993) and any other putative domain. The last 200 base pairs of *matK*, which correlates with the region denoted in this study as sector 7, were shown previously to have a high nucleotide substitution rate and indels (Hilu and Liang 1997; Hilu and Alice 1999). Unlike more conserved regions, this sector appears to reflect the same high variation in side chain composition as the high substitution rate seen in the nucleotide sequence. This suggests much less functional or structural importance for this region and highlights the constraints of function placed on more conserved regions of MatK.

Although there is no crystal structure for the MatK protein, we can predict from these data that MatK has two very hydrophobic regions and one to six putative transmembrane segments that could anchor this protein into a membrane. Low variation in amino acid and/or side chain composition for various sectors of the MatK ORF points to the presence of three conserved domains, two in the N-terminal region, one with unknown function (sector 2) and the

RT domain (sector 5), and one which is part of domain X (sector 6). In addition, the conserved side chain composition observed for MatK implies that this protein is under evolutionary constraint in spite of high nucleotide and amino acid substitution rates.

The effect of indels on MatK structure and function

One of the hallmarks of MatK is the high number of insertion and deletion events found across its entire reading frame (Johnson and Soltis 1995; Hilu and Liang 1997). These indels range in size from 1bp in the grasses (Hilu and Alice 1999) to 204 bp in the holoparasitic plant *Epifagus* (Ems et al. 1995; Hilu and Liang 1997), and are generally in multiples of three (Johnson and Soltis 1995; Hilu and Liang 1997; Soltis and Soltis 1998; Hilu and Alice 1999). Indels in most highly conserved genes such as *rbcL,* cause deleterious effects to the protein product and may result in pseudogene formation (Wolfe and dePamphilis 1998; Imsande et al. 2001; Ni, Dong, and Wei 2005). In fact, many null mutants are generated for study by forming an indel with T DNA insertion (Hsieh and Goodman 2005; Lee et al. 2006). However, the pattern of side chain composition and variation in MatK when examining the whole protein, domains, or sectors (Table 3.2), did not change significantly when comparing species lacking indels and with various number and sizes of indels. The only exception to this was a statistically significant difference in SD for sectors 2, 3, and 7. Even with a statistically significant change in variation in these three sectors, the pattern of variation and composition for MatK for all seven sectors did not change. Thus, the indels found within MatK do not appear to affect the overall function of this protein. This implies an unusual tolerance for alteration in amino acid sequence for a functional gene. As expected with the insertion or removal of amino acids, the variation in amino acid composition did change (data not shown). This suggests some structural changes are occurring with indels in

the MatK ORF even though function is being maintained. However, sector 6 remained highly conserved in amino acid variation for both species with and without indels (data not shown). This was not surprising since previous studies have mentioned that domain X lacks indels (Hilu and Liang 1997). Thus, even if indels may be present in the MatK ORF of a species, these indels will not occur in domain X but in other regions of the protein. Regardless, the high amino acid conservation observed in this study for sector 6 reemphasizes both the functional and structural importance of domain X.

MatK in gnetophytes

Gnetophytes are a group of gymnosperms that contain 70 species grouped into the genera *Gnetum, Ephedra,* and *Welwitschia* (Donoghue and Doyle 2000). This group of genera possess features that are shared with flowering plants, such as net-veined leaves in *Gnetum*, flower-like structures, double fertilization (Carmichael and Friedman 1995), and presence of vessels in the wood (Donoghue and Doyle 2000). However, the gnetophytes also have features similar to conifers including tracheids with circular bordered pits, a lack of scalariform pitting in primary xylem, and scale-like and strap-shaped leaves similar to conifers, in *Welwitschia* and *Ephedra* (Donoghue and Doyle 2000). These morphological features, as well as molecular data, have made the phylogenetic position of Gnetales controversial, as they have been regarded as sister to angiosperms, sister to gymnosperms, sister to seed plants, or sister to the pine group within gymnosperms (Bowe, Coat, and dePamphilis 2000; Chaw et al. 2000; Donoghue and Doyle 2000; Magallon and Sanderson 2002). It is rather intriguing to see in Gnetophyta, a significant increase (16%) in hydrophobic amino acid content in MatK compared to gymnosperms, angiosperms, and green plants in general (Figure 3.1B). A corresponding decrease occurred in

the uncharged amino acids, while both basic and acidic amino acid content remained stable. Evaluation of amino acid sequence from members of Gnetophyta demonstrated an increase in the number of predicted transmembrane regions for this plant group. After dividing the MatK reading frame into seven sectors, a transmembrane segment was predicted in sector 3 unique to members of Gnetophyta. This additional putative transmembrane segment possibly represents a unique evolutionary adaptation to MatK in this plant group. Transmembrane regions tend to act as anchors for protein localization. However, other functions have been associated with these regions apart from their purpose of anchoring to a membrane (Kaykas, Worringer, and Sugden 2002). Thus, the additional transmembrane segment predicted in the Gnetophyta may suggest an increased need for membrane association in this plant group or an additional function for this transmembrane region separate from the function of MatK in other plants.

Variation in side chain composition and genetic distance

Examination of nucleotide substitution rates in *matK* sequences at various genetic distances from intrageneric level to across green plants indicated peaks and valleys of deviation in side chain composition. Average variation in MatK side chain composition remained constant from the intrageneric to the intergeneric levels (CV= 0.014 and 0.012, respectively), using *Oryza* and the Poaceae family as our model. As expected, variation augmented as genetic distance increased above these levels (Figure 3.4B). However, average variation in side chain composition of MatK declined by 20% when orders of monocots were compared. Above this interordinal level, evolution of MatK drastically accelerated, with a linear increase in the coefficient of variation that almost doubled at each subsequent level. Therefore, it appears that MatK has gone through a punctuated pattern of evolution in side chain composition. This could indicate conserved

adaptations of structural motifs in the protein during speciation. However, MatK overall side chain composition is constrained at all levels noting that absolute variation is very low although the apparent percent of variation is magnified in Figure 3.4B.

The mutational peaks and valleys are the same for most amino acid categories except for nonpolar amino acids. Uncharged, basic, and acidic amino acids have peaks of variation at the interfamilial level, whereas nonpolar decreased by almost 30% from the previous level (Figure 3.4B). Past this level, all amino acid categories followed the same pattern. Although there is high variation in the amount of nonpolar amino acids (SD = 1.63) compared to acidic and basic amino acids (SD = 0.89, 1.23, respectively) in total MatK protein, this variation is highly constrained over different taxonomic levels (Figure 3.4B). A similar pattern was observed when dicot species (*Arabidopsis*, Brassicales, eurosids II; Hilu et al. 2003) were used as the model (data not shown). The only difference in variation of side chain composition found using this dicot model compared to the monocot model was a decrease in the coefficient of variation for acidic amino acids at the interfamilial level that was not observed when examining *Oryza* (data not shown). The discrepancy in CV for acidic amino acids can be attributed to inadequate sampling for the dicots used in this assessment. Once past the level of interfamilial comparisons, few complete MatK sequences were available for members of other eurosid II orders. The particular sequences selected from orders available may have skewed the analysis.

As indicated earlier, nonpolar amino acids comprise almost 50% of MatK protein. The relatively stable composition of this group of amino acids in MatK observed across all hierarchical levels signifies the importance of this amino acid catagory for MatK structure and function. Nonpolar amino acids appear to form possible hydrophobic pockets (Figure 3.3A) and putative transmembrane domains in MatK. Secondary structure prediction of the LtrA group II

intron maturase from *Lactococcus lactis* and comparison to HIV-1 RT indicated conserved α-helical structure in the N-terminal region and domain X of this maturase and HIV RT (Blocker et al. 2005). These helices are highly hydrophobic. The N-terminal region of these enzymes contains an α-helix that tends to be variable in length but always present (Blocker et al. 2005). This implies a minimum threshold need for nonpolar amino acids in this region to form the α-helix but relaxed constraint on the size of the helix. This form of relaxed constraint is evident in the variability of nonpolar amino acids in MatK. Although there is relatively high variability in the amount of nonpolar amino acids in this protein (SD = 1.63) compared to the other amino acid categories, once a certain level of phylogenetic evolution has occurred, this deviation in nonpolar composition is constrained. This could reflect evolution of a specific length of α-helical structure for particular phylogenetic groups.

JPRED and MatK Secondary Structure

A crystal structure does not exist for any group II intron maturase. However, secondary structure prediction of the LtrA group II intron maturase from *L. lactis* has proven effective for elucidating structural elements (Blocker et al. 2005). One of the main features of structure identified in LtrA were several conserved alpha helices (Blocker et al. 2005). Specifically, domain X of LtrA contained three α-helices that were similar in size and spacing to α-helices αH, αI, and αJ in the thumb region of HIV-1 RT, with the exception of an insertion of 16 amino acids called the "ti" insertion between αH and αI in LtrA (Blocker et al. 2005). JPRED prediction of MatK from *Oryza sativa* (GenBank: NP_039361) identified comparable motifs of three alpha helices similar to those of αH, αI, and αJ in domain X (Figure 3.5). A long stretch of 16 amino acids was also

found between αH and αI of domain X in MatK, analogous to the "ti" insertion of LtrA (Figure 3.5).

The thumb region of HIV-1 RT forms part of the template-primer binding track needed for reverse transcription in this virus and is analogous to domain X of LtrA, which forms part of the nucleic acid-binding track required for RNA splicing (Blocker et al. 2005). Identification of the same three conserved α-helices found in HIV-1 RT and LtrA in domain X of MatK further underscores the evolutionary constraint suggested by our analysis of side chain composition on both function and structure in this chloroplast enzyme and supports that MatK functions as a group II intron maturase. Moreover, the "ti" insertion of LtrA was found to be structurally conserved in some group II intron lineages (Blocker et al. 2005) and to be important for RNA splicing (Cui et al. 2004). Finding this same insertion in MatK reflects the lineage of this enzyme to LtrA and other mitochondrial group II intron maturases, and a conserved mechanism of intron splicing.

Two other regions of structural similarity between MatK and LtrA based on secondary structure are located in the N-terminal region. These two regions include two α-helices at the very beginning (Figure 3.5, box a) and two α-helices in region 3a of LtrA (Figure 3.5, box b). Both of these helical segments were observed for MatK, but were of slightly different lengths (Figure 3.5). These structural differences are very minute and should have minimal impact on overall protein structure. The region denoted as 3a by Blocker et al. (2005) of the LtrA secondary structure is an insertion not found in HIV-1 RT but conserved structurally in other group II intron maturases as well as non-LTR-retrotransposons (Blocker et al. 2005). Region 3a has been hypothesized to contribute to the specificity of binding of the RNA substrate to the maturase (Blocker et al. 2005). This region was also identified through unigenic evolution

analysis to be hypomutable and, thus, an essential element of intron splicing (Cui et al. 2004). It is interesting to note that MatK contains both the "ti" insertion and structural similarity to region 3a, both considered hypomutable elements of RNA splicing (Cui et al. 2004). The presence of the conserved helices and the insertion element 3a in the N-terminal region, as well as the structural conservation of the α-helical regions αH, αI, and αJ and the ti insertion in domain X, strongly support a high degree of evolutionary constraint on MatK structure regardless of the high nucleotide and amino acid substitution rates found in this gene.

Comparison of side chain composition in MatK to InfA, RbcL, and Mat-r

The consistent pattern of side chain composition in MatK across green plants despite the high nucleotide (Johnson and Soltis 1995; Soltis and Soltis 1998) and amino acid substitution rate (Olmstead and Palmer 1994) and the presence of indels (Soltis and Soltis 1998; Whitten, Williams, and Chase 2000; Hilu et al. 2003) clearly demonstrate the evolutionary constraint on MatK. To further support this conclusion, we compared variation in side chain composition in MatK with that of the functionally conserved protein RbcL (Kellogg and Juliana 1997; Wolfe and dePamphilis 1998) and the pseudogene *infA* (Wolfe et al. 1992; Millen et al. 2001). Although the relative side chain composition of each protein (RbcL, InfA, and MatK) was similar, with nonpolar amino acids comprising the highest percentage and acidic amino acids comprising the lowest percentage, the three proteins differ in degree of variation in side chain composition (Figure 3.6). Similarity in side chain composition among the three proteins can be attributed to the TA/CG-deficiency-TG/CT excess rule of coding sequences. This rule states that as long as the nucleotide base composition of the coding sequence is balanced, proteins of very divergent functions may still have very similar amino acid composition (Ohno 1992). However,

total variation of side chain composition in MatK was found to be more similar to that of RbcL than the pseudogene InfA (Table 3.3, Figure 3.6A). Calculation of statistical significance using a Student's T-test showed that the difference in variation in side chain composition was statistically significant between MatK and InfA, but not between MatK and RbcL at the $p < 0.05$ level. The *infA* gene encodes protein translation factor 1, an essential protein for proper chloroplast function. Plants that have a defective copy of *infA* in the chloroplast have had this gene transferred to the nucleus (Millen et al. 2001). The residual copy of *infA* in the chloroplast genome has accumulated random mutations including frame shifts and premature stop codons, conferring its status as a pseudogene (Millen et al. 2001). If the mutations in MatK were not under functional constraint but were random as seen in pseudogenes, then variation in side chain composition should have been more similar to that of InfA than RbcL. The variation in side chain composition observed for InfA was almost three times the variation observed for either RbcL or MatK (Table 3), underscoring the constraint imposed on functional verses nonfunctional proteins.

To see if the level of variation observed in MatK was specific to this protein or common to group II intron maturases, we also compared MatK to another maturase, Mat-r. Mat-r is a mitochondrial group II intron maturase that also contains domain X (Farré and Araya 1999). We found that although the distribution of side chain composition was the same for both proteins (Figure 3.6B), variation of side chain composition differed slightly. Although not statistically significant at the $p = 0.05$ level, the standard deviation in side chain composition of Mat-r was almost half that observed for MatK (Table 3.4). However, RbcL, a functionally conserved chloroplast gene (Kellogg and Juliana 1997; Wolfe and dePamphilis 1998), did not differ significantly in variation of side chain composition from MatK. Thus, this difference in variation

of side chain composition between MatK and Mat-r may not be due to lack of evolutionary constraint on MatK but most likely reflects the difference in rate of evolution of genes in the mitochondria and the chloroplast organelles, as mitochondrial genes are known to evolve at a third of the rate of chloroplast genes (Wolfe, Li, and Sharp 1987; Muse 2000).

Alternate start codon of matK

Several orchid papers have noted that *matK* is a pseudogene in the genera *Stenorrhynchos, Sacoila, Schiedeella, Svenkoeltzia, Platylepis, Spathoglottis,* and *Manniella* (Whitten, Williams, and Chase 2000; Goldman et al. 2001; Kores et al. 2001; Salazar et al. 2003). Initial nucleotide alignment of the putative *matK* open reading frame for 49 angiosperm and gymnosperm taxa identified a consensus sequence, ATG*GAAGAA*, that marked the start codon and first two amino acids of MatK. However, another start codon was identified 10 nucleotides upstream of this consensus sequence in some orchid taxa, specifically those noted to contain *matK* as a pseudogene. Translation from the upstream start codon resulted in a full-length reading frame for MatK that aligned well with the putative amino acid reading frame from other angiosperms (Figure 3.8B). Members of Orchidaceae that have not been noted to contain *matK* as a pseudogene, such as *Cynorkis* and *Neuwiedia*, have a one base-pair insertion 28 base pairs downstream from the consensus ATG*GAAGAA* used by other angiosperms (Figure 3.8A). This one base-pair insertion pushes the reading frame of MatK by one base-pair. Normally, a one base-pair insertion would result in a frame-shift, but in the case of *Cynorkis* and *Neuwiedia*, it aligns the reading frame of MatK correctly for these orchid taxa, suggesting that the out-of-frame start codon in the orchids is actually the correct start codon. Thus, orchid taxa that start with the

consensus ATG*GAAGAA* typical of other angiosperms must have an insertion to keep the MatK ORF in the correct reading frame.

Both in-frame and out-of-frame alternate start codons have been observed for various proteins in mammals and viruses (Dinesh-Kumar and Miller 1993; Otulakowski et al. 2001; Byrd, Zamora, and Lloyd 2002). However, this is the first report for an alternate out-of-frame start codon in plant DNA. In this case, the out-of-frame start codon is the correct beginning of the reading frame and an insertion further downstream in orchids lacking the alternate start codon is required to maintain the correct reading frame.

Members of the genus *Spathoglottis* do not contain either the alternate start codon or the consensus start codon. Instead the MatK reading frame starts with isoleucine as the first amino acid. The start of the reading frame for *Spathoglottis* is also one base-pair out of frame of the ATG*GAAGAA* consensus start position. However, translation of this reading frame produces a full-length MatK amino acid sequence that is similar in sequence to other orchid species (Figure 3.8B).

RNA editing has been identified previously in *matK* sequences from maiden-hair fern (*Adiantum capillus-veneris*), and barley (*Hordeum vulgare*) (Vogel et al. 1997; Wolf, Rowe, and Hasebe 2004). Since the codon used for isoleucine at the beginning of *Spathoglottis matK* sequences is ATT, a single nucleotide change from U to G at the third codon position is all that would be required to convert it to the normal methionine start codon. RNA editing of initiator codons has been identified in the horwort *Anthoceros* for five different genes, *atpB, atpH, petA, cysA,* and *ccsA* (Kugita et al. 2003). However, the nucleotides edited in these particular genes followed the more common form of RNA editing, C to U transition (Schmitz-Linneweber et al. 2001; Kugita et al. 2003; Wolf, Rowe, and Hasebe 2004). Much less common, but still found,

are editing conversions of U to C (Wolf, Rowe, and Hasebe 2004). If editing in the *matK* gene sequence of *Spathoglottis* followed this general pattern, the start codon could change from AUU to AUC. Coincidently, AUC also codes for isoleucine but this particular codon for isoleucine has been shown to initiate translation for dihydrofolate reductase in mammalian cells (Peabody 1989). Even though the AUC codon designates isoleucine as the amino acid to be integrated into the growing polypeptide, when used as the initiator codon in mammalian cells with wild-type initiator tRNA, methionine was still integrated as the first amino acid (Peabody 1989). Furthermore, if RNA editing does not correct this initiator codon in *Spathoglottis*, translation can still occur using the native AUU. Studies of translation for the green alga *Chlamydomonas petD* chloroplast gene found reduced but active translation using the AUU initiator start codon (Chen, Kindle, and Stern 1995). Further studies of the *matK* transcript from *Spathoglottis* orchids would be necessary to examine if RNA editing is taking place to restore this start codon or if isoleucine is being used as an alternate initiator amino acid. Aside from the irregularity of start codon in *Spathoglottis,* we have found no evidence of premature stop codons in members of the Orchidaceae.

Evolutionary constraint on amino acid composition

Evolutionary constraint is generally determined by the rate of nucleotide and amino acid substitution (Garcia-Maroto et al. 1991; Wolfe and dePamphilis 1998; Ophir et al. 1999; Halligan et al. 2004). Synonymous substitutions are regarded as neutral or silent mutations whereas nonsynonymous substitutions are indicative of selective pressures (Ophir et al. 1999; Young and dePamphilis 2005). Therefore, genes that display a high rate of amino acid substitution, such as MatK, would be characterized as to be undergoing positive selection,

whereas genes with a low rate of nonsynonymous substitution, such as RbcL, are characterized to be under purifying selection (Wolfe and dePamphilis 1998; Randle and Wolfe 2005). This model of selection, however, focuses on nucleotide substitutions and their consequent amino acid mutations but does not take into consideration an important component, the biochemical relationships of amino acids. To gain a full understanding of evolutionary constraint in genes and their corresponding protein products, and how selection would actually affect protein function and structure, we need to look beyond sequences and into the physical and chemical properties of these amino acids. Replacement of biochemically similar amino acids for each other may not affect protein structure and function (Wolfe and dePamphilis 1998). This kind of conserved substitution in protein would correspond to silent mutations at the biochemical level. Therefore, three levels of mutation exist, synonymous, nonsynonymous, and biochemical. The pattern of nonsynonymous substitution may impact structure, whereas biochemical mutation in side chain composition may have a greater affect on function (Graur 1985; Zhu and Karlin 1996; Wolfe and dePamphilis 1998; Aloy et al. 2001; Cheng et al. 2005). Further, estimation of overall rate of substitution across the gene is not a reliable measure of the dynamics of structural or functional evolution in the protein. A gene has to be dissected into its fundamental domains to reflect the differential selection pressure that impact these regions. Our analyses of MatK alone, and in comparison with other proteins, have clearly demonstrated that the high rate of nonsynonymous substitution in *matK* (Olmstead and Palmer 1994) is not random but under constraint to maintain structure and function. Analysis of substitution rates in the second and third codon position between *matK* and *rbcL* also support purifying selection for *matK* (Young and dePamphilis 2000). Examination of the side chain composition of MatK and elements of predicted secondary structure have shown very strong evolutionary constraint on overall

structure and function of this protein with very definite features of selection for functional domains verses possibly less important regions of function. Further, in the orchids, the *matK* gene has been suggested to be a pseudogene, implying a loss of functional constraint. We have demonstrated that this gene contains an alternate start codon, alleviating previously suggested premature stop codons in the reading frame. Thus, we have found no evidence suggestive of *matK* forming a pseudogene in the orchids. Fast evolving genes such as *matK* offer a great model of evolutionary change over short periods of time. Further investigation of the chemical properties of proteins from fast evolving genes would enhance our understanding of evolution beyond sequence to the functional relationships between mutation and the effects of these changes on protein structure.

