MOLECULAR ANALYSIS OF AN RNA EDITING CIS-ELEMENT AND ITS TRANS-FACTOR

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Aziana Ismail

August 2011

ABSTRACT

In angiosperm organelles, RNA editing alters specific cytidines to uridines. The mechanism involves recognition of *cis*-sequences surrounding specific Cs by nuclear-encoded proteins, but the particular molecular interactions and catalytic activities remain unclear. Functional analyses of the *cis*-elements suggest the upstream sequences act as binding sites for editing *trans*-factors. One trans-factor, *REQUIRED FOR ACCD RNA EDITING 1* (*RARE1*), is essential for RNA editing in the chloroplast *accD* transcript. This study examines 19 Brassicaceae species for editing patterns in the *accD* transcripts and utilizes comprehensive sequence analysis of *RARE1* homologs to analyze the evolutionary interaction between the *cis*-elements and *trans*-factors. The overall Ka/Ks ratio suggests all orthologous *RARE1* genes undergo negative selection although the varying Ka/Ks ratios for individual motifs indicate certain motifs are more conserved. In Brassicaceae species lacking editing at the *accD* site, *RARE1* orthologs show significant sequence variation indicating possible lost editing function or an alternate function.

BIOGRAPHICAL SKETCH

Aziana Ismail was born to Ismail Harun and Wan Mas Wan Senik of Kelantan, Malaysia in 1985. She attended Tengku Muhammad Faris Petra Science Secondary School and graduated in 2002. Afterward, Aziana attended Rochester Institute of Technology with Honors in May 2007. She moved to Ithaca in the fall of 2008 to begin graduate school at Cornell University in the Field of Genetics and Development.

ACKNOWLEDGMENTS

I would like to thank the Malaysian government for providing me the scholarship to pursue my study at Cornell. I would like to thank my advisor, Maureen Hanson, for the opportunity to have been a member of her lab for the past two years and for her guidance. I would also like to thank the former and current members of the Hanson lab for the experience. Lastly, I would also like to thank the National Science Foundation for providing the funding that has made this experience possible.

TABLE OF CONTENTS

Abstract

In angiosperms, RNA editing alters RNA sequences in both plastids and mitochondria from cytidine to uridine and less frequently from U to C. The mechanism is mediated by the recognition of *cis*-elements surrounding the targeted Cs by nuclear-encoded proteins, but the molecular interactions and the catalytic activities of the RNA editing apparatus remain unclear. Functional analyses of the *cis*-elements suggest that the upstream sequences of the C targets are important as binding sites for editing *trans*-acting factors. Most *trans*-factors that have been identified are members of the pentatricopeptide repeat (PPR) protein family, which are characterized by a tandemly repeated 35 amino acid motif. The PPR proteins involved in RNA editing are thought to be site-specific factors where each *trans*-factor recognizes at least one editing target. One *trans*-factor, *REQUIRED FOR ACCD RNA EDITING 1* (*RARE1),* has been shown to be required for RNA editing in the chloroplast *accD* transcript at C794. In this study, I examine 19 species of Brassicaceae for editing patterns in the *accD* transcripts, which encode the β–carboxyl transferase subunit of acetyl-coA carboxylase (ACCase). This study also involves comprehensive sequence analysis of RARE1 homologs in these species to analyze the evolutionary interaction between the *cis*-elements and the *trans*-factors. All orthologous RARE1 genes are under negative selection as indicated by the low substitution rate represented in K_a/K_s ratios. However, the K_a/K_s ratios for individual PPR motifs, the E- and DYW-motifs domain vary indicating that certain motifs are more conserved, perhaps due to the existence of elements that are necessary for RNA recognition. In Brassicaceae species where editing of the *accD* site is unnecessary, analysis of RARE1 orthologs shows significant sequence variation which suggests the protein might have lost editing function or may have an alternate function.

Introduction

RNA editing is a post-transcriptional modification of RNA transcripts that involves insertion, deletion, or conversion of particular nucleotide residues (Gott & Emeson 2000). The phenomenon was first observed in kinetoplastid protozoa where uridine residues were inserted and deleted from mitochondrial RNAs (Benne et al. 1986). Various RNA editing systems have been characterized in major eukaryotic lineages including animals, plants, and fungi as well as in viruses (Covello & Gray 1993; Smith, Gott, & Hanson 1997). Currently, all analyzed land plants are known to undergo cytidine (C)-to-uridine (U) editing in their organellar transcripts with the exception of marchantiid liverworts (Freyer et al. 1997). C-to-U RNA editing is observed both in the chloroplast and mitochondria and less frequently U-to-C has been reported (Freyer, Kiefer-Meyer, & Kössel, 1997; Maier et al., 1996; Mulligan, Chang, & Chou, 2007; Salmans et al., 2010; Shikanai, 2006; Zehrmann et al.,, 2008). In typical angiosperms, ~30-40 sites in chloroplast transcripts undergo C-to-U editing while more than 400 C targets have been identified in mitochondrial transcripts of a single plant species (Bentolila, Elliott, & Hanson, 2008; Handa, 2003; Notsu et al., 2002; Picardi & Quagliariello, 2008). Unlike RNA editing in animals, which results in functional protein diversity, for example in Apolipoprotein B (Chen et al. 1987; Powell et al. 1987), the purpose of plant RNA editing appears to be correction of defective genes at the RNA transcript level. The mechanism seems to be required for maintenance of gene function by restoring evolutionarily conserved amino acids or by creating a translational start or stop codon. This is supported by the observation that editing often changes the first and second codon positions with very few changes in third codon position (Cuenca et al. 2010; Giegé & Brennicke, 1999; Handa, 2003; Jobson & Qiu, 2008; Mower & Palmer, 2006)

In vivo transgenic analysis using plastid transformation and *in vitro* editing assays have identified *cis*-elements in the RNA sequence surrounding the editing targets (Bock et al. 1994; Chaudhuri et al. 1995; Hayes et al. 2006; Hayes & Hanson 2007; Hayes & Hanson 2008). Functional analysis of RNA editing *cis*-elements indicated the importance of upstream sequences of the target C for editing (Hayes & Hanson, 2007; Heller, Hayes, & Hanson, 2008; Miyamoto, Obokata, & Sugiura, 2004). Approximately 30 nucleotides upstream and 10 nucleotides downstream of the target C are essential for editing as binding sites for site-specific *trans*-factors (Hayes & Hanson, 2007; Miyamoto et al., 2004). Extensive *cis*-element sequence analysis at the tobacco chloroplast *psbE* editing site C nucleotide position 214 (NT*psbE* C214) identified specific nucleotides within the immediate upstream region that were critical for editing efficiency (Hayes & Hanson, 2007). Additionally, single nucleotide alteration within the cis-elements can greatly reduce the editing efficiency for a particular site (Chaudhuri & Maliga 1996; Bock et al. 1997; M L Reed et al. 2001; Hayes & Hanson 2007; Hayes et al. 2006). Sequence analysis of target C sites in divergent organisms showed editing *cis*-elements are highly conserved in homologous genes yet vary between transcripts of different genes within the same species suggesting certain *cis*-elements interact with specific a *trans*-factor (Hayes & Hanson 2008; Hammani et al. 2009).

While the catalytic deaminase itself responsible for conversion of C to U has not yet been identified, a number of *trans*-factors involved in RNA editing have been discovered. Most of the identified editing *trans*-factors belong to the pentatricopeptide (PPR) protein family. The PPR protein family is characterized by the presence of degenerate 35-amino acid repeats and is highly expanded in plants with more than 450 members found in the *Arabidopsis thaliana* genome (O'Toole et al., 2008; Schmitz-Linneweber & Small, 2008; Small & Peeters, 2000). Members of

the PPR protein family are known to be involved in various RNA metabolic processes, including RNA stabilization, translation, processing, splicing, and editing (Barkan et al.,1994; Beick et al., 2008; Kotera, Tasaka, & Shikanai, 2005; Schmitz-Linneweber et al., 2006). PPR proteins are also involved in the suppression of aberrant mitochondrial RNAs associated with cytoplasmic male sterility (Bentolila, Alfonso, & Hanson, 2002). In plants, the PPR protein family is divided into P and PLS subfamilies based on the type of repeats present (Lurin et al., 2004; Small & Peeters, 2000). The PLS subfamily is further divided into E and DYW classes based on the motifs present in the C-terminus of the protein. Most *Arabidopsis* PPR proteins are predicted to be targeted to organelles with ~54% and ~19% thought to be targeted to the mitochondria and plastids respectively (Lurin et al. 2004). All PPR proteins known to be involved in RNA editing are members of the PLS subfamily representing both the E class and DYW class (Lurin et al. 2004). A PPR protein, *CHLORORESPIRATORY REDUCTION 4* (CRR4), was the first chloroplast editing *trans*-factor identified and is involved in editing at the second C of the *ndhD* transcript to create a translational start codon. To date, 15 PPR proteins have been identified as editing *trans*-factors for 23 out of 34 editing sites in *Arabidopsis* chloroplasts (Cai et al., 2009; Chateigner-Boutin et al., 2008; Hammani et al., 2009; Kotera et al., 2005; Okuda et al., 2009; Okuda et al.,, 2008; Okuda et al., 2007; Robbins, Heller, & Hanson, 2009; Tseng et al., 2010; Yu et al., 2009; Zhou et al., 2008) and 11 PPR proteins have been identified to be involved in RNA editing for at least 17 editing sites in *Arabidopsis* mitochondria (Bentolila, Knight & Hanson, 2010; Sung, Tseng, & Hsieh, 2010; Takenaka, 2010; Takenaka et al., 2010; Verbitskiy et al., 2011; Verbitskiy et al, 2010; Zehrmann et al., 2009), indicating that a single *trans*-factor can recognize one site or may recognize multiple editing sites. Recombinant CRR4 expressed in *Escherichia coli* has been shown to specifically bind to 25 nucleotides upstream and 10