ACKNOWLEDGMENTS

The authors would like to thank NSF Deep Time, Sigma Xi, Virginia Academy of Science, and Virginia Tech for their support of this research. Special thanks to Scott Parker for his help in statistical analysis and Sabrina Majumder for her assistance downloading and sorting GenBank sequences.

Chapter 4

IMMUNOLOGICAL ASSESSMENT OF THE INFLUENCE OF LIGHT AND DEVELOPMENT ON MATK PROTEIN LEVELS

ABSTRACT

Group II intron maturases are essential to post-transcription processing in most organelles. However, detailed studies of maturases have mainly focused on the bacterial maturase LtrA. There are 15 genes in the chloroplast that contain group II introns. These genes encode tRNAs, ribosomal proteins, and at least one photosynthesis-related gene implying a vital requirement for maturase processing in chloroplast function. *matK* is the only chloroplast gene that has been found to encode putative group II intron maturase activity in land plants. Unlike other group II intron maturases, *matK* lacks two of the three domains thought to be involved in intron processing, suggesting a novel mechanism of splicing in the plastid. Only one study has documented a MatK protein of the expected size in any plant. To further investigate MatK protein in a variety of species and elucidate implications of function, we developed a anti-MatK specific antibody. Through immunoblot analyses, we have demonstrated that an immunoreactive protein of the expected size for full-length MatK protein exists in five different plant species that span monocots and eudicots. Further, we have demonstrated that MatK protein levels are influenced by light and developmental stage in rice. These results have suggested a relationship between MatK expression and the regulation of translation for intron-containing substrates.

INTRODUCTION

Only a limited number of studies have investigated expression and function of group II intron maturases. Most of these studies focused primarily on bacterial maturases, specifically that of *Lactococcus* (Matsuura et al. 1997; Saldanha et al. 1999; Wank et al. 1999; Matsuura, Noah, and Lambowitz 2001; Singh et al. 2002; Noah and Lambowitz 2003), yeast mitochondrial maturases (Moran et al. 1994), and a few nuclear-encoded maturases (Jenkins, Khulhanek, and Barkan 1997; Mohr and Lambowitz 2003). MatK has been proposed as the only group II intron maturase encoded in the chloroplast genome of higher plants (Neuhaus and Link 1987). The *matK* gene is approximately 1500 bp in length, nested between the 5' and 3' exons of *trnK* in the large single copy region of the chloroplast genome (Sugita, Shinozaki, and Sugiura 1985).

There are 16 group II introns encoded in 15 genes of the chloroplast that would require a maturase for proper splicing (Vogel et al. 1997). The nuclear-encoded maturase CRS2 is transported to the chloroplast where it processes nine of the ten chloroplast-encoded group IIB introns (Osteimer et al. 2003). This leaves eight group II introns for processing. Studies of the white-barley mutant *albostrians* demonstrated that a chloroplast-encoded maturase was responsible for processing at least five group II intron-containing transcripts (Hess et al. 1994; Vogel et al. 1997; Vogel, Borner, and Hess 1999). Since *matK* is the only gene encoded in the chloroplast genome of higher plants that contains domain X, the putative functional domain of mitochondrial group II intron maturases (Neuhaus and Link 1987), it was proposed that the protein product from this gene excises the group II intron in at least one and possibly all five of the transcripts unspliced in the *albsotrians* mutant (Vogel et al. 1997; Vogel, Borner, and Hess 1999). Immunoblot analysis of protein extract from the *albostrains* mutant using an anti-MatK antibody did not detect MatK protein in extracts of this mutant but did find this protein in green

barley extracts (Vogel, Borner, and Hess 1999), supporting the hypothesis of Vogel et al. (1999) that it may be a lack of MatK in *albostrians* that prevents splicing of some group II intron transcripts. Furthermore, sequence evidence from the residual plastid genome of the holoparasite *Epifagus virginiana* identified the *matK* gene in this genome but not *trnK* (Ems et al. 1995). The retainment of *matK* with the loss of *trnK* strongly supports that the protein from *matK* serves an essential function in the chloroplast of higher plants.

matK is considered to be fast-evolving gene with a nucleotide substitution rate three times higher than that of the large subunit of Rubisco (*rbcL*) and the β subunit of ATP synthase (*atpB*) (Soltis and Soltis 1998), and six-times higher amino acid substitution rate verses that of *rbcL* when comparing tobacco to rice (Olmstead and Palmer 1994). This gene has also been found to contain indels (Ems et al. 1995; Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003) and, in a few species, premature stop codons (Kores et al. 2000; Kugita et al. 2003). These observations have raised questions whether *matK* does produce a protein in most plants, and if so, is this a truncated protein? Although three studies noted finding a MatK protein in plant extracts (du Jardin et al. 1994; Liere and Link 1995; Vogel, Borner, and Hess 1999), only one study that used barley protein extracts identified a MatK protein of the expected size (Vogel, Borner, and Hess 1999). Although the other two studies, (Liere and Link 1995 and du Jardin et al. 1994), noted possible bands for MatK, these bands were at a much lower molecular mass than predicted by amino acid sequence. The discrepancy in size of protein found in these two studies may suggest that a premature stop codon or proteolysis in these species has led to the formation of a truncated, non-functional protein. Due to the essential role that MatK could play in post-transcriptional regulation in the chloroplast as a group II intron maturase, it is important to determine if this gene produces a full-length protein in plants.

To investigate expression and putative function of MatK, we developed an antibody against this protein from rice (*Oryza sativa*) that can be used against extracts from several related species. Previous studies have identified a *matK* transcript in several plant species (Kanno and Hirai 1993; Vogel et al. 1997; Kugita et al. 2003; Nakamura et al. 2003; Wolf, Rowe, and Hasebe 2004; Barthet and Hilu unpublished data). Identification of a MatK protein of the expected size in a range of plant species would complement transcript data and further support the hypothesis that this is a functional protein in the chloroplast. In addition, we examined the influence of light and plant development on MatK expression using Western blot analyses.

One of the proposed substrates for MatK activity is *atpF* (Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999), a subunit of ATP synthase. Although a nuclear maturase has been shown to splice the group II intron of atpF, a second factor from the chloroplast (MatK?) is though to required for complete excision of the intron (Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999). Other subunits of ATP synthase have been shown to be influenced by light (Jiao, Hilaire, and Guikema 2004; Mackenzie, Johnson, and Campbell 2005). This is expected as the protein complex is directly related to the cellular activities of photosynthesis. MatK may, therefore, act as a regulatory element for this and other substrate translation in the chloroplast similar to the regulation observed for *psbA*, a photosynthesis related chloroplast protein in which the splicing of the intron regulates proper translation (Lee and Herrin 2003). Thus, studying the factors that influence expression of MatK may provide insight into regulatory roles of this protein in the chloroplast.

MATERIALS AND METHODS

Plant material

The three grasses *Oryza sativa* (rice), *Avena sativa* (wheat), *Saccharum officinarum* (sugarcane), the orchid *Spathoglottis gracilis,* and the whisk fern *Psilotum nudum* were grown from seed stock at the Virginia Tech Biology greenhouse or obtained as potted plants. Leaf tissue from the orchids *Malleola ligulata* and *Holcoglossum kimbullianum* was generously provided by Dr. David Jarrell from Mary Washington University. *Arabidopsis thaliana* plants were provided by Dr. Eric Beers at Virginia Tech. Leaf material was collected for each plant, placed in Ziploc freezer bags, and stored at –80 ^{o}C. Herbarium vouchers are located in VPI for all specimens examined with the exception of *Malleola* and *Holcoglossum.* For development and etiolation experiments, rice seeds were stored at 4 ^{o}C, planted in vermiculite, and grown in the Virginia Tech Biology greenhouse under uniform light and water conditions. To assess developmental influence on MatK expression, plant leaves were collected at 2, 4, 6, and 8-weeks post-germination, placed in Ziploc freezer bags, frozen in liquid nitrogen, and stored at –80 ^{o}C. Rice seeds were also planted in vermiculate and grown for two-weeks either in the Virginia Tech Biology greenhouse under uniform light conditions or placed in a dark drawer at 24 ^{o}C for the two week duration to test etiolation. After two weeks, a sample of blades of rice grown in the dark were harvested and the remaining rice plants placed in a Percival growth chamber under uniform light conditions for up to twenty-four hours. Tissue samples were collected at 4 and 24 hours after light exposure. Tissues were frozen in liquid nitrogen and stored at –80 ^{o}C.

Production and purification of MatK antibody

An antibody was produced against the 15 amino acid rice MatK peptide sequence CPEEEKEIPKFQNLRS and synthesized by Cocalico Biologics Inc. (Reamstown, PA, USA). In order to increase specificity of this antibody against only MatK and no other maturase that may contain domain X, the peptide sequence used was chosen from the N-terminal region of MatK. The peptide included a natural N-terminal cysteine residue to facilitate conjugation to a KLH carrier. Examination of the amino acid sequence using a Kyte-Doolittle hydropothy plot (Kyte and Doolittle 1982; Pearson and Lipman 1988; Pearson 1990) and DS Gene© determined that this region was 60% hydrophilic. Further examination using Abie Pro version 3.0 (Chang Biosciences) identified this sequence as an antigenic peptide. Antibody production was performed by Cocalico Biologicals Inc. (Reamstown, PA, USA). Synthesized MatK peptide was injected into a rabbit for antibody production. Rabbit pre-bleeds were tested against total protein extract from rice to identify background binding from rabbit sera. Both the first and second bleeds were tested with an Elisa assay for reactivity by Cocalico Biologicals, Inc. (Reamstown, PA, USA).

The antibody was purified using a nitrocellulose absorption protocol (Asim Esen, Virginia Tech, personal communication) by running 1.4 mg of crude protein extract from rice and a protein molecular weight standard (NEB, Ipswitch, MA, USA) on a 7.5% SDS-PAGE gel followed by transfer to 0.22 μm nitrocellulose. Following transfer, the membrane was stained with Ponceau S and a strip with the size standards cut out. This strip was blocked in 5% Carnation non-fat dry milk in phosphate-buffered saline plus Tween-20 (PBS-T) and incubated with a 1:50 dilution of rabbit MatK antisera overnight at 4 ^{o}C. Subsequent, washes in PBS-T were performed and the membrane was incubated with horse-radish peroxidase (HRP)-

conjugated anti-rabbit antibody (1:2000, Cell Signaling Technology, Danvers, MA) for 1 hour at room temperature, washed again in PBS-T, and detected using ECL peroxidase/luminol system (Amersham Biosciences, Piscataway, NJ, USA) or West Pico chemiluminescent detection system (Pierce Biotechnology, Inc., Rockford, IL, USA). A band of the approximate size expected for MatK was observed and used as a measure to cut a correlating horizontal strip of protein from the original nitrocellulose membrane. This left two horizontal strips of membrane containing protein other than MatK. MatK antisera was incubated on these two strips to remove background. Following to each antisera incubation, membranes were washed 4X with PBS-T for 5 minutes each wash, rinsed 2X with dH_2O, and then bound background antibody removed by 0.1 M citric acid, pH 3.0. Membranes were then rinsed 2X with dH_2O and washed 2X with PBS-T before adding the antisera again for another removal of background. This procedure was repeated three times to obtain a clean antibody against MatK.

Protein extraction and immunoblotting

Approximately 200 mg of plant tissue was ground under liquid nitrogen. Five hundred microliters of 1 X Laemmli SDS sample buffer (62.5 mM Tris, pH 6.8, 2% SDS, 10% glycerol, and 5% β-mercaptoethanol) was then added to ground powder and protein samples were boiled at 95 ^{o}C for 15 minutes. Following protein dentaruring, samples were centrifuged 2X at 15,000 Xg for 10 minutes. The supernatant was kept after each spin. The final supernatant was stored at –20 ^{o}C as crude protein extract. Concentration of protein was determined by the BioRad Protien Microassay using BSA as the standard. Fifty or 75 μg of protein extract was fractionated by 7.5% SDS-PAGE, transferred to 0.22 μm nitrocellulose and stained by Ponceau S. The nitrocellulose membranes were blocked with 5% Carnation non-fat dry milk in PBS-T for one

hour at room temperature and incubated overnight at 4 ^{o}C on a nutating mixer with the nitrocellulose-purified rabbit anti-MatK IgG (diluted 1:50 or 1:300 in PBS-T). Following incubation with primary antibody, membranes were rinsed, washed with PBS-T for once for15 minutes, and twice for 5 minutes, followed by incubation with HRP-conjugated anti-rabbit (1:2000, Cell Signaling Technology, Danvers, MA, USA) for 1 hour at room temperature. Membranes were washed with PBS-T in the same manner as before and chemiluminescent signal was detected using ECL peroxidase/luminol system (Amersham Biosciences, Piscataway, NJ, USA) or West Pico chemiluminescent detection system (Pierce Biotechnology, Inc., Rockford, IL, USA).

RNA isolation

All solutions used for RNA isolation were treated with 0.1% DEPC followed by autoclaving to remove residual DEPC and RNases prior to use. Total RNA was isolated according to Altenbach and Howell (1981) by grinding tissue under liquid nitrogen followed by phenol/chloroform/LiCl extraction and ethanol precipitation. Pelleted RNA was dissolved in water and immediately quantified using a Beckman DU 520 UV spectrophotometer (Beckman-Coulter, Fullerton, CA, USA) with fixed absorption spectra of 260 nm and 280 nm, and stored at –80 ^{o}C.

Northern blots

Twenty micrograms of total RNA was separated along with a RNA size marker (Promega, Madison, WI, USA) on a 1% formaldehyde gel (Gerard and Miller 1986). Prior to use, the gel box, support, and combs for the formaldehyde gel were treated with 3% SDS to remove

contaminating nucleases. The RNA was then transferred to a nylon membrane according to Church and Gilbert (1984) with the exclusion of the denaturing and neutralizing steps, which result in RNA degradation. Membranes were hybridized with Digoxygenin (Dig, Roche, Indianapolis, IN, USA) -labeled probes in Church buffer (Church and Gilbert 1984) specific for the 5' exon of *trnK* or *matK*. Detection of hybridized probes was accomplished by incubation with an antibody conjugated with alkaline phosphatase against the Dig label and signal detected using the chemiluminescent substrate CDP-Star (Roche, Indianapolis, IN, USA) following the manufacture's instructions.

Probe synthesis

Two probes specific for the *trnK/matK* region of rice were constructed. The *matK* and *trnK* probes were generated from rice genomic DNA. Leaf material was ground under liquid nitrogen and extracted using CTAB/chloroform/ isoamyl alcohol followed by isopropanol precipitation (Doyle and Doyle 1990). A sequence region of 132 base pairs upstream to the end of the 5' exon of *trnK* was amplified using primers 5extrnKF (5' CCTTTTGGTATCTGAGTG 3') and trnK5exR (5' AGTACTCTACCATGAG 3'). Although this probe consisted of the *trnK* 5' exon, it has been shown previously to effectively hybridize to the 2.6 and 2.9 kb transcripts of *matK* (Barthet and Hilu, unpublished data). Dig-labeled probe synthesis was performed according to supplier instructions (Roche, Indianapolis, IN, USA) using the following PCR conditions: 1) starting cycle of 95 ^{o}C for 3 min., 48 ^{o}C for 3 min., and 72 ^{o}C for 3 min., 2) main cycle program of 95 ^{o}C for 30 sec., 48 ^{o}C for 1.30 min., and 72 ^{o}C for 3 min., with main cycles repeated 50 times, and 3) 72 ^{o}C for 20 min. to complete end extension.

RESULTS

Development of a MatK antibody

The 15 amino acid synthesized rice antigen used to illicit rabbit anti-MatK antibodies was demonstrated to be specific to MatK by Blastp search in GenBank using default settings, Viridiplantae as the organismal group, and an inclusion threshold of 0.005. Examination of Western blots containing rice crude protein extracts incubated with rabbit anti-rice MatK (anti-MatK) immune serum identified three bands unique to this serum that were not found when incubating with pre-immune serum. These bands had a molecular weight of ~55, 30, and 20 kDa (Figure 4.1A). The 55 and 30 kDa bands were observed to be specific to MatK anti-serum in several replicates of this experiment, whereas, the ~20 kDa band was not consistently observed with immunoblotting (data not shown).

MatK protein in land plants

We tested the sensitivity of our rabbit anti-MatK antibody against protein extracts from several different land plants. Although several bands were detected for the fern *Psilotum*, none of these bands were found to be unique to the anti-MatK antibody, but were observed with both pre-immune serum or anti-MatK immune serum (Figure 4.2A). A unique protein band of ~55 kDa was detected when anti-MatK immune serum was incubated with protein blots containing protein extracts from the monocots oat (*Avena sativa*, grasses) and arrowhead (*Saggittaria latifolia*) (Figure 4. 2A). A ~60 kDa band was detected in protein extracts from sugarcane (*Saccharum officinarum*, grasses) using this antibody. In addition, a band of ~55 kDa was observed for the eudicot *Arabidopsis* when incubated with the anti-MatK antibody. Incubation with pre-immune

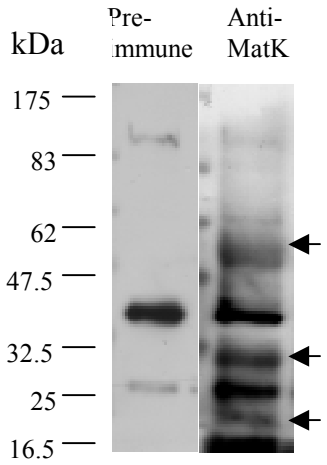
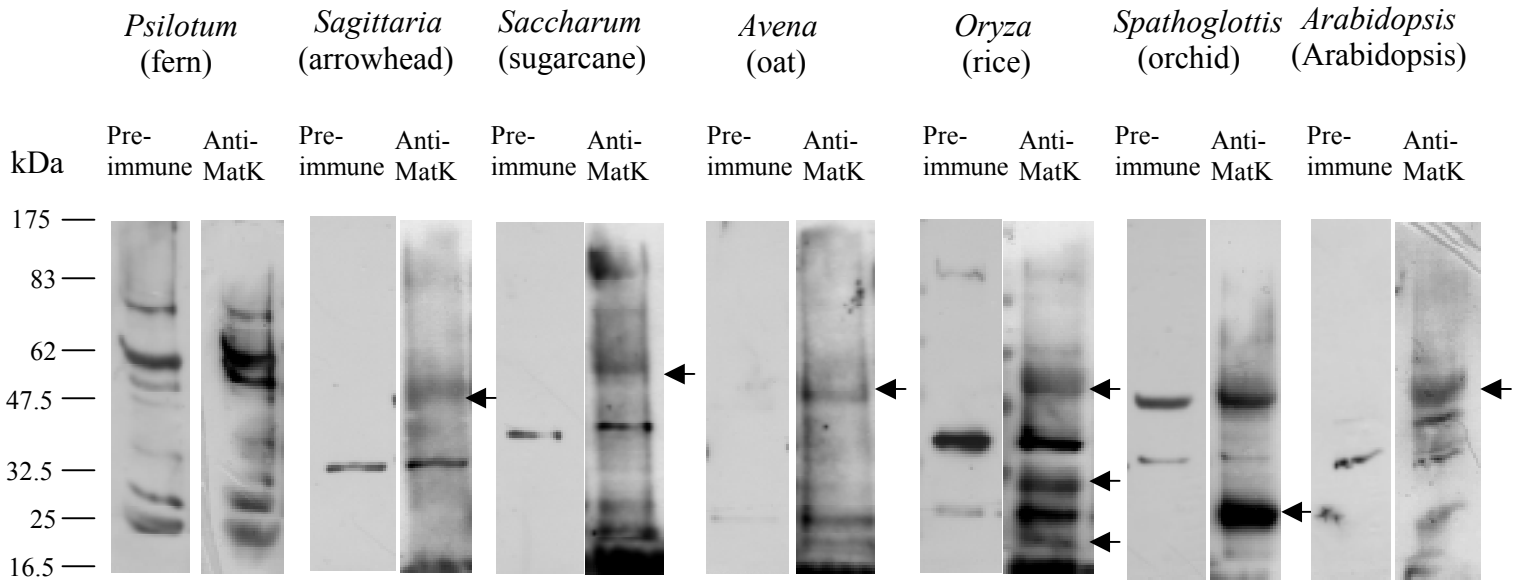


Figure 4.1. Detection of MatK protein in rice. To detect protein bands exclusive to binding to MatK peptide anti-serum, 50 μg of rice protein extract was immunoblotted with pre-immune or immune serum containing antibodies to MatK peptide. Bands found only with anti-MatK are indicated with arrows. Protein standards are noted.