nucleotides downstream of its C target with a greater preference for the pre-edited transcript than post-edited transcript (Okuda et al., 2006). Additionally, other PPR proteins involved in other mRNA processing events have been shown to directly bind to specific sequences (Beick et al., 2008; Nakamura et al., 2003; Pfalz et al., 2009; Schmitz-Linneweber et al., 2006; Schmitz-Linneweber, Williams-Carrier, & Barkan, 2005). One major question that arises is how do the editing factors recognize the specific sequences in the *cis*-elements? Several editing factors (Okuda et al. 2009; Chateigner-Boutin et al. 2008; Hammani et al. 2009) recognize multiple sites; however, the editing sites requiring the same factor lack substantial sequence identity in the *cis*-elements. A systematic bioinformatics analysis of all plastid editing sites in *Arabidopsis* suggested that the editing factors can distinguish specific bases at certain positions but generally distinguish pyrimidines from purines (Hammani et al. 2009).

The editing processes in both chloroplast and mitochondria share characteristics such as the conservation of the *cis*-elements and the requirement for nuclear-encoded *trans*-factors suggesting both systems originated from common evolutionary roots (Tillich et al. 2006). RNA editing events are simultaneously present or absent in both mitochondria and plastid in a certain phylogenetic group which further indicates that the mechanisms arose together (Mulligan et al. 2007). Despite the possible shared general mechanism for editing, the cytidines that are targeted for editing differ between the organelles and also between species (Freyer et al. 1997). However, there is some species-specific diversity, in that the same position sometimes requires editing or instead has a genomically encoded T (Covello & Gray, 1993; Tillich et al. 2006; Tillich et al., 2009). Therefore, comparisons of *cis*-elements between plants that edit or do not edit at a particular site could provide insight into the emergence or loss of editing sites within the same lineages. Because the *cis*-elements are recognized by specific *trans*-factors, there may be a

possible correlation between the acquisition or loss of editing sites and the selection that is acting on a specific *trans*-factor. C-to-T mutations at a site that is usually edited in the chloroplast could affect the substitution rate in the corresponding nuclear-encoded *trans*-factor, leading to a strong evolutionary interaction between the nuclear and chloroplast genomes.

In this study, I performed molecular analysis and proposed evolutionary interactions in the *cis*-elements of *accD* C at position 794 and its *trans*-factor RARE1 in Brassicaceae. Using an *in vitro* editing assay, I identified the *cis*-element of *Arabidopsis accD* C794 that is critical for editing efficiency. Sequence analysis of *accD* and *RARE1* orthologs in Brassicaceae suggest the presence of evolutionary interaction between the RNA editing *cis*-element and its *trans*-factor. The results indicate a correlation between the absence of editing sites and the substitution rates in the corresponding *trans*-factor.

Materials and Methods

Plant Material and Preparation of Nucleic Acids – The plant species used in this study are listed in Supplemental Table S2. Total cellular DNA for all plants was isolated using the pH 5.0 CTAB (Cetyl trimethylammonium bromide) protocol containing CTAB, 1 M Tris pH 8.0, 0.5 M EDTA (EthylenediaminetetraAcetic acid Di-sodium salt) pH 8.0, 5 M NaCl, 1.0 g PVP 40 (polyvinyl pyrrolidone (vinylpyrrolidine homopolymer) Mw 40,000) and distilled H_2O. ~200 mg of tissues were collected, ground in liquid nitrogen, and suspended in CTAB buffer for DNA extraction. Template DNAs were amplified using accD F and accD R for PCR amplification (*Taq* mastermix kit, Qiagen) of *accD* in the selected plants. DNA samples of these plants were also subjected to PCR amplification for *RARE1* with RARE1 F and RARE1 R primers. Primers used in this study are listed in Supplemental Table S1. Total RNA was isolated using Trizol (Invitrogen) from ~200mg of plant tissues. Contaminating DNA was removed using Turbo DNAse (Ambion) and cDNA was synthesized by reverse transcription (RT) (Omniscript; Qiagen) using degenerate hexamers. cDNA was PCR amplified using accD F and accD R for amplification of *accD* in the selected species. Amplified transcripts were assayed for editing extent using the poisoned primer extension assay and were sequenced for sequence analysis.

Synthesis of Editing Substrates In Vitro – The synthesis of RNA editing substrates was described previously (Heller, Hayes, & Hanson, 2008). DNA templates for *in vitro* editing were made by PCR amplification from *Arabidopsis thaliana* ecotype Columbia (Col-0) genomic DNA using gene-specific primers (Intergrated DNA Technology) containing overhanging bacterial fragments SK 5' and KS 3' (Supplemental Table S1). The T7 promoter sequence was then added by a subsequent PCR step to the 5' end of the templates. Truncated DNA templates were

made by the incorporation of the primers flanked with SK 5' and KS 3' for 125 nucleotides upstream and 50 nucleotides downstream of the *accD* target C. Truncations were performed at every 25 nucleotides on both directions and then at every 5 nucleotides for template with 35 nucleotides upstream or downstream of the C target. Scanning mutation templates were made by incorporation of mismatches for every 5 nucleotides in primers used for PCR (supplemental S1) within 35 nucleotides upstream or downstream of the target C. RNA substrates were then transcribed using the T7 MEGAshortscript kit (Ambion) from the PCR products and RNA transcripts were purified using the RNA clean-up kit-5 (Zymo Research).

Preparation of Pea Chloroplast Extracts and In Vitro Editing Reaction – Editing reactions for RNA editing substrate were as described previously (Hegeman, Hayes, & Hanson, 2005), 0.1 fmol of RNA was added to 4 μL of pea chloroplast extract in assay condition. Pea chloroplast extracts were prepared from 15 – 20 day old pea plants grown in short-day condition. Leaves were homogenized and plastids were isolated using a Percoll (Amersham Biosciences) gradient. Intact chloroplasts were lysed using Triton X-100, and dialyzed in Dialysis Buffer Extraction of the pea chloroplast is identical to the previously described protocol. Editing of RNA substrates was analyzed using the poisoned primer extension (PPE) assay (Hegeman, Hayes, & Hanson, 2005)

Analysis of RNA editing extent – Poisoned primer extension (PPE) was performed on RT-PCR products to determine the editing extent in truncated and scanning mutation *Arabidopsis accD* substrates (Hegeman, Hayes, & Hanson 2005). PPE was also performed on RT-PCR products

from the selected plant species to determine the editing extent of *accD* transcripts in each species.

Sequence analysis – PCR products from amplification of DNA and cDNA with accD F and accD R primers were sequenced to determine the editing pattern of *accD* transcripts in the selected species. The sequences from the DNA and cDNA templates in individual species were compared to determine whether the C at position 794 is edited to U, remained unedited or genomically encoded as T. PCR products from amplification of DNA with *Arabidopsis* RARE1 F and RARE1 R primers were sequenced in all selected species with additional internal primers. The sequences were compiled using Sequencher v4.10.1 (Gene Codes) and aligned using the ClustalW program (Thompson, Higgins, and Gibson 1994).

Analysis of accD cis-element and RARE1ortholog sequences – Additional genomic *accD* sequences from various species across the taxa were obtained from Genbank at the National Center for Biotechnology Information (Supplemental Table S3) to create a dataset for *cis*-element and flanking sequence analysis. DNA sequence alignments for *accD* and *RARE1* were performed with MEGA5 software (Tamura et al. 2011). Phylogenetic trees were constructed for Brassicaceae family and also for all the species in the *accD* dataset. The calculation for nucleotide substitutions and K_a/K_s ratios in *RARE1 orthologs* were calculated using the synonymous and non-synonymous substitution calculation program from DnaSP v5.

Results

In vitro editing assay identifies the minimal sequences that are required for editing and the cluster of sequences that are critical for editing efficiency.

An AtaccD C794 substrate (C at position 794 from A of ATG of the *Arabidopsis thaliana accD* gene), containing 125 nucleotides upstream and 50 nucleotides downstream of the C target flanked with SK and KS sequences on the 5' and 3' ends, respectively, was constructed for an *in vitro* editing assay. Additional substrates were constructed with truncation of every 25 nucleotides from either the 5' or 3' ends to determine the minimal sequence requirements for editing in AtaccD C794 (Figure 1a). In substrate (50-) with 50 nucleotides upstream and 50 nucleotides downstream of the target C, the editing efficiency was similar to substrate (125-). Truncation substrates containing more than 25 nucleotides on either the 5' or 3' ends retained editing efficiency, but substrates with less than 25 nucleotides on either direction had reduced editing efficiency to less than 15% relative to other edited substrates (Figure 1b). Editing extent in the RNA substrates were quantified using poisoned primer extension (PPE) assay by comparing the presence of edited substrates to the unedited substrates (Figure 1c). When the substrates were truncated to 25 nucleotides for both upstream and downstream, substrate (25-/25+), the editing activity was almost completely eliminated signifying the importance of the sequences for RNA editing (Figure 1b).

Based on the data above, the sequence requirement for editing should be less than 50 nucleotides on both directions from the target C. To further explore the minimal sequence requirements for AtaccD C794, we created a series of truncation substrates with every 5 nucleotides were deleted from the (50-) substrate (Figure 2a). The PPE assay for these series of truncation substrates indicate reduced editing efficiency of less than 20% in the cluster between

substrate (30-) to (35+). Once the minimal sequence requirement was determined, scanning mutation substrates were made with blocks for every 5 nucleotides mutated from the wild-type sequence to the complementary nucleotides on either directions from the target C794 (Figure 2b). The relative editing efficiency was quantified. Editing efficiency was reduced to less than 20% for substrates (-15_-11)*, (-10_-6)*, (-5_-1), (+1_+5)*, and (+11_+15)*, which collectively span the region from -15 to +15 from the target C. Sequences -15 to -6 and 5 nucleotides immediately upstream of the target C had no editing activity suggesting a potential role in recognition or binding sites for the *trans*-factor. Within this cluster, substrate (+6_+10)* had an editing efficiency similar to the controls (35-) and (35+) which suggests that the sequences mutated within the region are not essential for sequence recognition.

High sequence conservation of accD 794 cis-element in Brassicaceae and in various other plant species across the taxa

A *cis*-element for accD C794 has been identified in *Arabidopsis* from the *in vitro* editing assay. A C corresponding to C794 in *Arabidopsis* chloroplast transcripts is edited to U in several other species in Brassicaceae (Figure 3a). We investigated the editing pattern, i.e., whether a C or T is present at position 794 in several species in Brassicaceae (supplementary Table S2). Of the analyzed Brassicaceae species, all have a C present at the corresponding position and therefore are expected to undergo editing except for *Lobularia maritima* and *Draba nemorosa* which have a genomically encoded T (Figure 3a). To determine whether a C794 in a Brassicaceae species is actually edited in chloroplast transcripts, cDNAs from *accD* transcripts from leaf tissues of those species were generated. All analyzed plants with C794 were modified to U794, which changes the genomically encoded serine to leucine (Robbins, Heller, & Hanson 2009) except for

Moricandia arvensis which has an unedited C at the corresponding position (Figure 3a). A serine codon is conserved at the same position in *accD* transcripts in other taxa (Robbins, Heller & Hanson 2009); however *Arabidopsis* mutants with unedited C794 have no obvious mutant phenotype. Quantification of editing extent in the species listed in Figure 3 indicated the presence of the editing event in Brassicaceae species with C794 except for *M.arvensis* (data not shown).

The sequence conservation of the *accD* C794 *cis*-element was analyzed by aligning 15 nucleotides upstream and 10 nucleotides downstream of the target site to determine the selection acting on these sequences. The *accD* sequences in 19 species that belong to the Brassicaceae family were further analyzed (Figure 3a). Figure 3 shows the sequence logos of *accD* cis-element in Brassicaceae species with a genomically encoded C at 794 (Figure 3b). The analysis was also performed on Brassicaceae species with a genomically encoded T at position 794 (Figure 3c). When compared, the two sequence logos for the Brassicaceae *accD cis*-element showed significant similarity, suggesting that the *accD cis*-element is highly conserved even though two species with a genomically encoded T do not require editing. The information could not elucidate the importance of the *cis*-element for editing site recognition nor suggest sequence elements that are specific to each group. Thus, the *cis*-element sequence analysis was expanded to species outside of the Brassicaceae family in order to determine whether a certain sequence profile is present in the *cis*-element of *accD* that is critical for editing. A dataset of *accD* sequences in various species across the taxa from the NCBI database was created (Figure 3g; Supplemental Table S3). Figure 3d shows the sequence logo for the *cis*-element of species with a genomically encoded C at 794 while Figure 3e is the sequence logo for species across the taxa with a genomically encoded T at the corresponding position. There is a slight sequence variation

for both sequence logo; those variations are located at the third codon position. When the *accD* *cis*-elements in all the species in the dataset were compared, combining those with a genomically encoded C and those with a genomically encoded T, the sequence logo (Figure 3f) shows a high degree of conservation, likely due to the fact that the editing site is located in the coding region of the *accD* gene. The *accD* gene encodes the β-carboxyl transferase subunit of acetyl-coA carboxylase (Wakasugi, Tsudzuki, & Sugiura, 2001). The ratio of the number of non-synonymous (K_a) to the number of synonymous (K_s) substitutions (K_a/K_s ratio) were calculated for the *accD* gene in all of the analyzed species. The *accD* gene is under negative selection with a K_a/K_s ratio of less than 1 (data not shown).

Negative selection is acting on RARE1 PPR motifs in Brassicaceae

A major question in plant organellar RNA editing is how editing factors recognize their editing target sites. As mentioned, the target sites are recognized by the presence of the *cis*-element that is usually concentrated within 15 nucleotides upstream of the target C (Hirose & Sugiura 2001; Miyamoto et al., 2002, 2004; Kobayashi et al., 2008). The mechanism for the recognition of the specific *cis*-element by PPR proteins is not understood (Hammani et al., 2009). If the PPR proteins and *cis*-elements are both under selective pressure for their role in RNA editing, loss of the target Cs should reduce or eliminate the selection pressure allowing for variations in the PPR proteins as well as the *cis*-elements. To investigate this possibility, I sequenced *RARE1* in Brassicaceae species with various pattern of editing in the *accD* transcripts (Figure 3a). *RARE1* orthologs in Brassicaceae were sequenced and analyzed for 1) presence or absence of the complete *RARE1*; 2) presence of selection on the full *RARE1* sequence; 3) patterns of selection pressure that are acting on individual PPR motifs in *RARE1*.

The 19 species listed in Figure 3a were selected for further analysis for *RARE1* based on the following hypotheses that 1) *RARE1* in species requiring editing will be highly conserved; 2) *RARE1* in species with a genomically encoded T will have higher sequence variation when compared to species requiring editing. The sequence results show the presence of *RARE1* orthologs in Brassicaceae species requiring editing at C794 and also in the species with a genomically encoded T at the corresponding position (Figure 4a). I reasoned that since the Brassicaceae species are closely related, changes in the species that have lost its editing target are limited and all the components (i.e. *cis*-element, *trans*-factor, catalytic enzyme and other co-factors) involved are still present in the genome. To determine the selective pressure that is acting on *RARE1* orthologs, the K_a/K_s ratios were calculated for Brassicaceae species requiring editing and Brassicaceae species with a genomically encoded T. The K_a/K_s ratio for the non-edited dataset was about 0.11 (Figure 4b) indicating it is under a strong negative selection. However, this dataset consisted of only consist of two species, *L .maritima* and *D. nemorosa*. When both groups were combined and analyzed, the Ka/Ks ratio gave a value of less 1 indicating the gene is under strong negative selection (data not shown). *RARE1* has 15 PPR motifs, an E-motif and a DYW-motif each of which may have different evolutionary pressure based on the importance of the sequence in RNA editing. I looked at the K_a/K_s ratio for the individual motifs and obtained various values for the K_a/K_s ratio (Figure 4b). A K_a/K_s ratio greater than one implies positive selection where the gene should have more sequence variation and a K_a/K_s ratio less than one indicates negative (purifying) selection where the sequences in the homologous genes are conserved. The various degree of selective pressure that is acting on the PPR motifs, the E- and the DYW-motifs may indicate the significance of each element for *cis*-element sequence recognition and possibly interaction with other editing co-factors.

In the sequence analysis, one species in Brassicaceae family, *Moricandia arvensis* was found to have an unedited C in the *accD* transcript. Thus, I examined *M. arvensis* for the presence of a *RARE1* ortholog, in case it was present but had lost its editing ability for the target C (Figure 3a). A gene fragment orthologous to *RARE1* is present in *M. arvensis*. However, the gene fragment is a pseudogene that has a nucleotide insertion causing frameshifts that create internal stop codons (Figure 5).

Discussion

The plant organellar RNA editing machinery consists of multiple components: the *cis*-elements, which are specific sequences in the chloroplast transcripts surrounding the edited C and the *trans*-acting factors, which are proteins that recognize the *cis*-elements and are likely to recruit other co-factors such as cytidine deaminase to catalyze the deamination reaction (Shikanai 2006). The sequences surrounding the various editing sites do not show apparent similarity to each other apart from the one or two bases immediately surrounding the target sites (Mulligan et al. 1996; Cummings & Myers 2004) which suggests that at least one *trans*-factor is required for editing site recognition. *In vivo* and *in vitro* studies using organelle extracts have identified the *cis*-elements in several chloroplast transcripts that are essential for editing efficiency (reviewed in Shikanai 2006). The *in vitro* editing assay using sequences surrounding the editing target in *accD* C794 in plastid extracts suggests that about 25 nucleotides upstream and 10 nucleotides downstream of the target site are important for editing efficiency (Figure 1). Analysis of the substrates in which each successive 5-nucleotide block was substituted with its complementary nucleotide sequence revealed that the region from -15 to +5 of the target C is critical for editing (Figure 2). The simplest explanation for the absence of editing in the substrates would be that particular nucleotides within the region are important for recognition by the *trans*-factor. Binding of the *trans*-factor to the *cis*-element has been shown in the *ndhD* transcript where CRR4, the first editing *trans*-factor to be identified, preferably binds to its RNA substrate (Okuda et al., 2006). All known editing *trans*-factors are nuclear-encoded and belong to the pentatricopeptide repeat (PPR) protein family.