(A)



(B)

arabidopsis (*Arabidopsis*)	S	F	Q	G	K	-	-	Q	L	K	K	S	Y	N	L	Q	S
ground orchid (*Spathoglottis*)	S	L	E	E	K	-	-	R	I	L	K	S	Q	N	L	R	S
Malleola orchid (*Malleola*)	F	L	E	E	K	-	-	R	I	P	K	S	Q	N	L	R	S
pine-needle orchid (*Holcoglossum*)	S	L	E	E	K	-	-	R	I	P	K	S	Q	N	L	R	S
→ rice (*Oryza*)	C	P	E	E	K	-	-	E	I	P	K	F	Q	N	L	R	S
oat (*Avena*)	C	P	K	E	K	-	-	E	I	P	K	F	Q	N	L	R	S
sugarcane (*Saccharum*)	C	P	E	E	K	-	-	E	I	P	K	F	Q	N	L	Q	S
arrowhead (*Alisma*)	-	-	E	E	K	K	K	E	I	P	K	S	Q	N	L	R	S
	-	-	-	-	Q	-	-	Y	Q	S	E	W	N	S	L	Q	S

Figure 4.2. Detection of MatK in other plant species. (A) Fifty micrograms of protein extract was separated by 7.5% SDS-PAGE followed by immunoblotting and incubation with either pre-immune or rabbit anti-MatK serum. Protein molecular weight standards are noted. Arrows indicate protein bands observed only when blots were incubated with the anti-MatK antibody. (B) Amino acid alignment of the MatK peptide antigen region in species examined for MatK protein. An arrow indicates the rice sequence used to generate the MatK antibody for immunoblotting. Dark gray shading = 100% identity, light gray shading = consensus match, white = mismatch.

serum did not detect these same bands for oat, arrowhead, sugarcane, and *Arabidopsis* (Figure 4.2A).

The anti-MatK antibody did not bind to a protein band of the expected size for a full-length MatK (~55 kDa) protein when tested against protein extracts from the orchid *Spathoglottis gracilis*. Instead, a band of approximately 23 kDa was observed on the immunoblots that was not apparent in pre-immune controls (Figure 4.2A). Protein bands beyond those observed when immunoblots with crude protein extract from the orchids *Malleola* and *Holcoglossum* were incubated with pre-immune serum were not detected when blots were incubated with the MatK antibody (data not shown). Homology of the peptide region used to generate the MatK antibody to species examined is shown in Figure 4.2B.

Influence of light and developmental stage on MatK protein

Anti-MatK immune serum incubated with protein extracts from rice tissue bound to a ~55 kDa protein in greenhouse control, dark grown, and etiolated rice samples (Figure 4.3A). Protein levels of this band were decreased by about half in dark-grown (0 hours of light exposure) rice plants (Figure 4.3A). A slight increase in protein content was observed for the ~55 kDa band in rice samples exposed to light for 4 and 24 hours (Figure 4.3A).

Examination of RNA from the same rice plants used for protein analysis indicated that *matK* transcript levels are influenced by light. However, unlike protein levels, which decreased by about half in dark-grown plants, RNA levels of *matK* from these same tissues of dark-grown plants were almost not apparent (Figure 4.3B). Also, once the dark-grown plants were exposed to light, the amount of *matK* RNA almost doubled (Figure 4.3B), while protein only increased slightly (Figure 4.3A).

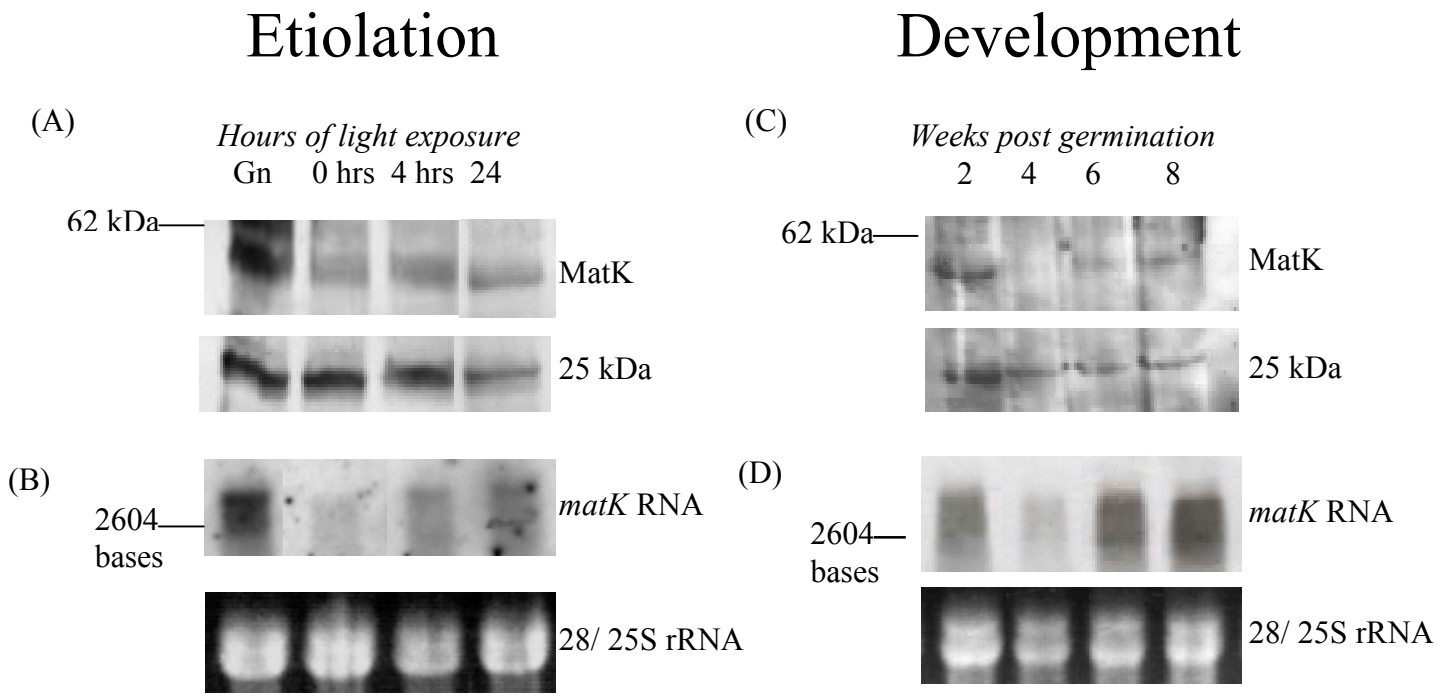


Figure 4.3. The effect of light and developmental stage on *matK* RNA and protein. (A and C) Immunoblot assays using seventy-five micrograms of rice extract separated by 7.5% SDS-PAGE and immunoblotted with rabbit anti-MatK serum. Shown in top panel is the ~55 kDa MatK protein, with a molecular weight band noted, and in bottom panel a 25 kDa background band used as a loading control. Gn = rice plants grown in the greenhouse under uniform light and water conditions as a control. (B and D) Northern blot analysis of (B) etiolated rice RNA followed by light exposure or (D) rice RNA from 2, 4, 6, and 8-weeks post-germination. RNA was separated on a 1% formaldehyde gel, transfered to membrane, and hybridized with a probe specific for the (B) 5' *trnK* exon probe or (D) *matK*. Shown in the top panel is the 2.6 and 2.9 kb *matK* RNA bands, with a size stardard noted, and in bottom panel the 28/ 25S rRNA loading control.

Western blot analysis of protein extracts from rice collected at 2, 4, 6, and 8-weeks post-germination using the MatK anti-serum revealed a pattern of protein levels corresponding to those of *matK* RNA (Figure 4.3 C and D). A protein band of ~55 kDa was identified from immunoblots after incubation with the rabbit-anti-MatK IgG in extracts from 2, 6, and 8 weeks post germination. However, only a very faint band of this size was identified from 4-week post-germination rice extracts (Figure 4.3C).

DISCUSSION

MatK has been proposed as the only group II intron maturase encoded in the chloroplast genome (Neuhaus and Link 1987; Vogel, Borner, and Hess 1999), impling an essential requirement for the expression and function of this protein in the chloroplast. However, identification of MatK protein from plant extracts is greatly lacking. A protein product of the expected size for MatK has only been identified in one previous study from barley extracts (Vogel, Borner, and Hess 1999). Two other studies that noted a MatK protein, found a band very different in size than expected (du Jardin et al. 1994; Liere and Link 1995), leaving some doubt on their findings. Premature stop codons found in the gene sequence for *matK* in the hornwort *Anthoceros formosae* (Kugita et al. 2003) and some orchids (Kores et al. 2000) has led to further doubt if this gene produces a functional protein product. Protein data is needed to confirm that this gene is translated into a protein product in the chloroplast.

Immunoblot protein detection using the MatK antibody

The protein sequence for MatK from rice (GenBank accession: P12175) is 511 amino acids, which correlates to a predicted protein molecular weight of 61.4 kDa. The ~55 kDa protein

band observed on immunoblots of rice crude protein extract incubated with the anti-MatK antibody is in good agreement with this predicted size of MatK. Examination of pre-bleed immunoblots in contrast to immunoblots incubated with the rabbit anti-MatK antiserum verified that the three bands of ~55 and 30, and 20 kDa detected with the anti-MatK antibody were specific to this antibody (Figure 4.1). Therefore, we strongly propose that the ~55 kDa band we have detected from these immunoblot experiments represents the full-length MatK protein product from this plant. It is possible that the lower molecular weight bands identified as specific to anti-MatK are the products of protein proteolysis.

Three studies have previously identified a MatK protein from plant extracts (du Jardin et al. 1994; Liere and Link 1995; Vogel, Borner, and Hess 1999). However, two of these studies reported a full-length MatK protein that was substantially less in molecular weight (~40 kDa , Liere and Link 1995, ~43 kDa, du Jardin et al. 1994) then predicted (~63 kDa, Liere and Link 1995, ~60 kDa, du Jardin et al. 1994). Although there is a slight molecular weight discrepancy between the predicted size of MatK from rice and the actual band detected with our antibody, this difference is only 6 kDa and can be attributed to rough estimation of the size of MatK protein on Western blots based on the relative position of bands for the protein standard. Interestingly, Vogel et al. (1999) predicted that the MatK protein from barley would have a molecular weight of ~56 kDa, very close to the size we have found for the related grass, rice. Nevertheless, the antibody used by Vogel et al. (1999) recognized a MatK protein of ~60 kDa in green barley, a band larger than expected. Thus, a slight discrepancy between the predicted and actual size of MatK is not unusual. Our results are in strong agreement with those of Vogel et al. (1999) and identified an immunoreactive protein in rice of close to the predicted size for full-length MatK.

In contrast to our results and those of Vogel et al. (1999), du Jardin et al. (1994) and Liere and Link (1995) observed a protein suggested to be MatK of approximately 43 kDa. This is almost 20 kDa less then expected. du Jardin et al. (1994) proposed that the discrepancy between the size of MatK observed in their studies and the predicted size is due to hydrophobic tendencies of this protein resulting in a faster-migrating polypeptide. However, a polypeptide of similar size to that found for Liere and Link (1995) and du Jardin et al. (1994) was never detected exclusively with our rabbit anti-MatK serum in any of our Western blot analyses. The only band we detected that is similar to their ~43 kDa band is a 40 kDa band that was found with both the anti-MatK antibody and rabbit pre-immune sera. Therefore, this is a background band and cannot be considered indicative of MatK.

MatK protein in land plants

The rice MatK antibody used in this study was generated against a conserved peptide sequence in MatK. Comparison of this peptide sequence to the same region in other plant genera found almost 100 % homology in closely related grass genera to rice, such as oat and sugarcane, 80% conserved in the basal monocot lineage of Alismataceae (Hilu et al. 2003) that includes arrowhead, and 67 and 73% similarity to the same region in orchids (Figure 4.2B). However, this rice peptide sequence only has 40% identity to the corresponding region of the eudicot *Arabidopsis* (Figure 4.2B). In order to confirm that bands observed using this antibody on Western blots of various plant extracts were specific to the immunogen and not background, pre-immune controls and a negative control were included in all assays. Vogel et al. (1999) included extract from the white barley mutant *albostrians* as their negative control. This mutant lacks chloroplast ribosomal activity and, therefore, was not expected to produce any chloroplast

proteins including MatK. However, in place of the *albostrians* mutant, we used protein extracts from the whisk fern *Psilotum nudum.* Although this fern may express MatK protein, the peptide sequence used to produce our MatK antibody contains a deletion of four amino acids in this fern compared to the rice sequence (Figure 4. 2B). The remaining 11 amino acid sequence in this region of MatK from this fern has only 13% identity to the sequence found in rice. Thus, binding of our antibody to protein from this fern would be the same as any antibody to any random protein in the plant and would suggest a lack of specificity to MatK protein. As predicted, only background bands from rabbit pre-immune serum were observed when crude protein extract from the whisk fern *Psilotum* was incubated with our rabbit anti-MatK antibody (Figure 4.2A).

Examining protein extracts from oat (*Avena*), and sugarcane (*Saccharum officinarum*), close relatives to rice, we identified a band that was unique to our anti-MatK antibody. A similar band was found in extracts from the basal monocot arrowhead (*Sagittaria latifolia*) (Figure 4.2A). The MatK protein is predicted to have a molecular weight of 61 kDa in oat, 63.7 kDa in sugarcane, and 60 kDa in arrowhead based on amino acid sequence. In sugarcane, as in rice, there are several amino acids listed before the actual start codon of MatK that were included in the GenBank protein sequence. Removal of these extra amino acids reduced the expected size of MatK to 62 kDa in sugarcane. Western blot analysis with our rabbit anti-MatK antibody detected a band of ~55 kDa in protein extracts from oat and arrowhead, and ~60 kDa from sugarcane. These bands were not found with pre-immune serum controls, indicating that the bands are specific products of anti-MatK binding. The molecular mass of these bands correspond well to the predicted MatK molecular weights for these species, supporting that these bands represent full-length MatK protein in these species. Although the putative MatK amino

acid sequence for *Arabidopsis* has very low homology to the rice peptide sequence used to generate our antibody, a band of ~55 kDa was observed with anti-MatK serum but not with pre-immune serum. This band again corresponds well with the predicted size of MatK, 60 kDa, from *Arabidopsis*.

We examined three orchid species for MatK protein because of reports which have suggested that *matK* is a pseudogene in many members of the orchid family (Kores et al. 2000; Whitten, Williams, and Chase 2000; Goldman et al. 2001; Salazar et al. 2003). *MatK* is considered a pseudogene in these species due to the presence of indels, lower substitution at the third codon position relative to the first two, premature stop codons, and high transition/ transversion ratios (Kores et al. 2000; Whitten, Williams, and Chase 2000; Goldman et al. 2001; Salazar et al. 2003). Western blot analysis of protein extracts from the three orchid species, *Spathoglottis gracilis, Malleola ligulata* and *Holcoglossum kimbullianum*, using our anti-MatK antibody did not detect any bands specific for MatK protein with the exception of one band of ~23 kDa in *S. gracilis*. This does not necessarily suggest that MatK protein is not produced in these orchids. Although the homology of our antibody to the same peptide region in these orchids was higher than that for *Arabidopsis*, several factors may have hindered detection of a band in these plants. For one, crude protein extracts from *Malleola* and *Holcoglossum* repeatedly contained high polysaccharides, preventing accurate loading onto SDS-PAGE gels. Thus, the amount of extract loaded for these two orchids on gels may have been too low to be detected with our antibody. It is also possible that the MatK protein in orchids may be more rapidly degraded than in other plants thereby preventing detection. In support of this second hypothesis, a small band of ~23 kDa was strongly recognized with our antibody in extract of *S. gracilis* (Figure 4.2A). This same band was not observed in pre-immune controls. The only way

to positively verify the hypothesis that MatK is expressed in orchids is to develop an antibody specific for this group of plants. However, the identification of immunoreactive protein highly probable to be MatK in several other plant species, both monocots and dicots, from this study strongly supports that this protein is made in the orchids as well. Further, a *matK* transcript has been found in the orchid *Spathoglottis plicata* (Barthet and Hilu, unpublished data), an orchid species noted to lack the *matK* gene (Goldman et al. 2001). In addition, an alternate start codon has been identified that could alleviate premature stop codons in several orchid taxa (Barthet and Hilu, unpublished data), supporting the possible existence of MatK protein in orchids.

The influence of light on MatK protein

Light exposure has been shown to increase expression for several genes in the chloroplast, specifically those involved with the photosynthetic apparatus and the formation of mature chloroplasts (Klein and Mullet 1990; Lodish et al. 2000; Jiao, Hilaire, and Guikema 2004). Although protoplasts exists in dark-adapted cells, the formation of the thylakoid, synthesis of chlorophyll and proteins of the stroma, require light induction (Lodish et al. 2000). This suggests a great increase in chloroplast protein translation with an increase in light exposure.

Transcripts of 2.6 and 2.9 kb indicative of *matK* (Barthet and Hilu, unpublished data) were observed in all rice samples tested. The level of *matK* RNA increased when rice plants, which had been kept in the dark for two weeks, were exposed to light (Figure 4.3B). To see if MatK protein levels corresponded to findings with RNA, the same tissues were examined for both *matK* RNA and protein. Similar to the results of *matK* RNA, the ~55 kDa immunoreactive protein we propose to be MatK had protein levels higher in greenhouse-grown rice plants compared to rice plants grown in the dark (Figure 4.3A and B). However, a substantial increase

in protein levels was not observed for plants exposed to light after being grown in the dark for two weeks (Figure 4.3A). This result is inconsistent with observations of RNA levels for *matK* in which there was a very prominent induction by light exposure (Figure 4.3B). The differential control of *matK* expression or stability at RNA and protein levels may be due to the required activity of MatK for substrate regulation at each of these levels.

Studies of intron splicing in *psbA* have indicated that it was the splicing of the intron in this RNA transcript that regulated translation of PsbA protein in the algae *Chlamydomonas reinhardtii* (Lee and Herrin 2003). MatK may act in a similar manner, if it does function as a group II intron maturase as proposed (Neuhaus and Link 1987; Liere and Link 1995; Vogel et al. 1997), to regulate the expression of its intron-containing substrates. In other words, the expression of MatK, would regulate the splicing of group II introns out of transcripts that need to be translated into protein. Potential substrates for MatK maturase activity include introns found within three tRNA genes, *trnK, trnA, trnI,* two ribosomal protein genes, *rpl2* and *rps12,* and one gene related to photosynthesis, *atpF* (Hess et al. 1994; Ems et al. 1995; Jenkins, Khulhanek, and Barkan 1997; Vogel, Borner, and Hess 1999). Although only *atpF* is directly related to photosynthesis via its role as a subunit in the formation of ATP synthase (Knauf and Hachtel 2002), the other potential substrates are required for the chloroplast translation apparatus and, therefore, would be required in greater levels during chloroplast maturation when an increase in chloroplast protein is occurring. Thus, it was expected that MatK levels would increase with light exposure after etiolation. *matK* RNA levels do appear to correspond to the putative function of this protein as a group II intron maturase with these potential substrates, showing a decrease in the dark followed by a large increase with light induction (Figure 4.3B). Although a decrease in MatK protein levels was observed when greenhouse control plants were contrasted

with those grown in the dark, the fact that light exposure did not significantly enhance those levels even after 24 hours was not expected (Figure 4.3A). Instead, MatK levels appeared to stay the same at 0 and 4 hours of light exposure, and only increase slightly after 24 hours of light exposure. It is apparent from this data that *matK* RNA and protein levels are not being regulated to the same extent by light.

Even as a protoplast, some protein translation would have to occur in this organelle, which would require properly spliced tRNAs and ribosomal proteins. Therefore, a continuous requirement for MatK may exist at all times in the chloroplast. This requirement, however, may increase during chloroplast maturation and active photosynthesis. Thus, the main form of regulation of MatK expression may be at the RNA level, but a persistent requirement for this protein overwhelms some of this control and mandates that at least a minimal amount of MatK protein needed for chloroplast function is available at all times.

Developmental regulation of MatK

Both light and development of a plant can regulate the expression and stability of certain genes in the chloroplast. Examination of *matK* RNA levels from 2, 4, 6, and 8-weeks post-germination in rice demonstrated a decrease in RNA at the 4 week point suggesting developmental control of this transcript (Figure 4.3D). A similar pattern was observed at the protein level for the ~55 kDa putative MatK protein (Figure 4.3C and D). This is unlike the differential control of *matK* RNA and protein expression with light, in which RNA levels did not correspond with protein (Figure 4.3A and B). The pattern observed for *matK* RNA and protein in development supports controlled regulation of MatK translation or protein stability in response to a developmental

factor occurring at 4 weeks post-germination in rice. This raises the question of what is this factor?

Between 3-4 weeks post-germination, the panicle of the rice plant is evident and flowering begins (Miller 1994). Thus, the plant as a whole is changing its developmental stages from the vegetative state to the reproductive one. This change in stage of development may temporarily reduce the requirement for certain enzymes and proteins in the chloroplast, such as the substrates for MatK. This would reduce the need for MatK expression in the plastid.

Changes in gene expression have been shown to correlate with panicle development in rice such as in the case of the shikamate pathway (Kasai et al. 2005). Gene expression for this pathway is increased at this stage of development (Kasai et al. 2005). Converse to the shikamate pathway, expression of MatK is decreased at this point of rice development. This decrease in the amount of MatK protein available in the chloroplast may be required to down-regulate translation of one of its substrates. Further experimentation is necessary to deduce the factor regulated by MatK activity during plant development.