C-to-U RNA editing may have evolved from a common ancestor in land plants and the editing target in homologous transcripts may have been independently lost throughout evolution

in several species within closely related families (Tillich et al., 2006; Tillich et al., 2009). Therefore, comparisons of *cis*-elements between related species with either a genomically encoded C or T may be informative in understanding the requirement for editing *cis*-element for species requiring editing. From all analyzed *accD* genes in the species across the taxa, several species were found with a genomically encoded T while most species have a genomically encoded C at the corresponding position. There is no obvious pattern that distinguishes between the *cis*-elements in species requiring editing when compared to species with the genomically encoded T. The conservation of the *accD* C794 *cis*-element observed in species across taxa is likely due to conservation of the protein coding region where the editing site is located. The *cis*-element within the *accD* transcript may be more conserved than the flanking sequences of other cytidines within the same transcript, but because the editing site is within the coding region, the rate of evolution in the *accD* C794 *cis*-element could not be discerned. It is likely that an editing site within a non-coding region might be a more useful experimental system in which to observe sequence variation between species with lack of editing when compared to species requiring editing.

I reasoned that species requiring editing at *accD* C794 would conserve in their genomes the gene for the corresponding *trans*-factor, RARE1, in order to allow target site recognition. The interaction of nuclear-encoded *trans*-factors with chloroplast sequences is an example of essential cooperation of nuclear and chloroplast genomes that might constrain the evolution of both genomes. I proposed that when an editing site is lost, the PPR *trans*-factor may lose its function or acquire a new function. This hypothesis was addressed by sequencing various genomic DNA and cDNA of *accD* transcripts in Brassicaceae to obtain a dataset with various editing requirements and *RARE1* homologs. Species with a genomically encoded T have *RARE1*

sequence conservation that is similar to species requiring editing. PPR editing factors in these species may still be functional in recognizing the *cis*-element and recruiting other co-factors. One explanation for the lack of detectable differences in *RARE1* between most editing vs. unedited species is that analysis was performed in closely related species where the genomes not have had sufficient time to undergo evolutionary changes resulting from a loss of a C target of editing.

One one species, *M. arvensis,* has a genomically encoded C at the target site but does not exhibit a T in the cDNA, indicating a lack of RNA editing. The *cis*-element sequence conservation between *M. arvensis* and other species is not sufficient to explain why the target site is not edited. There was no disruption in the *cis*-element that would have suggested that the *trans*-factor might be unable to recognize and bind. However, the homologous *RARE1* sequence was changed in this species by an indel creating a frameshift resulting is multiple stop codons within the gene. The disrupted *RARE1* gene in *M. arvensis* is likely to account for the loss of editing at the corresponding target site in *accD* transcript. If RARE1 normally binds to the *accD* site and recruits an editing complex, lack of expression of a functional RARE1 homolog can explain why the *accD* transcript is not edited.

PPR proteins contain helical repeats that are similar to another protein family, tetratricopeptide repeat (TPR) protein family (Small & Peeters 2000). TPR proteins have been shown to be involved in protein-protein interactions (reviewed in Blatch & Lässle 1999). Nonetheless, with the presence of hydrophilic side-chains and a positively charged groove, the PPR proteins bind to RNA. The putative RNA recognition motif is similar to another helical repeat protein family, the PUF proteins. The human PUF protein, Pumilio1, binds to its RNA target such that each base within the RNA sequence is contacted by a single Pumilio domain/motif (Wang et al. 2001). Thus, the helical repeats of PPR may functionally resemble the

PUF RNA-binding proteins where each PPR motifs may have the ability to bind to specific nucleotide in the *cis*-element. In the preliminary analysis of the *accD cis*-element and RARE1 protein, there was no obvious pattern of sequence and motif features. This analysis is not sufficient to define the RNA-binding properties between the PPR *trans*-factor and the *cis*-element. Nevertheless, the analysis of the Ka/Ks ratio for each PPR motif, and the E- and DYW-motifs in *RARE1* showed that there are differences in the selective pressures that are acting on each motif, suggesting that each feature has a different specificity or interaction state.

Acknowledgment

This work was supported by the National Science Foundation (NSF) grants-MCB0716888 and MCB 0929423. The author thanks the USDA Germplasm Resources Information Network for providing the seeds. The author also thanks the R.M. Mulligan lab at UC Irvine for seeds and plant materials. The author also thanks the Malaysian Government for the fellowship.

REFERENCES

Barkan, A. et al., 1994. A nuclear mutation in maize blocks the processing and translation of several chloroplast mRNAs and provides evidence for the differential translation of alternative mRNA forms. *The EMBO Journal*, 13(13), pp.3170-81.

Beick, S. et al., 2008. The pentatricopeptide repeat protein PPR5 stabilizes a specific tRNA precursor in maize chloroplasts. *Molecular and Cellular Biology*, 28(17), pp.5337-47.

Benne, R. et al., 1986. Major transcript of the frameshifted coxll gene from Trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, 46, pp.819-826.

Bentolila, S., Alfonso, A. & Hanson, M.R., 2002. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16), pp.10887-92.

Bentolila, S., Elliott, L.E. & Hanson, M.R., 2008. Genetic architecture of mitochondrial editing in Arabidopsis thaliana. *Genetics*, 178(3), pp.1693-708.

Bentolila, S., Knight, H. & Hanson M.R., 2010. Natural variation in Arabidopsis leads to the identification of REME1, a pentatricopeptide repeat-DYW protein controlling the editing of mitochondrial transcripts. *Plant Physiology*, 154 (4), pp.1966-82.

Blatch, G.L. & Lässle, M., 1999. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *BioEssays*, 21(11), pp.932-9.

Bock, R., Hermann, M. & Fuchs, M., 1997. Identification of critical nucleotide positions for plastid RNA editing site recognition. *RNA*, 3(10), pp.1194-200.

Bock, R., Kössel, H. & Maliga, P., 1994. Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. *The EMBO Journal*, 13(19), pp.4623-8.

Cai, W. et al., 2009. LPA66 is required for editing *psbF* chloroplast transcripts in Arabidopsis. *Plant Physiology*, 150(3), pp.1260-71.

Chateigner-Boutin, A-L. et al., 2008. CLB19, a pentatricopeptide repeat protein required for editing of *rpoA* and *clpP* chloroplast transcripts. *The Plant Journal*, 56(4), pp.590-602.

Chaudhuri, S & Maliga, P., 1996. Sequences directing C to U editing of the plastid *psbL* mRNA are located within a 22 nucleotide segment spanning the editing site. *The EMBO Journal*, 15(21), pp.5958-64.

Chaudhuri, S., Carrer, H. & Maliga, P., 1995. Site-specific factor involved in the editing of the psbL mRNA in tobacco plastids. *The EMBO Journal*, 14(12), pp.2951-2957.

Chen, S.H. et al., 1987. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science*, 238, pp.363-6.

Covello, P. & Gray, M., 1993. RNA editing in plant mitochondria and chloroplasts. *The FASEB Journal*, 7, pp.64-71.

Cuenca, A. et al., 2010. Are substitution rates and RNA editing correlated? *BMC Evolutionary Biology*, 10(1), p.349.

Cummings, M.P. & Myers, D.S., 2004. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics*, 5, p.132.

Freyer, R., Kiefer-Meyer, M.C. & Kössel, H., 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 94(12), pp.6285-90.

Giegé, P. & Brennicke, A., 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26), pp.15324-9.

Gott, J.M. & Emeson, R.B., 2000. Functions and mechanism of RNA editing. *Molecular Physiology*, (34), pp.499-531.

Hammani, K. et al., 2009. A study of new Arabidopsis chloroplast RNA editing mutants reveals general features of editing factors and their target sites. *The Plant Cell*, 21(11), pp.3686-99.

Handa, H., 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus L.*): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Research*, 31(20), pp.5907-5916.

Hayes, M.L. & Hanson, M.R., 2008. High conservation of a 5' element required for RNA editing of a C target in chloroplast psbE transcripts. *Journal of Molecular Evolution*, 67(3), pp.233-45.

Hayes, M.L. & Hanson, M.R., 2007. Identification of a sequence motif critical for editing of a tobacco chloroplast transcript. *RNA*, 13(2), pp.281-8.

Hayes, M.L. et al., 2006. Sequence elements critical for efficient RNA editing of a tobacco chloroplast transcript in vivo and in vitro. *Nucleic Acids Research*, 34(13), pp.3742-54.

Hegeman, C.E., Hayes, M.L. & Hanson, M.R., 2005. Substrate and cofactor requirements for RNA editing of chloroplast transcripts in Arabidopsis in vitro. *The Plant Journal*, 42(1), pp.124-32.