CONCLUSION

MatK is proposed as the only group II intron maturase encoded in the chloroplast genome of higher plants (Neuhaus and Link 1987). The presence of 16 group II introns in 15 chloroplast genes, at least five of which have been shown to not be spliced by a nuclear-imported maturase (Vogel et al. 1997; Vogel, Borner, and Hess 1999), implies a critical function of this protein in the chloroplast. However, the fast evolution of this gene at both nucleotide (Johnson and Soltis 1995; Soltis and Soltis 1998) and amino acid levels (Olmstead and Palmer 1994; Hilu and Liang

1997; Whitten, Williams, and Chase 2000) suggests a possible loss of function. The presence of indels (Johnson and Soltis 1995; Hilu and Liang 1997; Whitten, Williams, and Chase 2000; Hilu et al. 2003) and premature stop codons (Kores et al. 2000; Kugita et al. 2003) has further led to speculation regarding if this gene is translated into a functional protein. However, the presence of *matK* in the residual chloroplast genome of the holoparasitic plant *Epifagus* (Ems et al. 1995), the occurance of indels mostly in multiples of three conserving the reading frame (Hilu and Liang 1997; Soltis and Soltis 1998; Hilu, Alice, and Liang 1999; Whitten, Williams, and Chase 2000), and the identification of a transcript for this gene from several plant species (Vogel et al. 1997; Wolf, Rowe, and Hasebe 2004) has supported the functionality of MatK. Immunoblot analyses of pre-immune verses anti-MatK binding and the correlation between *matK* RNA levels and those of the ~55 kDa protein band found using the anti-MatK antibody supports that this protein represents full-length MatK in rice. Thus, this study directly demonstrates that *matK* is translated into a full-length protein of the expected size in a variety of plants and that the stability or expression of this protein is affected by light and the developmental stage of the plant. The fact that light and development influence this protein product strongly support a function for MatK in the chloroplast. Although we have speculated on possible substrates that may be regulated by MatK protein activity in response to these factors, further studies will need to address MatK's activity as a group II intron maturase and determine the exact substrates that this maturase processes.

ACKNOWLEDGEMENTS

The authors would like to thank NSF Deep Time, Sigma Xi, Virginia Academy of Science, and Virginia Tech for their support of this research. Special thanks to Dr. David Jarrell of Mary Washington College for providing orchid leaf material and Dr. David Bevan of Virginia Tech for his help designing the peptide antigen. Thanks Sunny Drysdale for all her help in the Western blot analyses.

Chapter 5

CONCLUDING REMARKS

The chloroplast *matK* gene is well known for its rapid rate of substitution and abundance of phylogenetically informative characters (Olmstead and Palmer 1994; Johnson and Soltis 1995; Hilu and Liang 1997; Soltis and Soltis 1998; Hilu et al. 2003). These features have been utilized to produce robust phylogenies for land plants at several taxonomic levels (Johnson and Soltis 1995; Whitten, Williams, and Chase 2000; Hilu et al. 2003). Analysis of *matK* gene structure identified another interesting feature of this chloroplast gene, the presence of a conserved domain called domain X, which contains sequence homology to a domain in mitochondrial group II intron maturases (Neuhaus and Link 1987). This is the only gene encoded in the chloroplast genome of higher plants identified to contain a putative group II intron maturase domain (Neuhaus and Link 1987; Vogel et al. 1997). However, our knowledge of *matK* function is very poor. Although several studies have now noted a *matK* transcript (Kanno and Hirai 1993; Vogel et al. 1997; Kugita et al. 2003; Nakamura et al. 2003; Wolf, Rowe, and Hasebe 2004), only a few studies have examined details of *matK* RNA (Vogel et al. 1997; Nakamura et al. 2003) and protein (du Jardin et al. 1994; Liere and Link 1995; Vogel et al. 1997; Vogel, Borner, and Hess 1999). The most in-depth study by Vogel et al. (1997) regarding *matK* transcription indicated that that this gene is not transcribed independently from the surrounding *trnK* gene. Further, protein data have been limited, and a full-length protein for MatK of the expected size has only been found in one of three protein studies (du Jardin et al. 1994; Liere and Link 1995; Vogel et al. 1997; Vogel, Borner, and Hess 1999). The conflicting data regarding gene expression for *matK* have strengthened the sequence-based hypothesis that this may be a pseudogene (Kores et al. 2000; Whitten, Williams, and Chase 2000; Kugita et al. 2003; Salazar et al. 2003).

Chapters 2 and 4 of this dissertation have provided clear and direct molecular evidence of *matK* gene expression and refute the notion that *matK* is a pseudogene in most plant species.

These studies have demonstrated that *matK* transcription occurs independent of *trnK* and that *matK* RNA and protein are influenced by light and developmental stage in rice plants. The results of these experiments suggest a function for MatK protein related to photosynthesis and plant development. To further support that MatK is translated into protein, an antibody against the MatK protein from rice was developed and used in immunological experiments. The immunoblot data demonstrated that an immunoreactive protein of a molecular mass consistent with full-length MatK protein is present in extracts from five plant species (Chapter 4).

Possible MatK-regulated substrate RNAs that would be translated into proteins involved in photosynthesis and plant development were hypothesized in Chapters 2 and 4. A maturase assay was developed to test MatK activity on three potential chloroplast substrates, *rpl2, trnA,* and *atpF*. These substrates were chosen based on studies of the *albostrians* chloroplast mutant, in which it was demonstrated that a nuclear maturase was not able to process the group II introns in these RNAs (Hess et al. 1994; Vogel, Borner, and Hess 1999). However, sufficient protein needed to complete the assay could not be produced prior to completion of this dissertation. Therefore, further experimentation is required to substantiate these hypotheses and confirm the maturase activity of MatK. The protocol for performing the maturase assay can be found in Appendix B.

Although the molecular information from Chapters 2 and 4 provided experimental evidence of *matK* expression, the question evolutionary constraint still operating on MatK remained unknown. How is it possible for a gene with such a rapid rate of evolution and high number of amino acid substitutions able to maintain structure and function? Since a crystal structure does not exist for any group II intron maturase, a bioinformatics approach was used to examine the amino acid side chain composition and varibility of putative MatK protein

sequences from 122 green plant species (Chapter 3). The information gathered from this analysis provided a schematic of structurally and functionally important regions in MatK, identified putative transmembrane domains, and further supported, by conserved structure to the LtrA group II intron maturase (Blocker et al. 2005), that MatK functions as a group II intron maturase. It is evident through this computation analysis that although the nucleotide and amino acid sequence of MatK may be under the influence of rapid evolution, evolutionary constraint still exists for this protein and confines the degree of mutation into specific amino acid categories to maintain function of this important chloroplast protein.

In conclusion, the studies presented in this dissertation on *matK* gene expression and evolutionary constraint provide evidence of functional and structural conservation in this protein. Further, the computational analysis of MatK protein offers the first framework for a model of protein structural evolution in a fast-evolving gene.

LITERATURE CITED

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. 1994. The Cell. Pp. 56-57, 376-377. The Cell. Garland Publishing, Inc., New York.

Aloy, P., E. Querol, F. X. Aviles, and M. J. E. Sternberg. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. Journal of Molecular Biology 311:395-408.

Altenbach, S. B., and S. H. Howell. 1981. Identification of satellite RNA associated with turnip crinkle virus. Virology 112:25-33.

Anderson, L. E., and K. Manabe. 1979. Disulfide-linked peptides in the chloroplast thylakoid membrane. Biochimica et Biophysica Acta 579:1-9.

Bailey, C. D., T. G. Carr, S. A. Harris, and C. E. Hughes. 2003. Characterization of angiosperm nrDNA polymorphism paralogy, and pseudogenes. Molecular Phylogenetics and Evolution 29:435-455.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Pp. 28-36. Proceedings of the Second International Conference on Intelligen Systems for Molecular Biology. AAAI Press, Menlo Park, CA.

Balakirev, E. S., and F. J. Ayala. 2003. Pseudogenes: Are they "junk" or funcitonal DNA? Annual Review of Genetics 37:123-151.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Research 30:276-280.

Bhattacharya, D., D. Simon, J. Huang, J. J. Cannone, and R. R. Gutell. 2002. The exon context and distribution of Euascomycetes rRNA spliceosomal intron. BMC Evolutionary Biology 3:7.

Blocker, F. J. H., G. Mohr, L. H. Conlan, L. Qi, M. Belfort, and A. M. Lambowitz. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11:14-28.

Bowe, L. M., G. Coat, and C. W. dePamphilis. 2000. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relative are conifers. Proceedings of the National Academy of Science USA 97:4092-4097.

Boyer, S. K., and J. E. Mullet. 1988. Sequence and transcript map of barley chloroplast *psbA* gene. Nucleic Acids Research 16:8184.

Bremer, B., K. Bremer, N. Heidar, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Kallersjo, and E. Barkhordarian. 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Molecular Phylogenetics and Evolution 24:274-301.

Brendel, V., P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, and S. Karlin. 1992. Methods and algorithms for statistical analysis of protein sequences. Proceedings of the National Academy of Science USA 89:2002-2006.

Byrd, M. P., M. Zamora, and R. E. Lloyd. 2002. Generation of multiple isoforms of eukaryotic translation initiation factor 4GI by use of alternate translation initiation codons. Molecular and Cellular Biology 22:4499-4511.

Cameron, K. M. 2005. Leave it to the leaves: A molecular phylogenetic of Malaxideae.

(Epidendroideae, Orchidaceae). American Journal of Botany 92:1025-1032.

Carmichael, J. S., and W. E. Friedman. 1995. Double fertilization in *Gnetum gnemon*: the relationship between the cell cycle and sexual reproduction. The Plant Cell 7:1975-1988.

Chase, M. W., D. E. Soltis, R. G. Olmstead, and e. a. coauthors). 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Annals of the Missouri Botanical Gardens 80:528-580.

Chaw, S. M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proceedings of the National Academy of Science USA 97:4086-4091.

Chen, X., K. L. Kindle, and D. B. Stern. 1995. The initiation codon determines the efficiency but not the site of translation initiation in *Chlamydomonas* chloroplasts. The Plant Cell 7:1295-1305.

Cheng, G., B. Qian, R. Samudrala, and D. Baker. 2005. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Research 33:5861-5867.

Cho, Y., Y.-L. Qiu, P. Kuhlman, and J. D. Palmer. 1998. Explosive invasion of plant mitochondria by a group I intron. Proceedings of the National Academy of Science USA 81:1991-1995.

Church, G. M., and W. Gilbert. 1984. Genome sequening. Proceedings of the National Academy of Science USA 81:1991-1995.

Clarke, B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. Nature 228:159-160.

Coin, L., and R. Durbin. 2004. Improved techniques for the identification of pseudogene. Bioinformatics 20:i94-i100.

Cui, X., M. Matsuura, Q. Wang, H. Ma, and A. M. Lambowitz. 2004. A group II intron-encoded maturase functions preferentially *in cis* and requires both the reverse transcriptase and X domains to promote RNA splicing. Journal of Molecular Biology 340:211-231.

Cuthbertson, J. M., D. A. Doyle, and M. S. P. Sansom. 2005. Transmembrane helix prediction: a comparative evaluation and analysis. Protein Engineering, Design &Selection 18:295-308.

Davis, J. I., D. W. Stevenson, G. Petersen, O. Seberg, L. M. Campbell, J. V. Freudenstein, D. H. Goldman, C. R. Hardy, F. A. Michelangeli, M. P. Simmons, C. D. Specht, F. Vergara-Silva, and M. Gandolfo. 2004. A phylogeny of the monocots, as inferred from *rbcL* and *atpA* sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. Systematic Botany 29:467-510.

Dinesh-Kumar, S. P., and W. A. Miller. 1993. Control of start codon choice on a plant viral RNA encoding overlapping genes. The Plant Cell 5:679-692.

Donoghue, M. J., and J. A. Doyle. 2000. Seed plant phylogeny: Demise of the anthophyte hypothesis? Current Biology 10:R106-R109.

Doyle, J. J., and J. L. Doyle. 1990. Isolation of plant DNA from fresh tissue. Focus 12:13-25.

du Jardin, P. , D. Portetelle, L. Harvengt, M. Dumont, and B. Wathelet. 1994. Expression of intron-encoded maturase-like polypeptides in potato chloroplasts. Current Genetics 25:158-163.

Ems, S. C., C. W. Morden, C. K. Dixon, K. H. Wolfe, C. W. dePamphilis, and J. D. Palmer. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. Plant Molecular Biology 29:721-733.

Engelman, D. M., Y. Chen, C.-N. Chin, A. R. Curran, A. M. Dixon, A. D. Dupuy, A. S. Lee, U. Lehnert, E. E. Matthews, Y. K. Reshetnyak, A. Senes, and J.-L. Popot. 2003. Membrane protein folding: beyond the two stage model. FEBS Letters 555:122-125.

Engelman, D. M., R. Henderson, A. D. McLachlan, and B. A. Wallace. 1980. Path of the polypeptide in bacteriorhodopsin. Proceedings of the National Academy of Science USA 77:2023-2027.

Engelman, D. M., T. A. Steitz, and A. Goldman. 1986. Identifying nonpolar transbilayer helices in amino acid seqences of membrane proteins. Annual Review of Biophysics and Biophysical Chemistry 15:321-353.

Farré, J., and A. Araya. 1999. The *mat-r* open reading fram is transcribed from a non-canonical promoter and contains an internal promoter to co-transcribe exons *nad1e* and *nad5III* in wheat mitochondria. Plant Molecular Biology 40:959-967.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively mis-leading. Systematic Zoology 27:401-410.

Felsenstein, J. 1985. Donfidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Ferat, J. L., and F. Michel. 1993. Group II self-splicing intron in bacteria. . Nature 364:358-361.

Freudenstein, J. V., C. V. D. Berg, D. H. Goldman, P. J. Kores, M. Molvray, and M. W. Chase. 2004. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. . American Journal of Botany 91:149-157.

Garcia-Maroto, F., A. Castangnaro, P. S. d. l. Hoz, C. Maraña, P. Carbonero, and F. García-Olmedo. 1991. Extreme variations in the ratios of non-synonymous to synonymous nucleotide substitution rates in signal peptide evolution. FEBS 287:67-70.

Geese, W. J., and R. B. Waring. 2001. A comprehensive characterization of a group IB intron and its encoded maturase reveals that protein-assisted splicing requires an almost intact intron RNA. Journal of Molecular Biology **308**:609-622.

Gerard, G. F., and K. Miller. 1986. Comparison of glyoxal and formaldehyde gels for sizing rRNAs. Focus 8:5.

Goldman, D. H., J. V. Freudenstein, P. J. Kores, M. Molvray, D. C. Jarrell, W. M. Whitten, K. M. Cameron, R. K. Jansen, and M. W. Chase. 2001. Phylogenetics of Arethuseae (Orchidaceae) based on plastid *matK* and *rbcL* sequences. Systematic Botany 26:670-695.

Gopalan, G., Z. He, Y. Balmer, P. Romano, R. Gupta, A. Heroux, B. B. Buchanan, K. Swaminathan, and S. Luan. 2004. Structural analysis uncovers a role for redox in regulating FKBP13, an immunophilin of the chloroplast thylakoid lumen. Proceedings of the National Academy of Science USA 101:13945-13950.

Graur, D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. Journal of Molecular Evolution 22:53-62.

Graur, D., and W.-H. Li. 1988. Evolution of protein inhibitors of serine proteinases: Positive Darwinian selection or compositional effects? Journal of Molecular Evolution 28:131-135.

Gravendeel, B., M. C. M. Eurlings, C. v. d. Berg, and P. J. Cribb. 2004. Phylogeny of *Pleione* (Orchidaceae) and parentage analysis of its wild hybrids based on plastid and nuclear ribosomal ItS sequences and morphological data. Systematic Botany 29:50-63.

Guo, Q., F. Zhao, S. Y. Guo, and X. Wang. 2004. The tryptophane residues of dimeric arginine kinase: roles of Trp-208 and Trp-218 in active site and conformation stability. Biochimie 86:379-386.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Research 14:273-279.

Hayashi, K., and S. Kawano. 2000. Molecular systematics of *Lilium* and allied genera (Liliaceae): phylogenetic relationships among *Lilium* and related genera based on the *rbcL* and *matK* gene sequence data. Plant Species Biology 15:73-93.

Herrmann, R. G., J. Steppuhn, G. S. Herrmann, and N. Nelson. 1993. The nuclear-encoded polypeptide Cfo-II from spinach is a real, ninth subunit of chloroplast ATP synthase. FEBS 326:192-198.

Hess, W. R., B. Hoch, P. Zeltz, T. Hübschmann, H. Kössel, and T. Börner. 1994. Inefficient *rpl2* splicing in barley mutants with ribosome-deficient plastids. Plant Cell 6:1455-1465.

Hidalgo, O., T. Garnatje, A. Susanna, and J. Mathez. 2004. Phylogeny of Valerianaceae based on *matK* and ITS markers, with reference to *matK* individual polymorphism. . Annals of Botany 93:283-293.

Hilu, K. W., and L. A. Alice. 1999. Evolutionary implications of *matK* indels in Poaceae. American Journal of Botany 86:1735-1741.

Hilu, K. W., L. A. Alice, and H. Liang. 1999. Phylogeny of Poaceae inferred from *matK* sequences. Annals of the Missouri Botanical Gardens 86:835-851.

Hilu, K. W., T. Borsch, K. Muller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, M. P. Powell, L. A. Alice, R. Evans, H. Sauquet, C. Neinhuis, T. A. B. Slotta, G. R. Jens, C. S. Campbell, and L. w. Chatrou. 2003. Angiosperm phylogeny based on *matK* sequence information. American Journal of Botany 90:1758-1776.

Hilu, K. W., and H. Liang. 1997. The *mat*K gene: sequence variation and application in plant systematics. American Journal of Botany 84:830-839.

Hsieh, M. H., and H. M. Goodman. 2005. Functional evidence for the involvement of Arabidopsis IspF homolog in the nonmevalonate pathway of plastid isoprenoid biosynthesis. Planta *in press*:1-6.

Imsande, J., J. Pittig, R. G. Palmer, C. Wimmer, and C. Gietl. 2001. Independent spontaneous mitochondrial malate dehydrogenase null mutants in soybean are the results of deletions. Journal of Heredity 92:333-338.

Jan, A., H. Nakamura, H. Handa, H. Ichikawa, H. Matsumoto, and S. Komatsu. 2005. Gibberellin regulates mitochondrial pyruvate dehydrogenase activity in rice. Plant and Cell Physiology *In press*.

Jenkins, B. D., D. J. Khulhanek, and A. Barkan. 1997. Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classess with distinct requirements for splicing factors. The Plant Cell 9:283-296.

Jiao, S., E. Hilaire, and J. A. Guikema. 2004. Identification and differential accumulation of two isoforms of the CF1-beta subunit under hight light stress in *Brassica rapa.* Plant Physiology and Biochemistry 42:883-890.

Johnson, L. A., and D. E. Soltis. 1994. *matK* DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str . Systematic Botany 19:143-156.

Johnson, L. A., and D. E. Soltis. 1995. Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *mat*K sequences. Annals of the Missouri Botanical Gardens 82:149-175.

Kanno, A., and A. Hirai. 1993. A transcription map of the chloroplast genome from rice (*Oryza sativa*). Current Genetics 23:166-174.

Kasai, K., T. Kanno, M. Akita, Y. Ikejiri-Kanno, K. Wakasa, and Y. Tozawa. 2005. Identification of three shikimate kinase genes in rice: characterization of their differential expression during panicle development and the enzymatic activities of the encoded proteins. Planta 222:438-447.

Kaykas, A., K. Worringer, and B. Sugden. 2002. LMP-1's transmembrane domains encode multiple funcitons required for LMP-1's efficient signaling. Journal of Virology 76:11551-11560.

Kelchner, S. A. 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. American Journal of Botany 89:1651-1669.

Kellogg, E. A., and N. D. Juliana. 1997. The structure and function of RuBisCO and their implications for systematic studies. American Journal of Botany 84:413-428.

Kiefer, E., W. Heller, and D. Ernst. 2000. A simple and efficient protocol for isolation of functional RNA form plant tissue rich in secondary metabolites. Plant Molecular Biology 18:33-39.

Klein, R. R., and J. E. Mullet. 1990. Light-induced transcription of chloroplast genes. The Journal of Biological Chemistry 265:1895-1902.

Knauf, U., and W. Hachtel. 2002. The genes encoding subuntis of ATP synthase are conserved in teh reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii*. Molecular Genetics and Genomics 267:492-497.

Kohchi, T., K. Umesono, Y. Ogura, Y. Komine, K. Nakahigashi, T. Komano, Y. Yamada, H. Ozeki, and K. Ohyama. 1988. A nicked group II intron and *trans*-splicing in liverwort, *Marchantia polymporpha.* Nucleic Acids Research 16:10025-10036.

Komine, Y., E. Kikis, G. Schuster, and D. Stern. 2002. Evidence for *in vivo* modulation of chloroplast RNA stability by 3' -UTR homopolymeric tails in *Chlamydomonas reinhardtii*. Proceedings of the National Academy of Science USA 99:4085-4090.

Komine, Y., L. Kwong, M. C. Anguera, G. Schuster, and D. B. Stern. 2000. Polyadenylation of three classes of chloroplast RNA in *Chlamydomonas reinhardtii*. RNA 6:598-607.

Kores, P. J., M. Molvray, P. H. Weston, S. D. Hopper, A. P. Brown, K. M. Cameron, and M. W. Chase. 2001. A phylogenetic analysis of Diurideae (Orchidaceae) based on plastid DNA sequence data. American Journal of Botany 88:1903-1914.

Kores, P. J., P. H. Weston, M. Molvray, and M. W. Chase. 2000. Phylogenetic relationships within the Diurideae (Orchidaceae): Inferences from plastid *mat*K DNA sequences. Pp. 449-455 *in* K. L. Wilson, and D. A. Morrison, eds. Monocots: Systematics and Evolution. CSIRO Publishin, Collingwood, Victoria Australia.

Kostrzewa, M., and K. Zetsche. 1993. Organizaiton of plastid-encoded ATPase genes and flanking regions including homologues of *infB* and *tsf* in the thermophilic red alga *Galdieria sulphuraria*. Plant Molecular Biology 23:67-76.

Kudla, J., R. Hayes, and W. Gruissem. 1996. Polyadenylation acceserates degradation of chloroplast mRNA. Embo Journal 15:7137-7146.

Kugita, M., A. Kaneko, Y. Yamamoto, Y. Takeya, T. Matsumoto, and K. Yoshinaga. 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the land plants. Nucleic Acids Research 31:716-721.

Kumar, S., K. Tamura, and M. Nei. 2004. *Mega3*: Integrated software for molecular evolutionary genetics analysis and sequence alignement. Briefings in Bioinformatics 5:150-163.

Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157:105-132.

Lee, J., and D. L. Herrin. 2003. Mutagenesis of a light-regulated *psbA* intron reveals the importance of efficient splicing for photosynthetic growth. Nucleic Acids Research 31:4361-4372.

Lee, C. F., H. Y. Pu, L. C. Wang, R. J. Sayler, C. H. Yeh, and S. J. Wu. 2006. Mutation in a homolog of yeast Vps53p accounts for the heat and osmotic hypersenstive phenotypes in Arabidopsis hit1-1 mutant. Planta *in press*:1-9.

Liere, K., and G. Link. 1995. RNA-binding activity of the *matK* protein encoded by the chloroplast *trnK* intron from mustard (*Sinapis alba* L.). Nucleic Acids Research 23:917-921.

Lisitsky, I., P. Klaff, and G. Schuster. 1996. Addition of destabilizing poly (A)-rich sequences to endonuclease cleavage sites during the degradation of chloroplast mRNA. . Proceedings of the National Academy of Science USA 93:13398-13403.

Lodish, H., A. Berk, L. S. Zipursky, P. Matsudaira, D. Baltimore, and J. E. Darnell. 2000. Molecular Cell Biology. Section: 17.11. Molecular Cell Biology. W. H. Freeman and Company, New York.

Mackenzie, T. D., J. M. Johnson, and D. A. Campbell. 2005. Dynamics of fluxes through photosynthetic complexes in response to changing light and inorganic carbon acclimation in *Synechococcus elongatus*. Photosynthesis Research 85:341-357.

Magallon, S., and M. J. Sanderson. 2002. Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. American Journal of Botany 89:1991-2006.

Maier, R. M., K. Neckermann, G. Igloi, and H. Kossel. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of Molecular Biology 251:614-628.

Matsuura, M., J. W. Noah, and A. M. Lambowitz. 2001. Mechanism of maturase-promoted group II intron splicing. The EMBO Journal 20:7259-7270.

Matsuura, M., R. Saldanha, H. Ma, H. Wank, J. Yang, G. Mohr, S. Cavanagh, G. M. Dunny, M. Belfort, and A. M. Lambowitz. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes & Development 11:2910-2924.

Michel, F., K. Umesono, and H. Ozeki. 1989. Comparative and functional anatomy of group II catalytic introns-a review. Gene 82:5-30.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham. 2000. Vertebrate pseudogenes. FEBS Letters 468:109-114.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe.

2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant Cell 13:645-658.

Miller, T. C. 1994. Mississippi Rice Growers Guide. Pp. 7. Mississippi Rice Growers Guide., Mississippi State University.

Mohr, G., and A. M. Lambowitz. 2003. Putative proteins related to group II intron reverse transcriptase/maturases are encoded by nuclear genes in higher plants. Nucleic Acids Research 31:647-652.

Mohr, G., P. S. Perlman, and A. M. Lambowitz. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Research 21:4991-4997.

Moran, J. V., K. L. Mecklenburg, P. Sass, S. M. Belcher, D. Mahnke, A. Lewin, and P. S. Perlman. 1994. Splicing defective mutants of the *COX1* gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Research 22:2057-2064.

Moran, J. V., S. Zimmerly, R. Eskes, J. C. Kennell, A. M. Lambowitz, R. A. Butow, and P. S. Perlman. 1995. Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. . Molecular and Cellular Biology 15:2828-2838.

Müller, J., and K. Müller. 2003. QuickAlign: A new alignment editor. Plant Molecular Biology Reporter 21:5.

Muse, S. V. 2000. Examining rates and patterns of nucleotide substitution in plants. Plant Molecular Biology 42:25-43.

Nakamura, T., Y. Furuhashi, K. Hasegawa, H. Hashimoto, K. Watanabe, J. Obokata, M. Sugita, and M. Sugiura. 2003. Array-based analysis on tobacco plastid transcripts: preparation of

a genomic microarray containing all genes and all intergenic regions. Plant and Cell Physiology 44:861-867.

Neuhaus, H., and G. Link. 1987. The chloroplast $tRNA^{Lys}$ (UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Current Genetics 11:251-257.

Ni, Z. L., H. Dong, and J. M. Wei. 2005. N-terminal deletion of the gamma subunit affects the stabilization and activity of chloroplast ATP synthase. FEBS Journal 272:1379-1385.

Nickelsen, J., and G. Link. 1991. RNA-protein interactions at transcript 3' ends and evidence for *trnK-psbA* cotranscription in mustard chloroplasts. Molecular and General Genetics 228:89-96.

Noah, J. W., and A. M. Lambowitz. 2003. Effects of maturase binding and Mg2+ concentration of group II intron RNA folding investigated by UV-cross-linking. Biochemistry 42:12466-12480.

Ohno, S. 1992. Universal constraint on evolution of coding sequences. Archives of Gerontology and Geriatrics 14:55-63.

Olmstead, R. G., and J. D. Palmer. 1994. Chloroplast DNA systematics: a review of methods and data analysis. American Journal of Botany 81:1205-1224.

Ophir, R., T. Itoh, D. Graur, and T. Gojobori. 1999. A simple method for estimating the intensity of purifying selection in protein-coding genes. Molecular Biology and Evolution 16:49-53.

Ostheimer, G. J., R. Williams-Carrier, S. Belcher, E. Osborne, J. Gierke, and A. Barkan. 2003. Group II intron splicing factors derived by diversification of an ancient RNA-binding domain. Embo Journal 22:3919-3929.

Ostheimer, G. J., M. Rojas, H. Hadivassiliou, and A. Barkan. 2005. Formation of the CRS2-CAF2 group II intron splicing complex is mediated by a 22 amino acid motif in the C-terminal region of CAF2. Journal of Biological Chemistry In Press.

Otulakowski, G., T. Freywald, Y. Wen, and H. O'Brodovich. 2001. Translation activation and repression by distinct elements within the 5'-UTR of ENaC alpha-subunit mRNA. American Journal of Physiology- Lung Cellular and Molecular Physiology 281:L1219-1231.

Peabody, D. S. 1989. Translation initiation at non-AUG triplets in mammallian cells. Journal of Biological Chemistry 264:5031-5035.

Pearson, W. R. 1990. Rapid and Sensitive Sequence Comparison with FASTP and FASTA. Methods in Enzymology 183:63-98.

Persson, B., and P. Argos. 1994. Prediciton of transmembrane segments in proteins utilizing multiple sequence alignments. Journal of Molecular Biology 237:182-192.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. Proceedings of the National Academy of Science USA 85:2444-2448.

Pryer, K. M., E. Schuettpelz, P. G. Wolf, H. Schneider, A. R. Smith, and R. Cranfill. 2004. Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. American Journal of Botany 91:1582-1598.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid, and nuclear genomes. Nature 402:404-407.

Rajarathnam, K., B. D. Sykes, B. Dewald, M. Baggiolini, and I. Clark-Lewis. 1999. Disulfide bridges in interleukin-8 probed using non-natural disulfide analogues: dissociation of roles in structure from function. Biochemistry 38:7653-7658.

Rambo, R. P., and J. A. Doudna. 2004. Assembly of an active group II intron-maturase complex by protein dimerization. Biochemistry 43:6486-6497.

Randle, C. P., and A. D. Wolfe. 2005. The evolution and expression of *RBCL* in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). American Journal of Botany 92:1575-1585.

Rivier, C., M. Goldschmidt-Clermont, and J.-D. Rochaix. 2001. Identification of an RNA-protein complex involved in chloroplast group II intron *trans*-splicing in *Chlamydomonas reihardtii.* Embo Journal 20:1765-1773.

Rogers, J. H. 1985. The origin and evolution of retroposons. International review of cytology 93:187-279.

Rudi, K., T. Fossheim, and K. S. Jakobsen. 2003. Nested evolution of a tRNA (Leu) (UAA) group I intron by both horizontal intron transfer and recombination of the entire tRNA locus. Journal of Bacteriology 184:666-671.

Salazar, G. A., M. W. Chase, M. A. S. Arenas, and M. Ingrouille. 2003. Phylogenetics of Cranichideae with emphasis on Spiranthinae (Orchidaceae, Orchidoideae): evidence from plastid and nucelar DNA sequences. American Journal of Botany 90:777-795.

Saldanha, R., B. Chen, H. Wank, M. Matsuura, J. Edwards, and A. M. Lambowitz. 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38:9069-9083.

Saldanha, R., G. Mohr, M. Belfort, and A. M. Lambowitz. 1993. Group I and group II introns. FASEB Journal 7:15-24.

Sanders, E. R., K. G. Karol, and R. M. McCourt. 2003. Occurence of *matK* in a *trnK* group II intron in charophyte green algae and phylogeny of the Characeae. American Journal of Botany 90:628-633.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19:101-109.

San Filippo, J. S., and A. M. Lambowitz. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease reigon of a group II intron-encoded protein. Journal of Molecular Biology 324:933-951.

Schmitz-Linneweber, C., M. Tillich, R. G. Herrmann, and R. M. Maier. 2001. Heterologous, splicing-dependent RNA editing in chloroplasts: allotetraploidy provides *trans*-factors. Embo Journal 20:4874-4883.

Schuster, G., and W. Gruissem. 1991. Chloroplast mRNA 3' end processing requires a muclear-encoded RNA-binding protein. Embo Journal 10:1493-1502.

Schuster, G., I. Lisitsky, and P. Klaff. 1999. Polyadenylation and degradation of mRNA in the chloroplast. Plant Physiology 120:937-944.

Shirley, B. W., and I. Hwang. 1995. The interaction trap: in vivo analysis of protein-protein associations. Methods in Cell Biology 49:401-416.

Singh, R. N., R. J. Saldanha, L. M. D'Souza, and A. M. Lambowitz. 2002. Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. Journal of Molecular Biology 318:287-303.

Soltis, D. E., and P. S. Soltis. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. Pp. 2-31 *in* P. S. S. Douglas E. Soltis, and J. J. Doyle., ed. Molecular Systematics of Plants II: DNA Sequencing. Kluwer Academic Publishers, Boston.

Stern, D. B., and W. Gruissem. 1989. Chloroplast mRNA 3' end maturation is biochemically distinct from prokaryotic mRNA processing. Plant Molecular Biology 13:615-625.

Sugita, M., K. Shinozaki, and M. Sugiura. 1985. Tobacco chloroplast tRNALys (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. . Proceedings of the National Academy of Science USA 82:3557-3561.

Swofford, D. L. 2001. PAUP: Phylogenetic analysis using parsimony, ver. 4.0b6. Sinauer Sunderland, Massachusetts, USA.

Thomson, M. C., J. L. Macfarlane, C. T. Beagley, and D. R. Wolstenholme. 1994. RNA editing of mat-r transcripts in maize and soybean increases similarity of the encoded protein to fungal and bryophyte group II intron maturases: evidence that mat-r encodes a functional protein. Nucleic Acids Research 22:5745-5752.

Till, B., C. Schmitz-Linneweber, R. Williams-Carrier, and A. Barkan. 2001. CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. RNA 7:1227-1238.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. Genome Research **13**:2559-2567.

Vanin, E. F. 1985. Processed pseudogenes: characteristics and evolution. Annual Review of Genetics 19:253-272.

Vogel, J., and T. Börner. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. Embo Journal 21:37943803.

Vogel, J., T. Borner, and W. Hess. 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Research 27:3866-3874.

Vogel, J., and W. R. Hess. 2001. Complete 5' and 3' end maturation of group II intron-containing tRNA precursors. RNA 7:285-292.

Vogel, J., T. Hubschmann, T. Borner, and W. R. Hess. 1997. Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for MatK as an essential splicing factor. Journal of Molecular Biology 270:179-187.

Wahleithner, J. A., J. L. MacFarlane, and D. R. Wolstenholme. 1990. A sequence encoding a maturase-related protein in a group II intron of a plant mitochondrial *nad1* gene. Proceedings of the National Academy of Science USA 87:548-552.

Walter, M., J. Kilian, and J. Kudla. 2002. PNPase activity determines the efficiency of mRNA 3'-end processing, the degradation of tRNA and the extent of polyadenylation in chloroplasts. Embo Journal 21:6905-6014.

Wang, X., Y. Xu, Y. Han, S. Bao, J. Du, M. Yuan, Z. Xu, and K. Chong. 2006. Overexpression of *RAN1* in rice and arabidopsis alters primordial meristem, mitotic progress, and sensitivity to auxin. Plant Physiology 140:91-101.

Wank, H., J. San Flippo, R. N. Singh, M. Matsuura, and A. M. Lambowitz. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. Molecular Cell 4:239-250.

Whitten, W. M., N. H. Williams, and M. W. Chase. 2000. Subtribal and generic relationships of Maxillarieae (Orchidaceae) with emphasis on Stanhopeinae: combined molecular evidence. American Journal of Botany 87:1842-1856.

Wolf, P. G., C. A. Rowe, and M. Hasebe. 2004. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339:89-97.

Wolf, P. G., C. A. Rowe, R. B. Sinclair, and M. Hasebe. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. . DNA research 10:59-65.

Wolfe, A. D., and C. W. dePamphilis. 1998. The effect of relaxed funcitonal constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. Molecular Biology and Evolution 15:1243-1258.

Wolfe, K. H., W.-H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Science USA 84:9054-9058.

Wolfe, K. H., C. W. Morden, S. C. Ems, and J. D. Palmer. 1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. Journal of Molecular Evolution 35:304-317.

Xia, X., and W.-H. Li. 1998. What amino acid properties affect protein evolution? Journal of Molecular Evolution 47:557-564.

Young, N. D., and C. W. dePamphilis. 2000. Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants Molecular Biology and Evolution 17:1933-1941.

Young, N. D., and C. W. dePamphilis. 2005. Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. BMC Evolutionary Biology 5:16-25.

Zhu, Z.-Y., and S. Karlin. 1996. Clusters of charged residues in protein three-dimensional structures. Proceedings of the National Academy of Science USA 93:8350-8355.

APPENDIX A

SUPPLEMENTAL TABLES

SUPPLEMENTAL TABLE 1. Nucleotide and amino acid sequences of *matK* used in various analyses. Plant family, species, GenBank accession number, and references are provided. Species are listed in taxonomic order.

Family	Species	Protein GenBank #	Reference
Chaetosphaeridiaceae	[abc]*Chaetosphaeridium globosum*	NP_683782	(Turmel, Otis, and Lemieux 2002)
Funariaceae	[abc]*Physcomitrella patens subspecies patens*	NP_904210	(Sugiura et al. 2003)
Marchantiaceae	[ab]*Marchantia polymorpha*	P12174	(Ohyama et al. 1986)
Anthocerotaceae	[ab]*Anthoceros formosae*	AB086179	(Kugita et al. 2003)
Lycopodiaceae	[abcf]*Huperzia lucida*	AY660566	(Wolf et al. 2005)
Psilotaceae	[abc]*Psilotum nudum*	NP_569609	(Wakasugi et al. Unpublished)
Pteridaceae	[abc]*Adiantum capillus-veneris*	NP848039	(Wolf, Rowe, and Hasebe 2004)
Gnetaceae	[abc]*Gnetum gnemon*	AAS18470	(Won and Renner Unpublished)
Gnetaceae	[b]*Gnetum gnemonoides*	AAS18474	(Won and Renner Unpublished)
Gnetaceae	[b]*Gnetum africanum*	AAS18479	(Won and Renner Unpublished)
Ephedraceae	[b]*Ephedra sarcocarpa*	AAS75352	(Huang, Giannasi, and Price 2005)
Ephedraceae	[b]*Ephedra sinica*	Q8MEX3	(Chaw and Hu 2000)
Welwitschiaceae	[abc]*Welwitschia mirabilis*	AAK69135	(Chaw and Hu Unpublished)
Ginkgoaceae	[abc]*Ginkgo biloba*	AAM46739	(Quinn, Price, and Gadek 2002)
Pinaceae	[b]*Pinus caribaea var. bahamensis*	BAC11944	(Lopez, Kamiya, and Harada Unpublished)
Pinaceae	[bcd]*Pinus koraiensis*	BAD32752	(Gernandt, Geada-Lopez, and Liston Unpublished)
Pinaceae	[ab]*Pinus thunbergii*	NP_042349	(Tsudzuki et al. 1992)
Amborellaceae	[abce]*Amborella trichopoda*	NP_904080	(Goremykin et al. 2003)
Cambombaceae	[ace]*Brasenia schreberi*	AAD05553	(Hilu, Mueller, and Borsch Unpublished)
Nymphaeaceae	[ace]*Nymphea alba*	YP_053136	(Goremykin et al. 2004)
Schisandraceae	[acde]*Illicium floridanum*	AAQ11856	(Hilu, Mueller, and Borsch Unpublished)

Canellaceae	[ace]*Canella winteriana*	AAQ11849	(Hilu, Mueller, and Borsch Unpublished)
Lactoridaceae	[ae]*Lactoris fernandeziana*	AAQ11857	(Hilu, Mueller, and Borsch Unpublished)
Aristolochiaceae	[ae]*Saruma henryi*	AAQ11866	(Hilu, Mueller, and Borsch Unpublished)
Myristicaceae	[cd]*Myristica maingayil*	AAP02935	(Sauquet et al. 2003)
Magnoliaceae	[ae]*Magnolia tripetela*	AAN59945	(Wang, Zhang, and Cui Unpublished)
Magnoliaceae	[ae]*Liriodendron chinense*	AAF43259	(Jin et al. 1999)
Himantandraceae	[ae]*Galbulimima belgraveana*	AAP02924	(Sauquet et al. 2003)
Annonaceae	[ae]*Cananga odorata*	AAP02921	(Sauquet et al. 2003)
Calycanthaceae	[ac]*Calycanthus florida var. glaucus*	AAS09901	(Li et al. Unpublished)
Araceae	[e]*Spirodela intermedia*	Q8WHM7	(Les et al. 2002)
Araceae	[e]*Lemna minuta*	Q8WHM4	(Les et al. 2002)
Araceae	[ae]*Orontium aquaticum*	AAQ11862	(Hilu, Mueller, and Borsch Unpublished)
Tofieldiaceae	[ae]*Pleea tenuifolia*	BAD60953	(Tamura et al. 2004a)
Tofieldiaceae	[e]*Tolfieldia coccinea*	BAD60950	(Tamura et al. 2004a)
Tofieldiaceae	[e]*Isidrogalvia schomburgkiana*	BAD60955	(Tamura et al. 2004a)
Alismataceae	[e]*Caldesia oligococca*	AAX84506	(Li and Zhou Unpublished)
Alismataceae	[e]*Ranalisma rostratum*	AAX84499	(Li and Zhou Unpublished)
Alismataceae	[e]*Alisma canaliculatum*	BAB16787	(Fuse and Tamura 2000)
Juncaginaceae	[ae]*Triglochin maritimum*	BAD20576	(Tamura et al. 2004b)
Cyclanthaceae	[ae]*Carludovica palmata*	BAD20587	(Tamura et al. 2004b)
Arecaceae	[ae]*Nypa fruticans*	AAQ11861	(Hilu, Mueller, and Borsch Unpublished)
Acoraceae	[cd]*Acorus gramineus*	Q9GHG8	(Fuse and Tamura 2000)
Restionaceae	[e]*Restio insignis*	AAT77491	(Hardy and Linder 2005)
Restionaceae	[e]*Rhodocoma foliosa*	AAT77496	(Hardy and Linder Unpublished)
Restionaceae	[e]*Thamnochortus stokoei*	AAV32367	(Hardy and Linder 2005)
Joinvilleaceae	[e]*Joinvillea ascendens*	AAF66167	(Hilu, Alice, and Liang 1999)
Poaceae	[ace]*Oryza sativa*	NP_039361	(Morton and Clegg 1993)
Poaceae	[e]*Oryza australiensis*	AAF37176	(Ge et al. 1999)
Poaceae	[e]*Oryza schlechteri*	AAF37177	(Ge et al. 1999)