Heller, W.P., Hayes, M.L. & Hanson, M.R., 2008. Cross-competition in editing of chloroplast RNA transcripts in vitro implicates sharing of trans-factors between different C targets. *The Journal of Biological Chemistry*, 283(12), pp.7314-9.

Jobson, R.W. & Qiu, Y.L., 2008. Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? *Biology Direct*, 3, p.43.

Kotera, E., Tasaka, M. & Shikanai, T., 2005. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature*, 433, pp.326-330.

Librado, P. & Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11), pp.1451-2.

Lurin, C. et al., 2004. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant Cell*, 16, pp.2089-2103.

Maier, R.M. et al., 1996. RNA editing in plant mitochondria and chloroplasts. *Plant Molecular Biology*, 32, pp.343-65.

Miyamoto, T., Obokata, J. & Sugiura, M., 2004. A site-specific factor interacts directly with its cognate RNA editing site in chloroplast transcripts. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), pp.48-52.

Mower, J.P. & Palmer, J.D., 2006. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. *Molecular Genetics and Genomics*, 276(3), pp.285-93.

Mulligan, R.M., Williams, M. & Shanahan, M.T., 1996. RNA editing site recognition in higher plant mitochondria. *The Journal of Heredity*, 90(3), pp.338-44.

Mulligan, R.M., Chang, K.L.C. & Chou, C.C., 2007. Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. *Molecular Biology and Evolution*, 24(9), pp.1971-81.

Nakamura, T. et al., 2003. RNA-binding properties of HCF152, an Arabidopsis PPR protein involved in the processing of chloroplast RNA. *European Journal of Biochemistry*, 270(20), pp.4070-4081.

Notsu, Y. et al., 2002. The complete sequence of the rice (*Oryza sativa L.*) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Molecular Genetics and Genomics*, 268(4), pp.434-45.

Okuda, K. et al., 2009. Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. *The Plant Cell*, 21(1), pp.146-56.

Okuda, K. et al., 2008. Amino acid sequence variations in Nicotiana CRR4 orthologs determine the species-specific efficiency of RNA editing in plastids. *Nucleic Acids Research*, 36(19), pp.6155-64.

Okuda, K. et al., 2007. Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19), pp.8178-83.

Okuda, K. et al., 2006. A pentatricopeptide repeat protein is a site recognition factor in chloroplast RNA editing. *The Journal of Biological Chemistry*, 281(49), pp.37661-7.

O'Toole, N. et al., 2008. On the expansion of the pentatricopeptide repeat gene family in plants. *Molecular Biology and Evolution*, 25(6), pp.1120-8.

Pfalz, J. et al., 2009. Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *The EMBO Journal*, 28(14), pp.2042-52.

Picardi, E. & Quagliariello, C., 2008. Is plant mitochondrial RNA editing a source of phylogenetic incongruence? An answer from in silico and in vivo data sets. *BMC Bioinformatics*, 9(Suppl 2), p.S14.

Powell, L.M. et al., 1987. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell*, 50(6), pp.831-40.

Reed, M.L., Peeters, N.M. & Hanson, M.R., 2001. A single alteration 20 nt 5' to an editing target inhibits chloroplast RNA editing in vivo. *Nucleic Acids Res.*, 29(7), pp.1507-1513.

Robbins, J.C., Heller, W.P. & Hanson, M.R., 2009. A comparative genomics approach identifies a PPR-DYW protein that is essential for C-to-U editing of the Arabidopsis chloroplast *accD* transcript. *RNA*, 15(6), pp.1142-53.

Salmans, M.L. et al., 2010. Editing site analysis in a gymnosperm mitochondrial genome reveals similarities with angiosperm mitochondrial genomes. *Current Genetics*, 56(5), pp.439-46.

Schmitz-Linneweber, C. & Small, I., 2008. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science*, 13(12), pp.663-70.

Schmitz-Linneweber, C. et al., 2006. A pentatricopeptide repeat protein facilitates the trans-splicing of the maize chloroplast *rps12* pre-mRNA. *The Plant Cell*, 18(10), pp.2650-63.

Schmitz-Linneweber, C., Williams-Carrier, R. & Barkan, A., 2005. RNA immunoprecipitation and microarray analysis show a chloroplast pentatricopeptide repeat protein to be associated with the 59 Region of mRNAs whose translation it activates. *The Plant Cell*, 17, pp.2791-2804.

Shikanai, T., 2006. RNA editing in plant organelles: machinery, physiological function and evolution. *Cellular and Molecular Life Sciences*, 63(6), pp.698-708.

Small, I.D. & Peeters, N., 2000. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. *Trends in Biochemical Sciences*, 25(2), pp.46-7.

Smith, H.C., Gott, J.M. & Hanson, M.R., 1997. A guide to RNA editing. *RNA*, 3(10), pp.1105-23.

Sung, T.Y., Tseng, C.C. & Hsieh, M.H., 2010. The SLO1 PPR protein is required for RNA editing at multiple sites with similar upstream sequences in Arabidopsis mitochondria. *The Plant Journal*, pp.499-511.

Takenaka, M., 2010. MEF9, an E-subclass pentatricopeptide repeat protein, is required for an RNA editing event in the *nad7* transcript in mitochondria of Arabidopsis. *Plant Physiology*, 152(2), pp.939-47.

Takenaka, M. et al., 2010. Reverse genetic screening identifies five E-class PPR proteins involved in RNA editing in mitochondria of *Arabidopsis thaliana*. *The Journal of Biological Chemistry*, 285(35), pp.27122-9.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Ne,i M., Kumar, S., (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distances, and maximum parsimony methods. *Molecular Biology and Evolution* (submitted).

Thompson, J.D., Higgins, D.G. & Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), pp.4673-80.

Tillich, M. et al., 2006. The evolution of chloroplast RNA editing. *Molecular Biology and Evolution*, 23(10), pp.1912-21.

Tillich, M. et al., 2009. Loss of *matK* RNA editing in seed plant chloroplasts. *BMC Evolutionary Biology*, 9, p.201.

Tseng, C.C. et al., 2010. Editing of accD and ndhF chloroplast transcripts is partially affected in the Arabidopsis vanilla cream1 mutant. *Plant Molecular Biology*, 73(3), pp.309-23.

Verbitskiy, D. et al., 2011. The DYW-E-PPR protein MEF14 is required for RNA editing at site matR-1895 in mitochondria of Arabidopsis thaliana. *FEBS letters*, 585(4), pp.700-4.

Verbitskiy, D. et al., 2010. The PPR protein encoded by the LOVASTATIN INSENSITIVE 1 gene is involved in RNA editing at three sites in mitochondria of *Arabidopsis thaliana*. *The Plant Journal*, 61(3), pp.446-55.

Wakasugi, T., Tsudzuki, T. & Sugiura, M, 2001. The genomics of land plant chloroplasts: Gene content and alteration of genomic information by RNA editing. *Photosynthesis Research*, 70(1), pp.107-18.

Wang, X., Zamore, P.D. & Hall, T.M., 2001. Crystal structure of a Pumilio homology domain. *Molecular Cell*, 7(4), pp.855-65.

Yu, Q.B. et al., 2009. AtECB2, a pentatricopeptide repeat protein, is required for chloroplast transcript accD RNA editing and early chloroplast biogenesis in *Arabidopsis thaliana*. *The Plant Journal*, 59(6), pp.1011-23.

Zehrmann, A. et al., 2008. Seven large variations in the extent of RNA editing in plant mitochondria between three ecotypes of *Arabidopsis thaliana*. *Mitochondrion*, 8, pp.319-327.

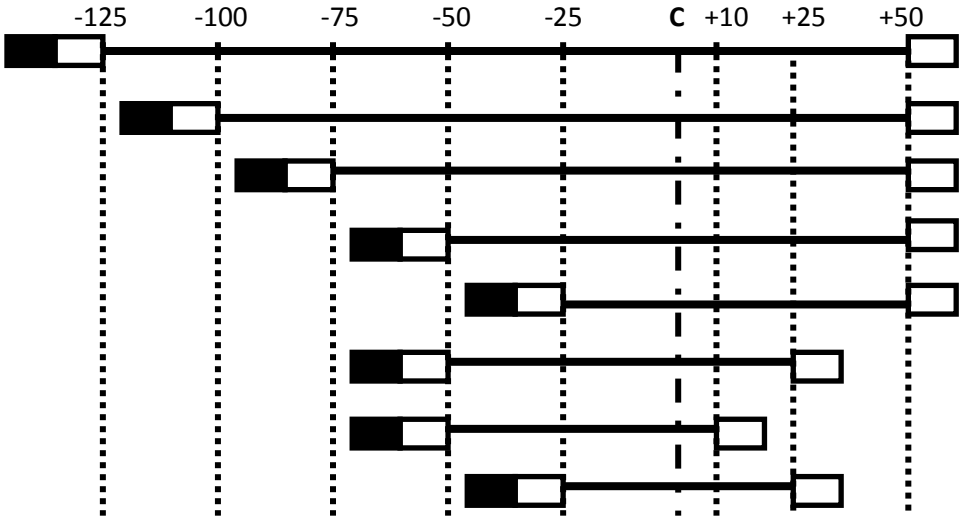
Zehrmann, A. et al., 2009. A DYW domain–containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of Arabidopsis thaliana. *The Plant Cell*, 21(2), pp.558-567.