Poaceae	[e]*Oryza ridleyi*	AAF37179	(Ge et al. 1999)
Poaceae	[e]*Oryza longiglumis*	AAF37180	(Ge et al. 1999)
Poaceae	[e]*Oryza meyeriana*	AAF37181	(Ge et al. 1999)
Poaceae	[e]*Oryza granulata*	AAF37182	(Ge et al. 1999)
Poaceae	[e]*Oryza malampuzhaensis*	AAO20963	(Ge et al. 2002)
Poaceae	[e]*Oryza eichingeri*	AAO43166	(Bao and Ge Unpublished)
Poaceae	[e]*Oryza nivera*	Q6ENJ6	(Ge et al. 1999)
Poaceae	[e]*Hordeum secalinum*	Q85ZS9	(Nishikawa et al. 2002)
Poaceae	[e]*Hordeum lechleri*	Q85ZU4	(Nishikawa et al. 2002)
Poaceae	[e]*Hordeum jubatum*	BAC54889	(Nishikawa et al. 2002)
Poaceae	[ace]*Hordeum vulgare*	P17158	(Boyer and Mullet 1988)
Poaceae	[e]*Sporobolus indicus*	AAF20357	(Hilu and Alice Unpublished)
Poaceae	[e]*Sporobolus heterolepis*	AAF66216	(Hilu, Alice, and Liang 1999)
Poaceae	[e]*Sporobolus contractus*	AAK60048	(Hilu and Alice 2001)
Poaceae	[ace]*Triticum aestivum*	NP_114240	(Ikeo and Ogihara 2000)
Poaceae	[ae]*Zea mays*	CAA60266	(Maier et al. 1995)
Poaceae	[ae]*Saccharum officinarum*	YP_054609	(Asano et al. 2004)
Liliaceae	[cd]*Lilium columbianum*	BAB08117	(Hayashi and Kawano 2000)
Orchidaceae	[ef]*Platylepis polyadenia*	AJ543946	(Salazar et al. 2003)
Orchidaceae	[ef]*Manniella gustavi*	AJ543944	(Salazar et al. 2003)
Orchidaceae	[ef]*Schiedeella faucisanguinea*	AJ543924	(Salazar et al. 2003)
Iridaceae	[e]*Watsonia angusta*	CAE45261	(Davies Unpublished)
Iridaceae	[e]*Orthrosanthus chimboracensis*	CAE45248	(Davies Unpublished)
Iridaceae	[e]*Pillansia templemannii*	CAE45249	(Davies Unpublished)
Ruscaceae	[e]*Maianthemum dilatatum*	Q9TNB2	(Yamashita and Tamura 2000)
Ruscaceae	[e]*Ruscus aculeatus*	Q9TN87	(Yamashita and Tamura 2000)
Ruscaceae	[e]*Campylandra watanabei*	BAD88433	(Yamashita and Tamura 2004)
Nelumbonaceae	[acde]*Nelumbo nucifera*	AAQ11858	(Hilu, Mueller, and Borsch Unpublished)
Platanaceae	[ae]*Platanus occidentalis*	AAQ11865	(Hilu, Mueller, and Borsch Unpublished)
Buxaceae	[acde]*Buxus sempervirens*	AAQ11846	(Hilu, Mueller, and Borsch Unpublished)
Rhizophoraceae	[ae]*Ceriops tagal*	AAG35557	(Shi et al. Unpublished)

Amaranthaceae	[acde]*Spinacea oleracea*	Q9M13NO	(Schmitz-Linneweber et al. 2001a)
Onagraceae	[acde]*Oenothera elata subspecies hookeria*	Q9MTQ1	(Hupfer et al. Unpublished)
Anacardiaceae	[e]*Pleiogynium timoriense*	Q646L1	(Harrington et al. 2005)
Sapindaceae	[e]*Cupaniopsis anacardioide*	Q646R8	(Harrington et al. 2005)
Sapindaceae	[e]*Aesculus pavia*	Q56B59	(Modliszewski et al. 2005)
Sapindaceae	[e]*Acer monspessulanum*	Q8SEL8	(Bittkau and Mueller-Starck 2002)
Brassicaceae	[ae]*Arabidopsis thaliana*	NP_051040	(Sato et al. 1999)
Brassicaceae	[e]*Arabidopsis lyrata*	AAG43311	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabidopsis halleri*	Q9GF51	(Koch and Mitchell-Olds 1999)
Brassicaceae	[e]*Arabidopsis wallichii*	AAG43336	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabidopsis korshinskyi*	AAG43327	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabidopsis himalaica*	AAG43325	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabidopsis griffithiana*	AAG43314	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabidopsis petraea*	AAG43305	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Cardamine rivularis*	AAG43334	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Cardamine penzesii*	AAG43333	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Cardamine amara*	AAG43306	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabis lignifera*	AAG43313	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabis pumila*	AAG43309	(Koch, Haubold, and Mitchell-Olds 2001)
Brassicaceae	[e]*Arabis procurrens*	AAG43308	(Koch, Haubold, and Mitchell-Olds 2001)
Gyrostemonaceae	[e]*Tersonia cyathiflora*	AAS77373	(Hall, Iltis, and Systsma 2004)
Gyrostemonaceae	[e]*Gyrostemon tepperi*	AAS77372	(Hall, Iltis, and Systsma 2004)
Resedaceae	[e]*Forchhammeria trifoliata*	AAS77380	(Hall, Iltis, and Systsma 2004)
Resedaceae	[e]*Oligomeris linifolia*	AAS77375	(Hall, Iltis, and Systsma 2004)
Malvaceae	[e]*Gossypium gossypioides*	Q95EF8	(Cronn et al. 2003)

Malvaceae	[e]*Kokia drynarioides*	Q95EF3	(Cronn et al. 2002)
Malvaceae	[e]*Hibiscus macrophyllus*	BAD89717	(Takayama et al. 2005)
Malvaceae	[ae]*Hibiscus tiliaceus*	BAD89715	(Takayama et al. 2005)
Malvaceae	[ae]*Heritiera littoralis*	AAQ84263	(Nyffeler et al. Unpublished)
Linaceae	[ae]*Reinwardtia indica*	BAB33400	(Kita and Kato 2001)
Malpighiaceae	[ae]*Dicella nucifera*	AAK71028	(Cameron et al. Unpublished)
Fabaceae	[cd]*Glycine max*	Q9TKS6	(Hu et al. 2000)
Theaceae	[ae]*Franklinia alatamaha*	AAL60378	(Prince and Parks 2001)
Pentaphylacaceae	[ae]*Pentaphylax euryoides*	CAD22187	(Bremer et al. 2002)
Symplocaceae	[ae]*Symplocos hookeri*	AAR13857	(Wang et al. 2004)
Ternstroemiaceae	[ae]*Ternstroemia gymnanthera*	CAD22198	(Bremer et al. 2002)
Araliaceae	[ace]*Panax ginseng*	YP_086947	(Kim and Lee 2004)
Solanaceae	[acde]*Nictotiana tabacum*	NP_054478	(Nadot et al. 1995)
Solanaceae	[cd]*Anthocercis viscosa*	CAE51504	(Clarkson et al. 2004)
Solanaceae	[cd]*Petunia axillaries*	CAE51521	(Clarkson et al. 2004)
Solanaceae	[acde]*Atropa belladonna*	Q8S8Y6	(Schmitz-Linneweber et al. 2002)
Orobanchaceae	[ae]*Epifagus virginiana*	P30071	(Wolfe et al. 1992)

[a]Data set A: species used for assessing variation in MatK across green plants.

[b]Data set B: species used for examining side chain composition in MatK highlighting the Gnetales.

[c]Data set C: the 31 taxa used for analyzing domains and sectors of MatK.

[d] Data set D: the 14 taxa containing only small indels.

[e]Data set E: species used for analysis of MatK variation at different genetic distances.

[f]Only nucleotide sequence available, translated in DS Gene©

SUPPLEMENTAL TABLE 2. Nucleotide and amino acid sequence data set used for analyzing *matK* in the Orchidaceae. Plant family, species, Genbank accession number, and references are provided. Species are listed in taxonomic order.

Family	Species	Protein GenBank #	Reference	Nucleotide GenBank #	Reference
Ginkgoaceae	*Ginkgo biloba*	AAM46739	(Quinn, Price, and Gadek 2002)	AF279806	(Chaw and Hu Unpublished)
Pinaceae	*Pinus caribaea var. bahamensis*	BAC11944	(Lopez, Kamiya, and Harada Unpublished)	AB080941	(Lopez, Kamiya, and Harada Unpublished)
Pinaceae	*Pinus koraiensis*	BAD32752	(Gernandt, Geada-Lopez, and Liston Unpublished)	NC_004677	(Noh et al. Unpublished)
Amborellaceae	*Amborella trichopoda*	NP_904080	(Goremykin et al. 2003)	NC_005086	(Goremykin et al. 2003)
Cambombaceae	*Brasenia schreberi*	AAD05553	(Hilu, Mueller, and Borsch Unpublished)	AF092973	(Les et al. 1999)
Nymphaeaceae	*Nymphea oderata*	AAQ11860	(Hilu, Mueller, and Borsch Unpublished)	AF543742	(Hilu, Mueller, and Borsch Unpublished)
Schisandraceae	*Illicium floridanum*	AAQ11856	(Hilu, Mueller, and Borsch Unpublished)	AF543738	(Hilu, Mueller, and Borsch Unpublished)
Canellaceae	*Canella winteriana*	AAQ11849	(Hilu, Mueller, and Borsch Unpublished)	AF543731	(Hilu, Mueller, and Borsch Unpublished)
Saururaceae	*Saururus cernuus*	AAQ11867	(Hilu, Mueller, and Borsch Unpublished)	AF543749	(Hilu, Mueller, and Borsch Unpublished)
Myristicaceae	*Myristica maingayil*	AAP02935	(Sauquet et al. 2003)	AY220452	(Sauquet et al. 2003)
Magnoliaceae	*Magnolia henryi*	AF209199	(Jin et al. 1999)	AF209199	(Jin et al. 1999)
Magnoliaceae	*Magnolia denudata*	AAF43243	(Jin et al. 1999)	AF123465	(Jin et al. 1999)
Lauraceae	*Umbellularia californica*	CAC05405	(Rohwer 2000)	AJ247190	(Rohwer 2000)
Chloranthaceae	*Chloranthus brachystachys*	AAQ11851	(Hilu, Mueller, and Borsch Unpublished)	AF543733	(Hilu, Mueller, and Borsch Unpublished)
Ceratophyllaceae	*Ceratophyllum demersum*	AAQ11850	(Hilu, Mueller, and Borsch	AF543732	(Hilu, Mueller, and Borsch

			Unpublished)		Unpublished)
Calycanthaceae	[a]*Calycanthus fertilis var. ferax*	This study		AJ428413	(Goremykin et al. Unpublished)
Arecaceae	*Nypa fruticans*	AAQ11861	(Hilu, Mueller, and Borsch Unpublished)	AF543743	(Hilu, Mueller, and Borsch Unpublished)
Acoraceae	*Acorus gramineus*	Q9GHG8	(Fuse and Tamura 2000)	ABO40155	(Fuse and Tamura 2000)
Poaceae	*Oryza sativa*	NP_039361	(Morton and Clegg 1993)	NC_001320	(Morton and Clegg 1993)
Poaceae	*Hordeum vulgare*	P17158	(Boyer and Mullet 1988)	X64129	(Ems et al. 1995)
Poaceae	*Triticum aestivum*	NP_114240	(Ikeo and Ogihara 2000)	NC_002762	(Ikeo and Ogihara 2000)
Poaceae	*Zea mays*	CAA60266	(Maier et al. 1995)	X86563	(Maier et al. 1995)
Poaceae	*Saccharum officinarum*	YP_054609	(Asano et al. 2004)	NC_006084	(Asano et al. 2004)
Liliaceae	*Lilium columbianum*	BAB08117	(Hayashi and Kawano 2000)	AB030847	(Hayashi and Kawano 2000)
Orchidaceae	[a]*Platylepis polyadenia*	This study		AJ543946	(Salazar et al. 2003)
Orchidaceae	[a]*Manniella gustavi*	This study		AJ543944	(Salazar et al. 2003)
Orchidaceae	[a]*Schiedeella faucisanguinea*	This study		AJ543924	(Salazar et al. 2003)
Orchidaceae	[a]*Stenorrhynchos speciosum*	This study		AJ543932	(Salazar et al. 2003)
Orchidaceae	[a]*Pleione delavayi*	This study		AF503731	(Gravendeel et al. 2004)
Orchidaceae	[a]*Sacoila lanceolata*	This study		AJ543933	(Salazar et al. 2003)
Orchidaceae	[a]*Neuwiedia veratrifolia*	AY557211	(Kocyan et al. 2004)	AY557211	(Kocyan et al. 2004)
Orchidaceae	[a]*Cynorkis species*	AY370656	(Freudenstein et al. 2004)	AY370656	(Freudenstein et al. 2004)
Orchidaceae	[a]*Svenkoeltzia congestiflora*	This study		AJ543921	(Salazar et al. 2003)
Orchidaceae	[b]*Spathglottis plicata*	This study		This study	
Orchidaceae	[b]*Spathglottis gracilis*	This study		This study	
Nelumbonaceae	*Nelumbo nucifera*	AAQ11858	(Hilu, Mueller, and Borsch Unpublished)	AF543740	(Hilu, Mueller, and Borsch Unpublished)

Trochodendracea	*Trochodendron aralioides*	AAQ11869	(Hilu, Mueller, and Borsch Unpublished)	U92848	(Manos and Steele 1997)
Buxaceae	*Buxus sempervirens*	AAQ11846	(Hilu, Mueller, and Borsch Unpublished)	AF543728	(Hilu, Mueller, and Borsch Unpublished)
Amaranthaceae	*Spinacea oleracea*	Q9M13NO	(Schmitz-Linneweber et al. 2001a)	NC_002202	(Schmitz-Linneweber et al. 2001a)
Onagraceae	*Oenothera elata subspecies hookeria*	Q9MTQ1	(Hupfer et al. Unpublished)	NC_002693	(Hupfer et al. Unpublished)
Brassicaceae	*Arabidopsis thaliana*	NP_051040	(Sato et al. 1999)	NC_000932	(Sato et al. 1999)
Fabaceae	*Glycine max*	Q9TKS6	(Hu et al. 2000)	AF142700	(Hu et al. 2000)
Araliaceae	*Panax ginseng*	YP_086947	(Kim and Lee 2004)	NC_006290	(Kim and Lee 2004)
Solanaceae	*Mandragora officinarum*	CAE51524	(Clarkson et al. 2004)	AJ585883	(Clarkson et al. 2004)
Solanaceae	*Nictotiana tabacum*	NP_054478	(Nadot et al. 1995)	NC_001879	(Nadot et al. 1995)
Solanaceae	*Nicotiana occidentalis*	CAE51530	(Clarkson et al. 2004)	AJ585889	(Clarkson et al. 2004)
Solanaceae	*Anthocercis viscosa*	CAE51504	(Clarkson et al. 2004)	AJ85863	(Clarkson et al. 2004)
Solanaceae	*Petunia axillaries*	CAE51521	(Clarkson et al. 2004)	AJ787880	(Clarkson et al. 2004)
Solanaceae	*Atropa belladonna*	Q8S8Y6	(Schmitz-Linneweber et al. 2002)	AJ585882	(Clarkson et al. 2004)
Orobanchaceae	*Epifagus virginiana*	P30071	(Wolfe et al. 1992)	NC_001568	(Wolfe et al. 1992)

[a]Only nucleotide sequence available, translated in DS Gene

[b]Nucleotide and protein sequence generated in this study.

SUPPLEMENTAL TABLE 3.

Sequences used for comparing MatK, RbcL, and InfA for 22 species of green plants.

Species	MatK Genbank #	Reference	RbcL Genbank #	Reference	InfA Genbank #	Reference
Chaetosphaeridium globosum	NP_683782	(Turmel, Otis, and Lemieux 2002)	AAM96553	(Turmel, Otis, and Lemieux 2002)	NC_004115	(Turmel, Otis, and Lemieux 2002)
Physcomitrella patens subspecies patens	NP_904210	(Sugiura et al. 2003)	BAC85044	(Miyata et al. 2002)	AP005672	(Miyata et al. 2002)
Marchantia polymorpha	P12174	(Ohyama et al. 1986)	P06292	(Ohyama et al. 1986)	NC_001319	(Ohyama et al. 1986)
Anthoceros formosae	AB086179	(Kugita et al. 2003)	BAC55357	(Kugita et al. 2003)	AB086179	(Ohyama et al. 1986)
Psilotum nudum	NP_569609	(Wakasugi et al. Unpublished)	BAB84224	(Wakasugi et al. Unpublished)	NC_003386	(Wakasugi et al. Unpublished)
Adiantum capillus-veneris	NP_848039	(Wolf, Rowe, and Hasebe 2004)	NP_848068	(Wolf et al. 2003)	NC_004766	(Wolf, Rowe, and Hasebe 2004)
Pinus korainsis	BAD32752	(Gernandt, Geada-Lopez, and Liston Unpublished)	AAO74041	(Noh et al. Unpublished)	NC_004677	(Noh et al. Unpublished)
Pinus thunbergii	NP_042349	(Tsudzuki et al. 1992)	BAA04368	(Tsudzuki et al. 1992)	NC_001631	(Tsudzuki et al. 1992)
Amborella trichopoda	NP_904080	(Goremykin et al. 2003)	CAD45115	(Goremykin et al. 2003)	NC_005086	(Goremykin et al. 2003)
Nymphea alba	YP_053136	(Goremykin et al. 2004)	YP_053163	(Goremykin et al. 2004)	NC_006050	(Goremykin et al. 2004)
Calycanthus florida var. glaucus	AAS09901	(Li et al. Unpublished)	CAD28729	(Goremykin et al. Unpublished)	AJ428413	(Goremykin et al. Unpublished)
Oryza sativa	NP_039361	(Morton and Clegg 1993)	NC_001320	(Morton and Clegg 1993)	NC_001320	(Morton and Clegg 1993)
Hordeum vulgare	P17158	(Boyer and Mullet 1988)	AAN27989	(Petersen and Seberg 2003)	AY743911	(Landau, Paleo, and Prina Unpublished)
Triticum aestiva	NP_114240	(Ikeo and Ogihara 2000)	AAP92166	(Niu Unpublished)	NC_002762	(Ikeo and Ogihara 2000)
Zea mays	CAA60266	(Maier et al. 1995)	CAA60294	(Maier et al. 1995)	X86563	(Maier et al. 1995)
Saccharum officinarum	YP_054609	(Asano et al. 2004)	YP_054639	(Asano et al. 2004)	NC_006084	(Asano et al. 2004)
Spinacea oleracea	Q9M3NO	(Schmitz-Linneweber et al. 2001a)	CAB88737	(Schmitz-Linneweber et al. 2001a)	NC_002202	(Schmitz-Linneweber et al. 2001a)
Oenothera elata subspecies	Q9MTQ1	(Hupfer et al. Unpublished)	CAB67126	(Hupfer et al. Unpublished)	NC_002693	(Hupfer et al.

hookeria					pseudogene	Unpublished)
Panax ginseng	YP_086947	(Kim and Lee 2004)	YP_086974	(Kim and Lee 2004)	NC_006290	(Kim and Lee 2004)
Nictotiana tabacum	NP_054478	(Nadot et al. 1995)	NC_001879	(Nadot et al. 1995)	Z00044 pseudogene	(Nadot et al. 1995)
Atropa belladonna	Q8S8Y6	(Schmitz-Linneweber et al. 2002)	CAC88052	(Schmitz-Linneweber et al. 2002)	AF347644 pseudogene	(Millen et al. 2001)
Epifagus virginiana	P30071	(Wolfe et al. 1992)	NC_001568 pseudogene	(Wolfe et al. 1992)	NC_001568	(Wolfe et al. 1992)

SUPPLEMENTAL TABLE 4.

Sequences used for comparing MatK and Mat-r across 32 land plant species.