Zhou, W. et al., 2008. The Arabidopsis gene YS1 encoding a DYW protein is required for editing of *rpoB* transcripts and the rapid development of chloroplasts during early growth. *The Plant Journal*, pp.82-96.

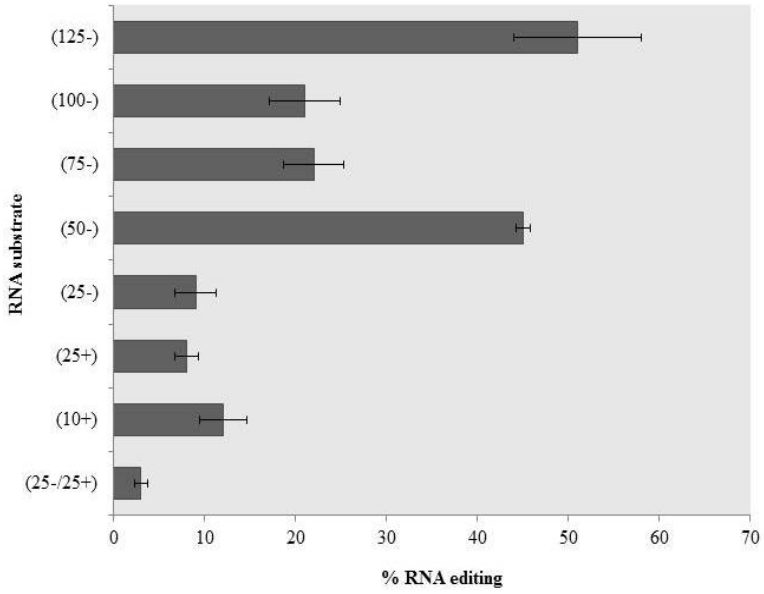
LIST OF FIGURES

Figure 1.

a)



b)



c)

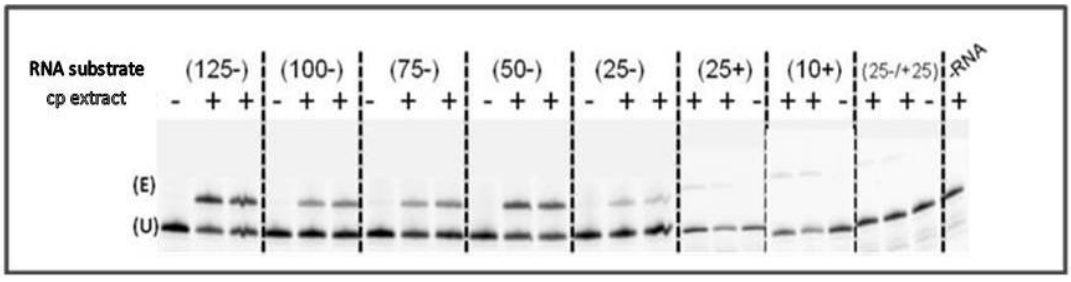
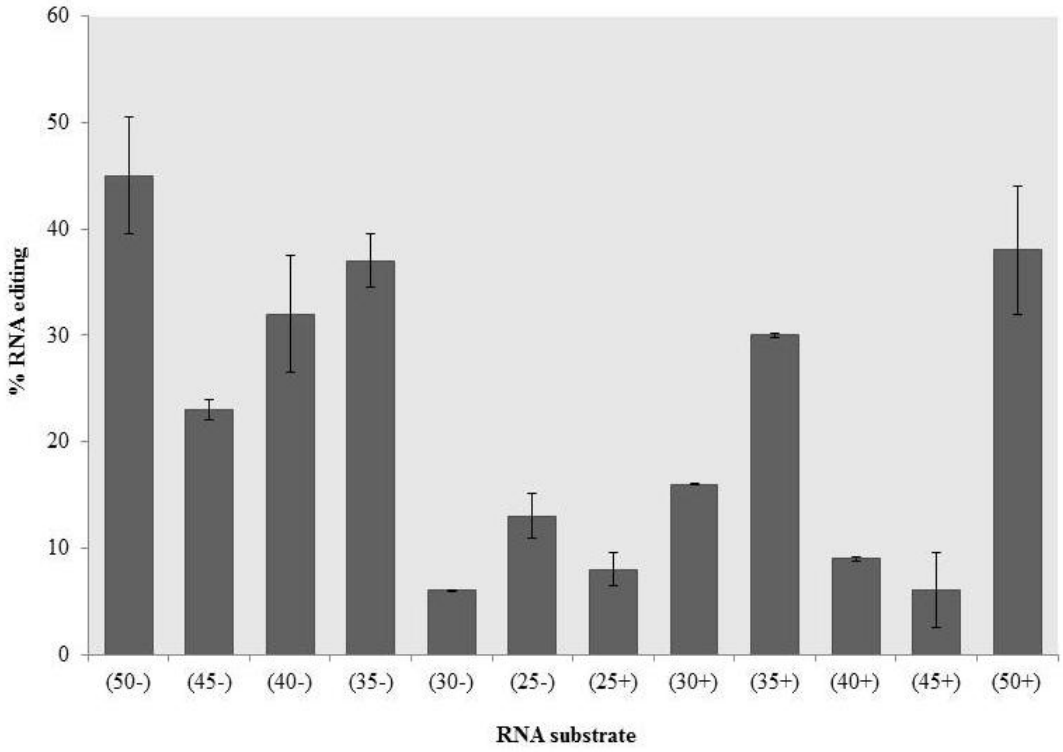


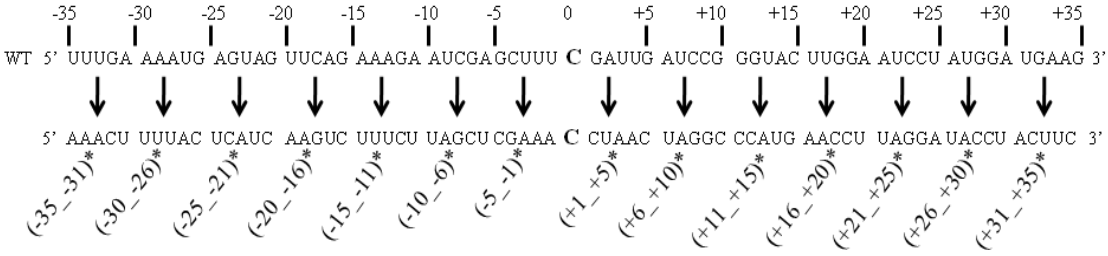
Figure 1. Editing efficiency *in vitro* of *Arabidopsis accD* C794 substrates with varying sequence lengths flanking the target C. (a) Diagram of subset of RNA substrates constructed with varying length of sequences upstream and downstream of the target C. *Open boxes* indicate sequences used for universal amplification of substrate during RT-PCR, SK on 5' end and KS on 3' end. *Black* boxes indicate the T7 promoter sequence used for *in vitro* transcription. (b) Editing efficiency for each substrates shown in (a). *Error bars* represent 1 S.D from the mean in replicate samples. (c) Polyacryalamide gel of poisoned primer extension (PPE) assay for truncation substrates listed in (a). *(E)* represents the edited substrates and *(U)* represent the unedited substrates.

Figure 2.

a)


60
50
40
30
20
10
0
% RNA editing
(50-)
(45-)
(40-)
(35-)
(30-)
(25-)
(25+)
(30+)
(35+)
(40+)
(45+)
(50+)
RNA substrate

b)


-35 -30 -25 -20 -15 -10 -5 0 +5 +10 +15 +20 +25 +30 +35
WT 5' UUUGA AAAUG AGUAG UUCAG AAAGA AUCGAGCUUU C GAUUG AUCCG GGUAC UUGGA AUCCU AUGGA UGAAG 3'
5' AAACU UUUAC UCAUC AAGUC UUUCU UAGCU CGAAA C CUAAC UAGGC CCAUG AACCU UAGGA UACCU ACUUC 3'
(-35_-31)*
(-30_-26)*
(-25_-21)*
(-20_-16)*
(-15_-11)*
(-10_-6)*
(-5_-1)*
(+1_+5)*
(+6_+10)*
(+11_+15)*
(+16_+20)*
(+21_+25)*
(+26_+30)*
(+31_+35)*

c)