Species	MatK Genbank #	Reference	Mat-r Genbank #	Reference
Huperzia lucida	AY660566	(Wolf et al. 2005)	AAK55491	(Qiu et al. 2001)
Gnetum gnemon	AAS18470	(Won and Renner Unpublished)	AAF14710	(Qiu et al. 1999)
Welwitschia mirabilis	AAK69135	(Chaw and Hu Unpublished)	AAF14711	(Qiu et al. 1999)
Ginkgo biloba	AAK69129	(Chaw and Hu Unpublished)	AAF14714	(Qiu et al. 1999)
Brasenia schreberi	AAQ11845	(Hilu, Mueller, and Borsch Unpublished)	AAF14720	(Qiu et al. 1999)
Illicium florida	AAQ11856	(Hilu, Mueller, and Borsch Unpublished)	AAF14732	(Qiu et al. 1999)
Canella winteri	AAQ11849	(Hilu, Mueller, and Borsch Unpublished)	AAF14749	(Qiu et al. 1999)
Lactoris fernandeziana	AAQ11857	(Hilu, Mueller, and Borsch Unpublished)	AAF14804	(Qiu et al. 1999)
Saruma henryi	AAQ11866	(Hilu, Mueller, and Borsch Unpublished)	AAF14744	(Qiu et al. 1999)
Magnolia tripetela	AAN59945	(Wang, Zhang, and Cui Unpublished)	AAF14762	(Qiu et al. 1999)
Liriodendron chinense	AAF43259	(Jin et al. 1999)	AAF14766	(Qiu et al. 1999)
Galbulimima belgraveana	AAP02924	(Sauquet et al. 2003)	AAF14765	(Qiu et al. 1999)
Cananga odorata	AAP02921	(Sauquet et al. 2003)	AAF14755	(Qiu et al. 1999)
Calycanthus florida	AAS09901	(Li et al. Unpublished)	AAF14769	(Qiu et al. 1999)
Orontium aquaticum	AAQ11862	(Hilu, Mueller, and Borsch Unpublished)	AAF14737	(Qiu et al. 1999)
Pleea tenuifolia	BAD60953	(Tamura et al. 2004a)	AAF14735	(Qiu et al. 1999)
Triglochin maritimum	BAD20576	(Tamura et al. 2004b)	AAF14717	(Qiu et al. 1999)
Carludovica palmata	BAD20587	(Tamura et al. 2004b)	AAF14726	(Qiu et al. 1999)
Nypa fruticans	AAQ11861	(Hilu, Mueller, and Borsch Unpublished)	AAQ56265	(Shi et al. 2005)
Triticum aestiva	NP_114240	(Ikeo and Ogihara 2000)	CAA41033	(Bonen 1991)
Nelumbo nucifera	AAQ11858	(Hilu, Mueller, and Borsch Unpublished)	AAF14787	(Qiu et al. 1999)
Platanus occidentalis	AAQ11865	(Hilu, Mueller, and Borsch Unpublished)	AAF14785	(Qiu et al. 1999)

Buxus sempervirens	AAQ11846	(Hilu, Mueller, and Borsch Unpublished)	AAF14778	(Qiu et al. 1999)
Ceriops tagal	AAG35557	(Shi et al. Unpublished)	AAQ56241	(Shi et al. 2005)
Arabidopsis thaliana	AAG43347	(Koch, Haubold, and Mitchell-Olds 2001)	CAA69736	(Giege and Brennicke 1999)
Hibiscus tiliaceus	BAD89715	(Takayama et al. 2005)	AAQ56248	(Shi et al. 2005)
Heritiera littoralis	AAQ84263	(Nyffeler et al. Unpublished)	AAQ56250	(Shi et al. 2005)
Reinwardtia indica	BAB33400	(Kita and Kato 2001)	AAU03345	(Davis and Wurdack 2004)
Dicella nucifera	AAK71028	(Cameron et al. Unpublished)	AAU03287	(Davis and Wurdack 2004)
Franklinia alatamaha	AAL60378	(Prince and Parks 2001)	AAO63600	(Yang, Yang, and Li Unpublished)
Pentaphylax euryoides	CAD22187	(Bremer et al. 2002)	AAO63618	(Yang, Yang, and Li Unpublished)
Symplocos hookeri	AAR13857	(Wang et al. 2004)	AAO63622	(Yang, Yang, and Li Unpublished)
Ternstroemia gymnanthera	CAD22198	(Bremer et al. 2002)	AAO63623	(Yang, Yang, and Li Unpublished)

LITERATURE CITED

Asano, T., T. Tsudzuki, S. Takahashi, H. Shimada, and K. Kadowaki. 2004. Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA research **11**:93-99.

Bao, Y., and S. Ge. Unpublished. Origin and phylogenetic relationships of the *Oryza* species with the CD genome based on multiple-gene sequence data. Unpublished.

Bittkau, C., and G. Mueller-Starck. 2002. Direct Submission.

Bonen, L. 1991. Direct Submission.

Boyer, S. K., and J. E. Mullet. 1988. Sequence and transcript map of barley chloroplast *psbA* gene. Nucleic Acids Research **16**:8184.

Bremer, B., K. Bremer, N. Heidar, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Kallersjo, and E. Barkhordarian. 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Molecular Phylogenetics and Evolution **24**:274-301.

Cameron, K. M., M. W. Chase, W. R. Anderson, and H. G. Hills. Unpublished. Molecular systematics of Mapighiaceae: evidence from plastid *rbcL* and *matK* sequences. Unpublished.

Chaw, S.-M., and S.-H. Hu. 2000. Direct Submission.

Chaw, S.-M., and S.-H. Hu. Unpublished. Chloroplast *matK* sequence data reconfirm the monophyly of extant gymnosperms and the coniferophytic origin of Gnetales.

Clarkson, J. J., S. Knapp, V. F. Garcia, R. G. Olmstead, A. R. Leitch, and M. W. Chase. 2004. Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. Molecular Phylogenetics and Evolution **33**:75-90.

Cronn, R., R. L. Small, T. Haselkorn, and J. F. Wendel. 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. Evolution **57**:2475-2489.

Cronn, R. C., R. L. Small, T. Haselkorn, and J. F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium:* Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany **89**:707-725.

Davies, T. J. Unpublished. Environmental energy and species richness in flowering plants. Unpublished.

Davis, C. C., and K. J. Wurdack. 2004. Host-to-parasite gene transfer in flowering plants: phylogenetic evidence from Malpighiales. Science **305**:676-678.

Ems, S. C., C. W. Morden, C. K. Dixon, K. H. Wolfe, C. W. dePamphilis, and J. D. Palmer. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. Plant Molecular Biology **29**:721-733.

Freudenstein, J. V., C. V. D. Berg, D. H. Goldman, P. J. Kores, M. Molvray, and M. W. Chase. 2004. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. . American Journal of Botany **91**:149-157.

Fuse, S., and M. N. Tamura. 2000. A phylogenetic analysis of the plastid *matK* gene with emphasis on Melanthiaceae sensu lato. Plant Biology **2**:415-427.

Ge, S., A. Li, B.-R. Lu, S.-Z. Zhang, and D.-Y. Hong. 2002. A phylogeny of the rice tribe Oryzeae (Poaceae) based on *matK* sequence data. American Journal of Botany **89**:1967-1972.

Ge, S., T. Sang, B. R. Lu, and D. Y. Hong. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proceedings of the National Academy of Science USA **96**:14400-14405.

Gernandt, D., G. Geada-Lopez, and A. Liston. Unpublished. Phylogeny and classification of *Pinus*. Unpublished.

Giege, P., and A. Brennicke. 1999. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. Proceedings of the National Academy of Science USA **96**:15324-15329.

Goremykin, V., K. Hirsch-Ernst, S. Wolfl, and F. Hellwig. Unpublished. Complete structure of the chloroplast genome of *Calycanthus fertilis*. Unpublished.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2004. The chloroplast genome of *Nymphea alba*: Whole-genome analysis and the problem of identifying the most basal angiosperm. Molecular Biology and Evolution **21**:1445-1454.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Molecular Biology and Evolution **20**:1499-1505.

Gravendeel, B., M. C. M. Eurlings, C. v. d. Berg, and P. J. Cribb. 2004. Phylogeny of *Pleione* (Orchidaceae) and parentage analysis

Hall, J. C., H. H. Iltis, and K. J. Systsma. 2004. Molecular phylogenetics of core Brassicales, placement of orphan genera Emblingia, Forchhammeria, and Tirania, and character evolution. Systematic Botany **29**:654-669.

Hardy, C. R., and H. P. Linder. 2005. Intraspecific variability and timing in ancestral ecology reconstruction: a test case from the Cape flora. Systematic Biology **54**:299-316.

Hardy, C. R., and H. P. Linder. Unpublished. Phylogeny and historical ecology of Rhodocoma (Restionaceae) from the Cape Floristic Region of southern Africa. Unpublished.

Harrington, M. G., K. J. Edwards, S. A. Johnson, M. W. Chase, and P. A. Gadek. 2005. Phylogenetic inference in Sapindaceae using plastid *matK* and *rbcL* sequences. Systematic Botany **30**:366-382.

Hayashi, K., and S. Kawano. 2000. Molecular systematics of *Lilium* and allied genera (Liliaceae): phylogenetic relationships among *Lilium* and related genera based on the *rbcL* and *matK* gene sequence data. Plant Species Biology **15**:73-93.

Hilu, K. W., and L. A. Alice. 2001. A phylogeny of Chloridoideae (Poaceae) based on *matK* sequences. Systematic Biology **26**:386-405.

Hilu, K. W., and L. A. Alice. Unpublished. Phylogenetic relationships in subfamily Chloridoideae (Poaceae) based on *matK* sequences: A preliminary assessment. Unpublished.

Hilu, K. W., L. A. Alice, and H. Liang. 1999. Phylogeny of Poaceae inferred from *matK* sequences. Annals of the Missouri Botanical Gardens **86**:835-851.

Hilu, K. W., K. F. Mueller, and T. Borsch. Unpublished. Fast evolving DNA and deep level phylogenetics: a case study in basal angiosperms. Unpublished.

Hu, J. M., M. Lavin, M. F. Wojciechowski, and M. J. Sanderson. 2000. Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast *trnK/matK* sequences and its implications for evolutionary patterns in Papilionoideae. American Journal of Botany **87**:418-430.

Huang, J., D. E. Giannasi, and R. A. Price. 2005. Phylogenetic relationships in *Ephedra* (Ephedraceae) inferred from chloroplast and nuclear DNA sequences. . Molecular Phylogenetics and Evolution **35**:48-59.

Hupfer, H., M. Swiatek, S. Hornung, R. G. Herrmann, R. M. Maier, W. L. Chiu, and B. Sears. Unpublished. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable euoenothera plastomes. Molecular and General Genetics **263**:581-585.

Ikeo, K., and Y. Ogihara. 2000. Direct submission. Unpublished.

Jin, H., S. Shi, H. Pan, Y. Huang, and H. Zhang. 1999. Phylogeny of *Michelia* (Magnoliaceae) and its related genera inferred from *matK* gene sequence. Ziran Kexueban **38**:93-97.

Kim, K. J., and H. L. Lee. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA research **11**:247-261.

Kita, Y., and M. Kato. 2001. Infrafamilial phylogeny of the aquatic angiosperm Podostemaceae inferred from the nucleotide sequences of *matK* gene. Plant Biology **3**:156-163.

Koch, M., B. Haubold, and T. Mitchell-Olds. 2001. Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. American Journal of Botany **88**:534-544.

Koch, M., and T. Mitchell-Olds. 1999. Direct Submission.

Kocyan, A., Y.-L. Qiu, P. Endress, and E. Conti. 2004. A phylogenetic analysis of Apostasioideae (Orchidaceae) based on ITS, *trnL-F*, and *matK* sequences. Plant Systematics and Evolution ***In press*.**

Kugita, M., A. Kaneko, Y. Yamamoto, Y. Takeya, T. Matsumoto, and K. Yoshinaga. 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the land plants. Nucleic Acids Research **31**:716-721.

Landau, A. M., A. H. D. Paleo, and A. R. Prina. Unpublished. Plastid *infA* gene in barley. Unpublished.

Les, D. H., D. J. Crawford, E. Landolt, J. Gabel, and R. T. Kimball. 2002. Phylogeny and systematics of Lemnaceae, the duckweed family. Systematic Botany **27**:221-240.

Les, D. H., E. L. Schneider, D. J. Padgett, P. S. Soltis, and M. Zanis. 1999. Phylogeny, classification and floral evolution of water lilies (Nymphaeaceae; Nymphaeales): A synthesis of non-molecular, *rbcL, matK*, and *18S rDNA* data. Systematic Biology **In press**.

Li, J., J. Ledger, T. Ward, and P. D. Tredici. Unpublished. Phylogenetics of Calycanthaceae based on molecular and morphological data, with special reference to divergent paralogs of the nrDNA ITS region. Unpublished.

Li, X. X., and Z. K. Zhou. Unpublished. Monocotyledons phylogeny based on three genes (*matK, rbcL,* and *18S rDNA*) sequences. Unpublished.

Lopez, G. G., K. Kamiya, and K. Harada. Unpublished. Phylogeny of the North American pines. Unpublished.

Maier, R. M., K. Neckermann, G. Igloi, and H. Kossel. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of Molecular Biology **251**:614-628.

Manos, P. S., and K. P. Steele. 1997. Phylogenetic analysis of 'higher' Hamamelididae based on plastid sequence data. American Journal of Botany **84**:1407-1419.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe.

2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant Cell **13**:645-658.

Miyata, Y., C. Sugiura, Y. Kobayashi, M. Hagiwara, and M. Sugita. 2002. Chloroplast ribosomal S14 protein transcript is edited to create a translation initiation codon in the moss *Physcomitrella patens*. Biochimica et Biophysica Acta **1576**:346-349.

Modliszewski, J. L., D. T. Thomas, C. Fan, D. J. Crawford, C. W. dePamhilis, and Q.-Y. Xiang. 2005. Direct Submission.

Morton, B. R., and M. T. Clegg. 1993. A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbcL* in the grass family (Poaceae). Current Genetics **24**:357-365.

Nadot, S., G. Bittar, L. Carter, R. Lacroix, and B. Lejeune. 1995. A phylogenetic analysis of monocotyledons based on the chloroplast gene *rps4*, using parsimony and a new numerical phenetics method. Molecular Phylogenetics and Evolution **4**:257-282.

Nishikawa, T., B. Salomon, T. Komatsuda, R. v. Bothmer, and K. Kadowaki. 2002. Molecular phylogeny of the genus *Hordeum* using three chloroplast DNA sequences. Genome **45**:1157-1166.

Niu, J. S. Unpublished. Wheat chloroplast gene *rbcL*. Unpublished.

Noh, E. W., J. S. Lee, Y. I. Choi, M. S. Han, Y. S. Yi, and S. U. Han. Unpublished. Complete nucleotide sequence of *Pinus koraiensis*. Unpublished.

Nyffeler, R., A. Yen, W. S. Alverson, C. Bayer, G. Blattner, B. Whitlock, M. W. Chase, and D. A. Baum. Unpublished. Phylogenetic analysis of Malvaceae sensu lato based on chloroplast and nuclear DNA sequences. Unpublished.

Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Omesono, Y. Shiki, M. Takeuchi, Z. Chang, S. Aota, H. Inokuchi, and H. Ozeki. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature **322**:572-574.

Petersen, G., and O. Seberg. 2003. Phylogenetic analysis of the diploid species of *Hordeum* (Poaceae) and a revised classification of the genus. Systematic Botany **28**:293-306.

Prince, L. M., and C. R. Parks. 2001. Phylogenetic relationships of Theaceae inferred from chloroplast DNA sequence data. American Journal of Botany **88**:2039-2320.

Qiu, Y.-L., J. Lee, B. A. Witlock, F. Bernasconi-Quadroni, and O. Dombrovska. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya? Molecular Biology and Evolution **18**:1745-1753.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid, and nuclear genomes. Nature **402**:404-407.

Quinn, C. J., R. A. Price, and P. A. Gadek. 2002. Familial concepts and relationships in the conifers based on *rbcL* and *matK* sequence comparisons. Kew Bulletin **57**:513-531.

Rohwer, J. G. 2000. Toward a phylogenetic classification of the Lauraceae: Evidence from *matK* sequences. Systematic Botany **25**:60-71.

Salazar, G. A., M. W. Chase, M. A. S. Arenas, and M. Ingrouille. 2003. Phylogenetics of Cranichideae with emphasis on Spiranthinae (Orchidaceae, Orchidoideae): evidence from plastid and nucelar DNA sequences. American Journal of Botany **90**:777-795.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA research **6**:283-290.

Sauquet, H., J. A. Doyle, T. Scharaschkin, T. Borsch, K. W. Hilu, L. W. Chatrou, and A. L. Thomas. 2003. Phylogenetic analysis of Magnoliales and Myristicaceae based on multiple data sets: implications for character evolution. Botanical Journal of the Linnean Society **142**:125-186.

Schmitz-Linneweber, C., R. M. Maier, J. P. Alcaraz, A. Cottet, R. G. Herrmann, and R. Mache. 2001. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Molecular Biology **45**:307-315.

Schmitz-Linneweber, C., R. Regel, T. G. Du, H. Hupfer, R. G. Herrmann, and R. M. Maier. 2002. The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Molecular Biology and Evolution **19**:1602-1612.

Shi, S., Y. Huang, K. Zeng, F. Tan, H. He, J. Huang, and Y. Fu. 2005. Molecular phylogenetic analysis of mangroves: independent evolutionary origins of vivipary and salt secreation. Molecular Phylogenetics and Evolution **34**:159-166.

Shi, S., Y. Zhong, Y. Huang, and H. Chang. Unpublished. Molecular phylogenies, evolution, and biogeography of Rhizophoraceae in China. Unpublished.

Sugiura, C., Y. Kobayashi, S. Aoki, C. Sugita, and M. Sugita. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. Nucleic Acids Research **31**:5324-5331.

Takayama, K., T. Ohi-Toma, H. Kudoh, and H. Kato. 2005. Origin and diversification of *Hibiscus glaber*, species endemic to the oceanic Bonin Islands, revealed by chloroplast DNA polymorphism. Molecular Ecology **14**:1059-1071.

Tamura, M. N., S. Fuse, H. Azuma, and M. Hasebe. 2004a. Biosystematic studies on the family Tofieldiaceae I. Phylogeny and circumscription of the family inferred from DNA sequences of *matK* and *rbcL*. Plant Biology **6**:562-567.

Tamura, M. N., J. Yamashita, S. Fuse, and M. Haraguchi. 2004b. Molecular phylogeny of monocotyledons inferred from combined analysis of plastid *matK* and *rbcL* gene sequences. Journal of Plant Research **117**:109-120.

Tsudzuki, J., K. Nakashima, T. Tsudzuki, J. Hiratsuka, M. Shibata, T. Wakasugi, and M. Sugiura. 1992. 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*. Molecular and General Genetics **232**:206-214.

Turmel, M., C. Otis, and C. Lemieux. 2002. The chloroplast and mitochondrial genome sequences of the charophyte Chaetosphaeridium globosum: Insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proceedings of the National Academy of Science USA **99**:11275-11290.

Wakasugi, T., A. Nishikawa, K. Yamada, and M. Sugiura. Unpublished. Complete nucleotide sequence of the chloroplast genome from a fern, *Psilotum nudum*. Unpublished.

Wang, Y., P. W. Fritsch, S. Shi, F. Almeda, and B. Cruz. 2004. Phylogeny and infrageneric classification of *Symplocos* (Symplocaceae) inferred from DNA sequence data. American Journal of Botany **91**:1901-1914.

Wang, Y., S. Zhang, and T. Cui. Unpublished. Molecular phylogeny of Magnoliaceae. Unpublished.

Wolf, P. G., K. G. Karol, D. F. Mandoli, J. Kuehl, K. Arumuganathan, M. W. Ellis, B. D. Mishler, D. G. Kelch, R. G. Olmstead, and J. L. Boore. 2005. The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene **350**:117-128.

Wolf, P. G., C. A. Rowe, and M. Hasebe. 2004. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. . Gene **339**:89-97.

Wolf, P. G., C. A. Rowe, R. B. Sinclair, and M. Hasebe. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. . DNA research **10**:59-65.

Wolfe, K. H., C. W. Morden, S. C. Ems, and J. D. Palmer. 1992. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. Journal of Molecular Evolution **35**:304-317.

Won, H., and S. S. Renner. Unpublished. Molecular phylogeny of *Gnetum* based on chloroplast *rbcL, matK,* and *tRNA Leu* (*UAA*) intron/ IGSs sequences. Unpublished.

Yamashita, J., and M. N. Tamura. 2004. Phylogenetic analysis and chromosome evolution in Convallarieae (Ruscaceae sensu lato), with some taxonomic treatments. Journal of Plant Research **117**:363-370.

Yamashita, J., and M. N. Tamura. 2000. Molecular phylogeny of the Convallariaceae (Asparagales). Pp. 387-400 *in* K. L. Wilson, and D. A. Morrison, eds. MONOCOTS: SYSTEMATICS AND EVOLUTION. CSIRO Publishing, Melbourne.

Yang, S.-X., J.-B. Yang, and D.-Z. Li. Unpublished. Phylogenetic relationships of Theaceae inferred from mitochondrial *matR* gene sequence data. Unpublished.

APPENDIX B

MATURASE ASSAY

The protein product from *matK* is proposed to be a group II intron maturase (Neuhaus and Link 1987; Vogel et al. 1997). Thus far, there is no direct biochemical evidence of this function for MatK. Vogel et al. (1997) and Vogel, Borner, and Hess (1999) proposed a set of six putative substrates for MatK activity. These RNAs were shown to lack intron splicing in the white barley chloroplast ribosomal mutant *albostrians* (Vogel et al. 1997; Vogel, Borner, and Hess 1999). In order to examine the maturase activity of MatK, a biochemical assay was developed. Unfortunately, a lack of a sufficient quantity of MatK protein to perform the assay prevented completion of this experiment. Below is the methodology used for generating three of the substrates to be tested, cloning the reading frame of *matK* from *Arabidopsis thaliana* into an expression vector, and transformation into a bacterial expression line. In addition, the *in vitro* translation and bacterial-induction of the recombinant *matK* reading frame clone for protein expression is described as well as the protocol for conducting the maturase assay. The maturase assay was attempted using the small quantities of protein that could be generated using *in vitro* translation. However, this still was not suffiecient to determine splicing activity of this putative maturase. Both non-radioactive and radioactive methods of detecting spliced transcripts were utilized and are described below.

RNA substrates

Examination of the chloroplast ribosomal mutant *albostrians* identified potential substrates for MatK activity (Vogel et al. 1997; Vogel, Borner, and Hess 1999). Three of these potential substrates *trnA*, *rpl2*, and *atpF*, were cloned for *in vitro* transcription. Total genomic DNA from *Arabidopsis* was isolated according to Doyle and Doyle (1990). Primers were designed to amplify *trnA* (trnAxhoI, 5' CCGCTCGAGGGGGGATATAGCTCAG 3', and

trnARNotII, 5' AAGGAAAAAGCGGCCGCTGGAGATAAGCGGACT 3'), *rpl2* (rpl2FxhoI, 5' CCGCTCGAGATGGCGATACATTTA 3', and rpl2RNotI, 5' AAGGAAAAAGCGGCCGCCTATTTACTACGGCGA 3'), and *atpF* (atpFNotI, 5' AAGGAAAAAGCGGCCGCTTAATCAGTTATTTC 3', and atpFxhoI, 5' CCGCTCGAGATGAAAAATTTAACCGAT 3') including all exons and the group II intron. These genes were amplified using *Pfx* polymerase (Invitrogen, Carlsbad, CA, USA) to ensure high fidelity. Primers included restriction sites for directional cloning. These restriction sites are underlined in the primer sequence. In addition, the primers contained a short linker sequence to increase efficiency of digestion after PCR. PCR was performed using the program 48 MatK with the following conditions: 1) starting cycle of 95 ^{o}C for 3 min., 40 ^{o}C for 3 min., and 72 ^{o}C for 3 min., 2) main cycle program of 95 ^{o}C for 30 sec., 40 ^{o}C for 1.30 min., and 72 ^{o}C for 3 min., with main cycles repeated 50 times, and 3) 72 ^{o}C for 20 min. to complete end extension. Amplified products were gel excised using the Qiagen Qiaquick gel extraction kit (Qiagen, Valancia, CA, USA), digested with *Not*I (Ipswich, MA, USA) and *Xho*I (Promega, Madison, WI, USA) restriction enzymes, ethanol precipitated, and cloned into the vector pGEM 11Zf+ (Promega, Madison, WI, USA) using the *Not*I and *Xho*I restriction sites for directional cloning. The ligation reaction proceeded using a 5:1 insert to vector ratio, 1 mM ATP, and 1 μ l of 4u/μl T4 DNA ligase (NEB, Ipswich, MA, USA) in a 10 μl reaction. Ligation reactions were incubated at room temperature overnight and then stored at 4 ^{o}C until transformation. Ligated plasmids were then transformed into *E. coli* JM109 by heat shock at 42 ^{o}C for 40 seconds, and plated on LB_{Amp50} agar. Positive colonies were determined by PCR screen using M13 universal primers that lay outside the multiple cloning site of pGEM 11Zf+ and PCR protocol described above but with a Tm of 52 ^{o}C. In addition, the recombinant *rpl2* plasmid was sequenced using the primer

rpl2ex1R (5' CAAAGGTAGGGCATTTCC 3"). Cycle sequencing was performed using 2.5 µl ABI BigDye Terminator sequencing dye (ABI, Framingham, MA, USA) and a final concentration of 1.25 µM of primer in a 15 µl reaction. Sequenced products were separated through capillary gel electrophoresis at the Virginia Bioinformatics Core Laboratory Facility (Blacksburg, VA, USA).

In vitro transcription of RNA substrates

Recombinant clones of the group II intron substrates of *rpl2*, *trnA*, and *atpF* in the pGEM 11Zf+ vector were digested using *Not*I (Ipswich, MA, USA) to produce linearized templeates for the *in vitro* transcription reaction. Linearized templates were treated with proteinase K and 0.5% SDS followed by incubation at 50 ^{o}C for 30 minutes. Proteinase K, as well as other proteins, were removed by phenol/chloroform extraction and ethanol precipitiation. One microgram of purified linearized template for each substrate was transcribed into RNA using Ambion's MEGAscript *in vitro* transcription kit (Ambion, Austin, TX, USA) according to the manufacturers' instructions for the non-radioactive assay. The *in vitro* transcription reaction proceeded for a total of 4 hours. RNA transcripts were purified and recovered by ammonium acetate/phenol:chloroform extraction followed by isopropanol precipitation at –20 ^{o}C for 1 hour (Ambion, Austin, TX, USA). The template DNA used to generate these RNA transcripts was then removed by adding 1 µl DNase I and incubating for 15 minutes at 37 ^{o}C. RNA concentration was determined by quantification using a Beckman DU 520 UV spectrophotometer with a fixed absorption spectra of 260 nm and 280 nm. In order to confirm transcripts of the expected size for these genes were generated from this *in vitro* transcription protocol and confirm purity of transcripts, *in vitro* transcribed RNAs were separated on a 1%

formaldehyde gel (Gerard and Miller 1986). An RNA size stardard (Promega, Madison, WI, USA) was included on the gel for size estimation. *In vitro* transcribed RNA substrates were stored at –20 ^{o}C.

^{32}P –UTP radioactively-labeled RNA substrates were generated in a similar manner as described above for non-radiactive substrates in the MEGAscript protocol (Ambion, Austin, TX) but included 12 µl of radioactive label. The *in vitro* transcription reaction then proceeded for 15 minutes and 30 ^{o}C. Template DNA was then degraded by adding 1 µl DNase I and incubating for 15 minutes at 37 ^{o}C in order to further purify the radioactively-labeled transcripts. DNase treated transcripts were purified on Bio-Rad BioSpin6 columns (Bio-Rad, Hercules, CA, USA) and stored at –20 ^{o}C. Amount of transcripts produced that incorporated radioactive label was determined by a scintillation counter.

Probe synthesis

Non-radioactive probes for detection of spliced and unspliced RNA substrates were generated from genomic DNA of *Arabidopsis* or plasmid clones. For genomic DNA isolation, leaf material was ground under liquid nitrogen and extracted using CTAB/chloroform/ isoamyl alcohol followed by isopropanol precipitation (Doyle and Doyle 1990). Primers specific for exons of RNA substrates were used to generate Dig-labeled probes through PCR. Primer pair rpl2xhoI and rpl2ex1R were used to amplify the entire 312 bp region of the 5' exon of *rpl2*. Primers atpFex2F (5' AGAAGAACTGCGTGAAGG 3') and atpFNotI amplified 354 bp of the 3' exon of *atpF*. A probe for the 5' exon of *trnA* was generated using the *trnA/* pGEM 11Zf+ recombinant plasmid as template DNA and primers trnAFxhoI and trnAex1R (5' GTAGAGTCTTTCAGTGGC 3'), which amplified all 38 bp of the 5' exon of this tRNA and

133 bp of the 5' end of the intron. Dig-labeled probe synthesis was performed according to supplier instructions (Roche, Indianapolis, IN, USA) using the following PCR conditions: 1) starting cycle of 95 ^{o}C for 3 min., 42 ^{o}C for 3 min., and 72 ^{o}C for 3 min., 2) main cycle program of 95 ^{o}C for 30 sec., 42 ^{o}C for 1.30 min., and 72 ^{o}C for 3 min., with main cycles repeated 50 times, and 3) 72 ^{o}C for 20 min. to complete end extension. Probes were PCR cleaned using Qiagen's QIAquick PCR purification protocol according to Qiagen instructions (Qiagen, Valancia, CA). The *atpF* and *rpl2* probe PCR products were then cycle sequenced using atpFex2F and rpl2xhoI forward primers using a Tm of 42 ^{o}C. Cycle sequencing was performed following the ABI cycle sequencing protocol (Framingham, MA, USA) using 2.5 µl ABI BigDye Terminator sequencing dye (ABI, Framingham, MA, USA) and a final concentration of 1.25 µM of primer in a 15 µl reaction. Sequenced products were separated through capillary gel electrophoresis at the Virginia Bioinformatics Core Laboratory Facility (Blacksburg, VA, USA).

Cloning of the matK reading frame into the expression vector pET32a

RNA was isolated from *Arabidopsis* tissue according to Altenbaach and Howell (1981) and stored at –80 ^{o}C. Twelve micrograms of total RNA was then DNase treated by adding 1 µl of RNAse inhibitor (40 u/µl, Invitrogen, Carlsbad, CA, USA) along with 6.7 µl of 5X Reverse Transcriptase Buffer (Invitrogen, Carlsbad, CA, USA), and 4.3 µl DNase (1 u/µl, Promega, Madison, WI, USA) in a 20 µl reaction. This reaction was incubated at 37 ^{o}C for 30 minutes. One microliter of DNase stop solution (Promega, Madison, WI, USA) was then added, followed by incubation at 65 ^{o}C for 10 minutes to denature DNase. The full-length reading frame of *matK* from *Arabidopsis* plus part of the 5' UTR was amplified from cDNA prepared by first strand cDNA synthesis using the TaKaRa 3'-Full RACE Core Set (TAKARA BIO Inc, Otsu, Shiga,

Japan) according to the manufacturers' instructions. PCR amplification followed using the primers 5UTRArabprimer (5' CGCACTATGTGTCATTTCAGAACT 3') and ArabmatKSacIstop (5' CGAGCTCTTATTCATGATTGACCA 3'). A no-RT control was included in the reaction to ensure that the RNA template used was free of DNA. PCR amplification was performed using *Pfx* high-fidelity polymerase (Invitrogen, Carlsbad, CA, USA) and the 48 MatK PCR program described above using a Tm of 45 ^{o}C. The amplified product was gel excised using the Qiagen Qiaquick gel extraction kit (Qiagen, Valancia, CA, USA) and blunt-end cloned into pT7-Blue using the Perfectly Blunt cloning kit (Novagen, San Diego, CA, USA). Subsequently, recombinant plasmids were transformed into Nova Blue cells (Novagen, San Diego, CA, USA) according to the manufacturers' instructions. Clones were identified as positive for the insert using a PCR screen incorporating U19 and T7 universal primers, which have complementary sequence outside the multiple cloning site of pT7-Blue, the 48 MatK PCR program described above, and a Tm of 55 ^{o}C. Positive plasmids were then extracted using the alkaline/SDS method (Birnboim and Doly 1979). The *Arabidopsis* MatK coding region was subcloned in pET-32a (Novagen, San Diego, CA, USA) using the primers ArabmatKNcoIstart (5' CATGCCATGGATAAATTTCAAGGA 3') and ArabmatKSacIstop (5' CGAGCTCTTATTCATGATTGACCA 3'). These primrs included *Nco*I and *Sac*I restriction sites for in-frame directional cloning into the pET-32a expression vector. Restriction sites are underlined in the primer sequence. The ligation reaction proceeded using a 3.8:1 insert to vector ratio, 1 mM ATP, and 1 μ l of 4u/μl T4 DNA ligase (NEB, Ipswich, MA, USA) in a 10 μl reaction. Ligation reactions were incubated at room temperature overnight and then stored at 4 ^{o}C until transformation. Recombinant plasmid clones were then transformed into Nova Blue cells by heat shock (Novagen, San Diego, CA, USA) after incubating the cells with ligated

clones on ice for 20 minutes. Heat shocking proceeded for 40 seconds at 42 ^{o}C, followed by a 2-minute incubation on ice. After incubation, cells were recovered by the addition of 80 µl of SOC followed by shaking for one hour at 150 rpm, 37 ^{o}C. Transformed cells were then plated on LB_{Amp50} agar plates and incubated overnight at 37 ^{o}C. Bacterial colonies were PCR screened using T7 universal primers that lay outside the multiple cloning site of pET-32a and the 48 MatK PCR program described above with a Tm of 55 ^{o}C. Positive clones were sequenced to confirm in-frame insertion of the *matK* reading frame into pET-32a. Plasmids were alkaline extracted according to Birnboim and Doly (1979) and transformed into *E. coli* BL21(DE3)pLysS bacterial cells (Novagen, San Diego, CA, USA) for according to the manufacturers' instructions with the exception of prolonging incubation with shaking after the addition of SOC to 1 hour at 150 rpm, 37 ^{o}C. Cells were then plated on $LB_{Amp50Cam34}$ plates for overnight incubation at 37 ^{o}C.

Bacterial induction of MatK expression

BL21(DE3)pLysS cells containing the pET-32a recombinant plasmid that included the *matK* reading frame (plasmid C18) were grown overnight at 37 ^{o}C with shaking at 150 rpm in a 5 ml culture of $LB_{Amp50Cam34}$. The following morning, a 1:100 dilution was made into 25 mls $LB_{Amp50Cam34}$. This subculture was incubated with shaking at 150 rpm for ~3 hours. Once the OD_{600} of the culture reached between 0.5-0.7, 1 mM IPTG was added for induction. Prior to induction a 100 ul pre-induction sample was taken and spun at 15,000 x g, 4 ^{o}C, for 10 minutes. The supernatant was removed after centrifugation and pellet stored at –20 ^{o}C. The induced culture was shaken at 150 rpm for either 3 hours at 37 ^{o}C or overnight at room temperature. After induction, a 100 µl post-induction sample was taken and spun in the centrifuge at 15,000 x g, 4 ^{o}C, for 10 minutes. The supernatant was removed and pellet stored as described for the pre-

induction sample. The rest of the 25 mls culture was spun at 10,000 x g, 4 ^{o}C, for 10 minutes. The pellet was stored at –20 ^{o}C. An induction control (BL21(DE3) cells with pET-32a plasmid including only the expression tags, Novagen, San Diego, CA, USA) as well as a non-induced control of the C18 plasmid were included in inductions tests. A non-induced control of C18 was included to ensure any new proteins found after induction were not the result of merely longer bacterial incubation but induction for MatK protein. Pre- and post-induction samples were analyzed by SDS-PAGE, transferred to 0.22 μm nitrocellulose and immunoblotted using an anti-His antibody (Qiagen, Valancia, CA, USA) targeted against the 6X Histidine tag of pET-32a according to Qiagen instructions. Immunoreactive signal was determined using HRP-conjugated Donkey anti-mouse (diluted 1:10,000 in 5% Carnation/PBS-T) secondary antibody (Jackson immunologics, West Grove, PA, USA) and ECL peroxidase/luminol system (Amersham Biosciences, Piscataway, NJ, USA) or West Pico chemiluminescent detection system (Pierce Biotechnology, Inc., Rockford, IL, USA) followed by exposure to film.

In vitro transcription/translation of recombinant plasmids

MatK protein was not produced by induction using the *E. coli* BL21(DE3)pLysS bacterial induction system described above. Therefore, the TNT® Quick Coupled Transcription/ Translation system (Promega, Madison, WI, USA) and Flexi® Rabbit Reticulocyte Lysate System (Promega, Madison, WI, USA) were utilized for non-radioactive *in vtiro* translation of MatK protein. One microgram of C18 alkaline-extracted plasmid was used for the TNT® Quick Coupled Transcription/Translation system (Promega, Madison, WI, USA). The reaction proceeded according to Promega instructions. Incubation of the reaction was performed at 30 ^{o}C for 90 minutes, followed by storage of the resulting products at 4 ^{o}C.

PCR product was used for *in vitro* expression of MatK using the Promega Flexi® Rabbit Reticulocyte Lysate System (Promega, Madison, WI, USA). The region of the C18 pET-32a plasmid incorporating the 6X Histidine tag through the full-length reading frame for *matK* was amplified using *Pfx* polymerase (Invitrogen, Carlsbad, CA, USA) and T7 universal primers. This region included a T7 promoter. The 48 MatK PCR program described above was used with a Tm of 55 ^{o}C. Amplified products were resolved on a 0.8% agarose gel. The ~2 kb band expected from this amplification for the *matK* reading frame and 6X Histidine tag was gel excised using the Qiagen Qiaquick gel extraction kit (Qiagen, Valancia, CA, USA). The cleaned PCR band was then *in vitro* transcribed using Ambion's T7 MEGAscript *in vitro* transcription kit (Ambion, Austin, TX, USA) according to the manufacturers' instructions. Background DNA was removed after completion of the *in vitro* transcription reaction by the addition of 1 µl DNase and incubation for 15 minutes at 37 ^{o}C. Transcripts were purified and recovered by ammonium acetate/phenol:chloroform extraction followed by isopropanol precipitation at –20 ^{o}C for 1 hour (Ambion, Austin, TX, USA). RNA concentration was determind using a Beckman DU 520 UV spectrophotometer with a fixed absorption spectra of 260 nm and 280 nm. In order to confirm a transcript of the expected size was generated from this protocol and confirm purity of transcripts, *in vitro* transcribed RNAs were separated on 1% formaldehyde gel (Gerard and Miller 1986). One microgram of this purified RNA transcript of the C18 region that includes from the T7 promoter through the *matK* reading frame was used in a 100 µl *in vitro* translation reaction using the Promega Flexi® Rabbit Reticulocyte Lysate System (Promega, Madison, WI, USA). The *in vitro* translation reaction proceeded according to Promega instructions with the following modifications: the reaction volume and all components were doubled from the manufacturers' protocol, and magnesium acetate was not included in the reaction.

Translated products from both *in vitro* translation systems were resolved on 7.5% SDS-PAGE gels, followed by transfer to nitrocellulose, and immunoblot detection with anti-His antibody (Qiagen, Valancia, CA, USA), HRP-conjugated Donkey anti-mouse (diluted 1:10,000 in 5% Carnation/PBS-T; Jackson immunologics, West Grove, PA, USA) as the secondary antibody, and ECL peroxidase/luminol system (Amersham Biosciences, Piscataway, NJ, USA) or West Pico chemiluminescent detection system (Pierce Biotechnology, Inc., Rockford, IL, USA). Chemiluminescent signals were identified by exposure to film.

Maturase assay

The maturase splicing assay was performed according to Noah and Lambowitz (2003) with the exception of using 100-200 nM protein and 10-20 nM of each RNA substrate in a 100 µl volume reaction. RNA was renatured by heating at 50 ^{o}C for 1 min. and then cooling slowly to 30 ^{o}C over 20 min. in the reaction buffer. The reaction buffer had a final concentration of 40 mM Tris (pH= 8.0), 5 mM $MgCl_2$, and 0.5 M NH_4Cl. Aliquots of the splicing reaction were taken at 30 min. and 1 hour after incubation at 30 ^{o}C. For self-splicing reactions, RNA was renatured as mentioned above in 40 mM Tris (pH= 8.0), 5 mM $MgCl_2$, and 0.5 M NH_4Cl, but following renaturing of the RNA, self-splicing was initiated by raising the $MgCl_2$ concentration to 50 mM. Self-splicing reactions were incubated at 30 ^{o}C. Aliquots were removed at 30 min. and 1 hour intervals. Phenol-CIA (25:24:1) was added to terminate self-splicing and maturase activity reactions followed by ethanol precipitation of products (Noah and Lambowitz 2003).

If non-radioactively-labeled substrates were used in the maturase assay, precipitated products were rehydrated in MOPS/EDTA/formaldehyde/deionized formamide buffer and resolved on a 1% formaldehyde gel (Gerard and Miller 1986). An RNA size marker (Promega,

Madison, WI, USA) was included on the gel for size estimation. Resolved RNA was transferred to nylon membrane and hybridized with the non-radioactively-labeled probes described above according to Church and Gilbert (1984) with the exception of using non-radioactive probes. Hybridized probes were detected by chemiluminescent signal after incubation with anti-Dig antibody conjugated with alkaline phosphatase (Roche, Indianapolis, IN, USA) and addition of the chemiluminescent substrate CDP-Star (Roche, Indianapolis, IN, USA) according to the manufactures' instructions.

Radioactively-labeled substrates were rehydrated after ethanol precipitation in 10 µl dH_20 and resolved on a 4% urea polyacrylamide gel. The polyacrylamide gel was then dried and spliced substrates detected directly by exposure of the gel to film.

LITERATURE CITED

Altenbach, S. B. , and S. H. Howell. 1981. Identification of satellite RNA associated with turnip crinkle virus. Virology 112:25-33.

Brinboim, H. C., and J. Doly. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Research 7:1513-1523.

Doyle, J. J., and J. L. Doyle. 1990. Isolation of plant DNA from fresh tissue. Focus 12:13-25.

Neuhaus, H., and G. Link. 1987. The chloroplast $tRNA^{Lys}$ (UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Current Genetics 11:251-257.

Noah, J. W., and A. M. Lambowitz. 2003. Effects of maturase binding and Mg^{2+} concentration on group II intron RNA folding investigated by UV-cross-linking. Biochemistry 42:12466-12480.

Vogel, J., T. Borner, and W. Hess. 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Research 27: 3866-3874.

Vogel, J., T. Hubschmann, T. Borner, and W. R. Hess. 1997. Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for MatK as an essential splicing factor. Journal of Molecular Biology 270:179-187.