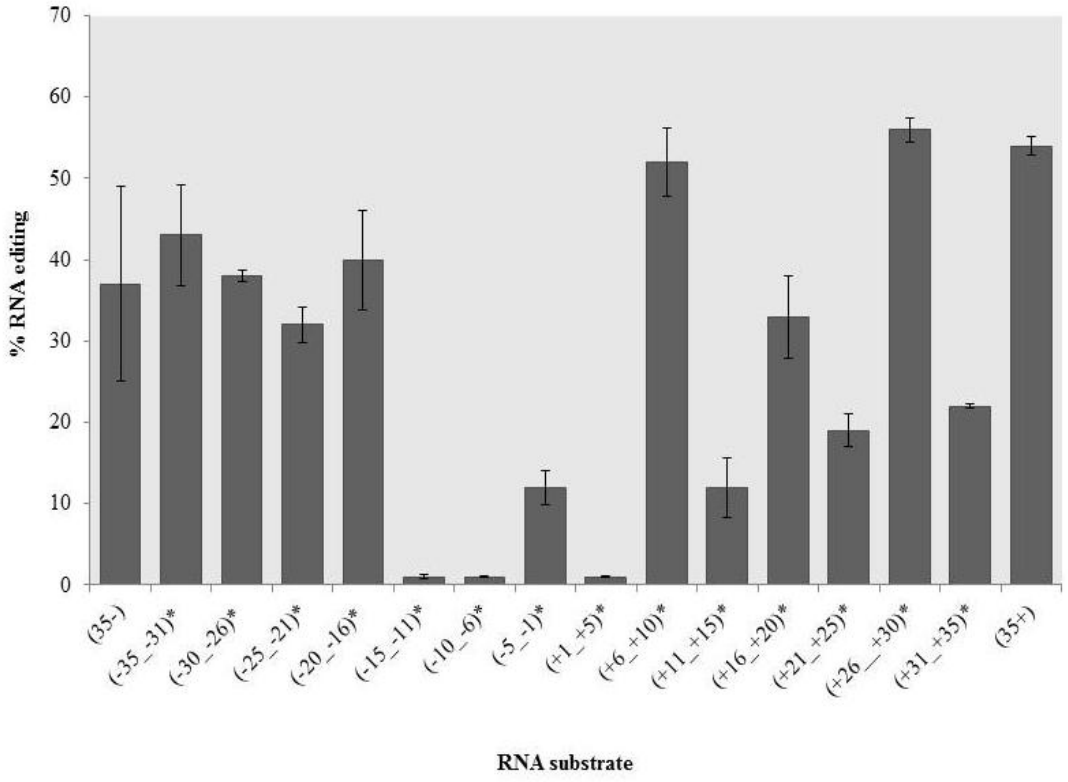


Figure 2. Editing efficiency *in vitro* of *Arabidopsis accD* C794 substrates with every 5 nt-truncated and mutated sequences flanking the target C. (a) Editing efficiency for substrates with truncation at every 5 nucleotides starting from 50 nucleotides upstream or downstream from the target C. (b) Editing efficiency for substrates with every 5 nucleotides mutated from the wild-type sequence to the complementary nucleotides on either directions from the target C. *Error bars* represent 1 S.D from the mean in replicate samples.

Figure 3.

a)

	794										
A.thaliana gDNA	…AGA	ATC	GAG	CTT	T**C**G	ATT	GAT	CCG	GGT…		
A.thaliana cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
T.arvense cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CCA	GGT…		
L.sativum cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
N.officinale cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
A.rusticana cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
D.californica cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
H.matronalis cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CGG	GGT…		
B.oleracea cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CTG	GGT…		
R.rugosum cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
L.annua cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
A.petiolata cDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CGG	GGT…		
R.sativus cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
M.incana cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CGG	GGT…		
B.rapa cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
C.hispanica cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CGG	GGT…		
S.arvensis cDNA	…AGA	ATT	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
M.arvensis cDNA	…AGA	ATC	GAG	CTT	T**C**G	ATT	GAT	CCG	GGT…		
D.nemorosa gDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		
L.maritima gDNA	…AGA	ATC	GAG	CTT	T**T**G	ATT	GAT	CCG	GGT…		

794

b) 

c) 

d) 

e) 

f) 

g)

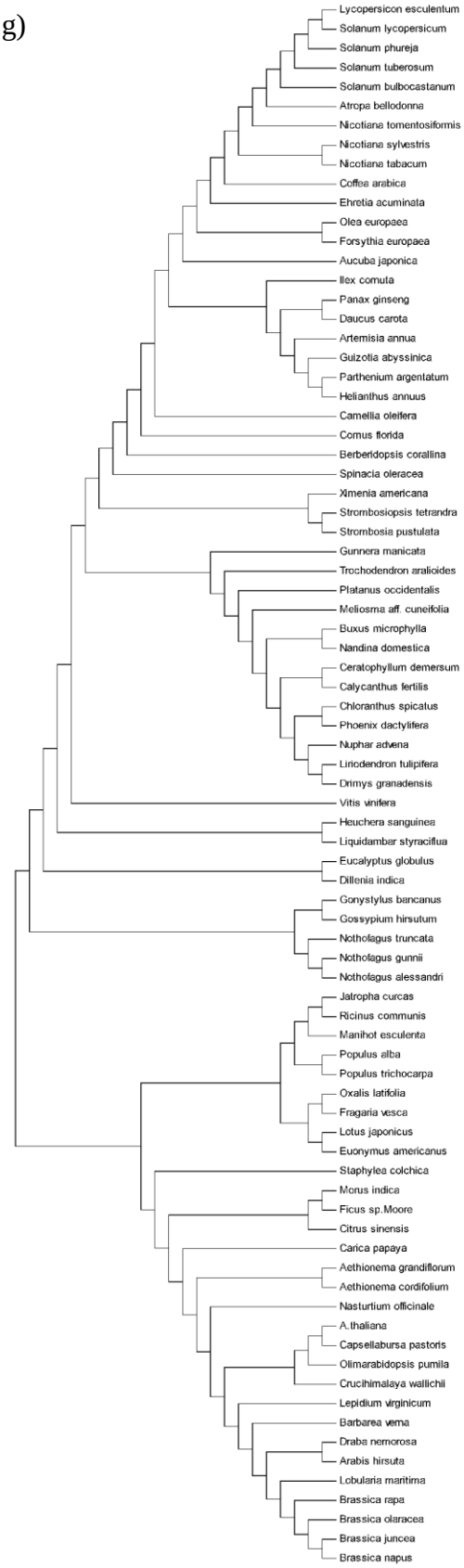
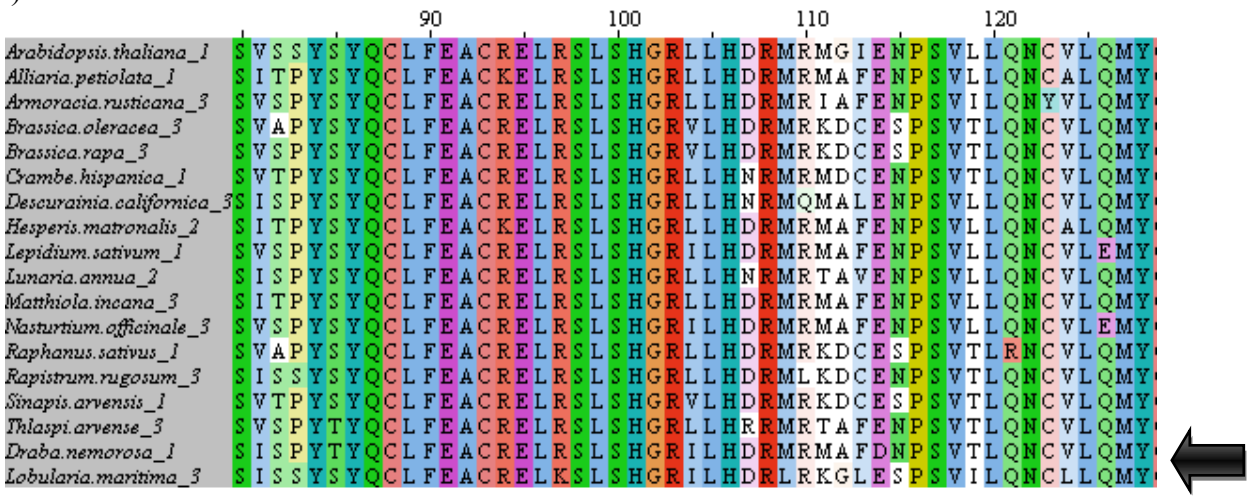

Lycopersicon esculentum
Solanum lycopersicum
Solanum phureja
Solanum tuberosum
Solanum bulbocastanum
Atropa bellodonna
Nicotiana tomentosiformis
Nicotiana sylvestris
Nicotiana tabacum
Coffea arabica
Ehretia acuminata
Olea europaea
Forsythia europaea
Aucuba japonica
Ilex comuta
Panax ginseng
Daucus carota
Artemisia annua
Guizotia abyssinica
Parthenium argentatum
Helianthus annuus
Camellia oleifera
Cornus florida
Berberidopsis corallina
Spinacia oleracea
Ximenia americana
Strombosiopsis tetrandra
Strombosia pustulata
Gunnera manicata
Trochodendron aralioides
Platanus occidentalis
Meliosma aff. cuneifolia
Buxus microphylla
Nandina domestica
Ceratophyllum demersum
Calycanthus fertilis
Chloranthus spicatus
Phoenix dactylifera
Nuphar advena
Liriodendron tulipifera
Drimys granadensis
Vitis vinifera
Heuchera sanguinea
Liquidamber styraciflua
Eucalyptus globulus
Dillenia indica
Gonystylus bancanus
Gossypium hirsutum
Notholagus truncata
Notholagus gunnii
Notholagus alessandri
Jatropha curcas
Ricinus communis
Manihot esculenta
Populus alba
Populus trichocarpa
Oxalis latifolia
Fragaria vesca
Lotus japonicus
Euonymus americanus
Staphylea colchica
Morus indica
Ficus sp.Moore
Citrus sinensis
Carica papaya
Aethionema grandiflorum
Aethionema cordifolium
Nasturtium officinale
A.thaliana
Capsellabursa pastoris
Olimarabidopsis pumila
Crucihimalaya wallichii
Lepidium virginicum
Barbarea vema
Draba nemorosa
Arabis hirsuta
Lobularia maritima
Brassica rapa
Brassica olaracea
Brassica juncea
Brassica napus

Figure 3. **Sequence alignment of *accD* C794 *cis*-element.** (a) Sequence alignment for 15 nucleotides upstream and 10 nucleotides downstream of target C for 19 species in Brassicaceae family. *gDNA* represents the sequence from genomic DNA of the species. *cDNA* represents the sequence obtained from the *accD* transcripts in the species. *794* → indicates the nucleotide position at 794 in *Arabidopsis accD* gene and the corresponding position in other Brassicaceae species. (b) Sequence logo for *accD* C794 *cis*-element in Brassicaceae species with genomically encoded C. (c) Sequence logo for *accD* 794 *cis*-element in Brassicaceae species with genomically encoded T at the corresponding position.(d) Sequence logo for *accD* C794 *cis*-element for species in the dataset with genomically encoded C. (e) Sequence logo for *accD* 794 *cis*-element for species in the dataset with genomically encoded T. (f) Sequence logo for *accD* 794 *cis*-element for all analyzed species across the taxa in the dataset (Table S3). (g) Cladogram of species in the dataset for *accD* sequence analysis.

Figure 4.

a)



b)

	K_a/K_s *ratio*
Full RARE1 in Brassicaceae species with genomically encoded T at 794 (2 species)	0.11
Full *RARE1* in Brassicaceae species requiring editing at C794 (15 species)	0.13
PPR motif 1	0.06
PPR motif 2	0.14
PPR motif 3	0.15
PPR motif 4	0.34
PPR motif 5	0.49
PPR motif 6	0.24
PPR motif 7	0.31
PPR motif 8	0.11
PPR motif 9	0.10
PPR motif 10	0.14
PPR motif 11	0.14
PPR motif 12	0.25
PPR motif 13	0.14
PPR motif 14	0.26
PPR motif 15	0.14
E-motif	0.15
DYW-motif	0.10

Figure 4. The Ka/Ks ratio for *RARE1* in Brassicaceae species. Table for the K_a/K_s ratio corresponding to the subset of species in the dataset. The K_a/K_s ratios were obtained for Brassicaceae species with a genomically encoded T in the *accD* transcript and species with a genomically encoded C in the *accD* C794 transcript. The K_a/K_s ratios were also obtained for the PPR motifs, E- and DYW-motifs for Brassicaceae species with a genomically encoded C in the *accD* transcript. The ratio of less than 1 indicates negative selection, ratio of more than 1 indicates positive selection and ratio of 1 indicates natural or neutral selection is acting on the gene sequence.

Figure 5.

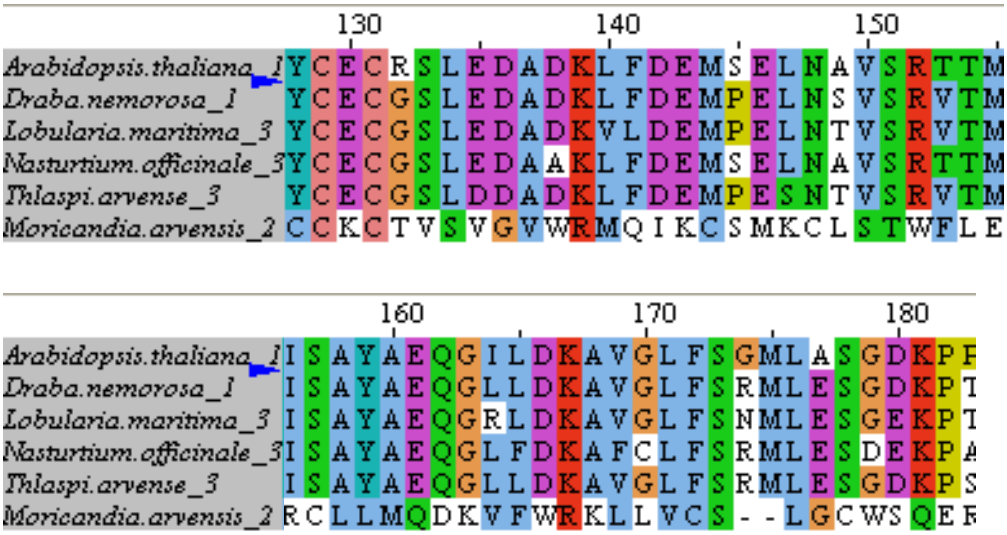


Figure 5. *RARE1* pseudogene is present in *Moricandia arvensis,* a Brassicaceae species which has lost editing ability at *accD* C794. Protein sequence alignment in *M. arvensis* shows more sequence variation when compared to other species in Brassicaceae family with pre-edited T or edited C. An indel which causes frameshift in the gene sequence creates internal stop codon.

LIST OF SUPPLEMENTAL TABLES

Table S1. The following oligonucleotides (Integrated DNA Technologies) were used in this study

Primer	**Sequence**
accD Forward	5' TGTGGATTCAATGCGACAAT
accD Reverse	5' GGAAGTTCAAATTAGACTAGACAAAC
RARE1 Forward	5' TCCATCAACTATGACGATTCTCACTG
RARE1 Reverse	5' CAACGATTACTGGTGACTGATGATCT
RARE1 int repeat1 Forward	5' CCGTCGTGGGTTTCTCTGAAGTC
RARE1 int Reverse	5' GGTGATAAACACCATCCACAAAC
T7SK	5' TAATACGACTCACTATAGGCGCTCTAGAACTAGTGGATC
SK	5' CGCTCTAGAACTAGTGGATC
KS	5' TCGAGGTCGACGGTATC
accD 125-	5' CTACAATCAATTGTGGATTCAATGCG
accD 100-	5' GACAATTGTTATGGATTAAT
accD 75-	5' AGAAAGTCAAAATGAATGTT
accD 50-	5' ACAATGTGGACATTATTTGA
accD 45-	5' GTGGACATTATTTGAAAATGAGTAG
accD 40-	5' CATTATTTGAAAATGAGTAGTTCAG
accD 35-	5' TTTGAAAATGAGTAGTTCAGAAAG
accD 30-	5' AAATGAGTAGTTCAGAAAGAATCG
accD 25-	5' AGTAGTTCAGAAAGAATCGA
accD 10+	5' TCGAGCTTTCGATTGATCCG
accD 25+	5' ATCCGGGTACTTGGAATCCT
accD 30+	5' CCGGGTACTTGGAATCCTATGGA
accD 35+	5' GGTACTTGGAATCCTATGGATGAAG
accD 40+	5' GGAATCCTATGGATGAAGACATG
accD 45+	5' CTATGGATGAAGACATGGTCTC
accD 50+	5' TGAAGACATGGTCTCTGCGG
accD (-35_-31)	5' AAACTAAATGAGTAGTTCAGAAAGA
accD (-30_-26)	5' TTTGATTTACAGTAGTTCAGAAAGAATCGA
accD (-25_-21)	5' TTTGAAAATGTCATCTTCAGAAAGAATCGAGCTTT
accD (-20_-16)	5' TTTGAAAATGAGTAGAAGTCAAAGAATCGAGCTTT

accD (-15_-11)	5' TTTGAAAATGAGTAGTTCAGTTTCTATCGAGCTTT
accD (-10_-6)	5' AGTAGTTCAGAAAGATAGCTGCTTTCGATTGATCCGGGTA
accD (-5_-1)	5' AGTAGTTCAGAAAGAATCGACGAAACGATTGATCCGGGTA
accD (+1_+5)	5' AAAGAATCGAGCTTTCCTAACATCCGGGTA CTTGGAATCCTATGGAT
accD (+6_+10)	5' AAAGAATCGAGCTTTCGATTGTAGGCGGTA CTTGGAATCCTATGGAT
accD (+11_+15)	5' ATCCGCCATGTTGGAATCCTATGGATGAAG
accD (+16_+20)	5' ATCCGGGTACAACCTATCCTATGGATGAAG
accD (+21_+25)	5' ATCCGGGTACTTGGATAGGAATGGATGAAG
accD (+26_+30)	5' ATCCGGGTACTTGGAATCCTTACCTTGAAG
accD (+31_+35)	5' GGTACTTGGAATCCTATGGAACTTC
accD C794 PPE-G	5' CCATAGGATTCCAAGTACCCGG
accD C794 PPE-C	5' TGAGTAGTTCAGAAAGAATCGAGC

Table S2. The following plants belong to the Brassicaceae family were used in the *accD* and *RARE1* sequence analyses.

Species	**Accession**		**Species**	**Accession**
Alliaria petiolata	Ames 16096		*Lepidium sativum*	Ames 29183
Arabidopsis thaliana	Local		*Lobularia maritima*	PI 288262
Armoracia rusticana	Local		*Lunaria annua*	PI 279717
Brassica oleracea	Local		*Matthiola incana*	Ames 26267
Brassica rapa	Local		*Moricandia arvensis*	GR 623
Crambe hispanica	PI 337996		*Nasturtium officinale*	NSL 69920
Descurainia californica	W6 30819		*Raphanus sativus*	PI 121031
Draba nemorosa	Local		*Rapistrum rugosum*	PI 388816
Hesperis matronalis	PI 443300		*Sinapis arvensis*	Ames 19280
			Thlaspi arvense	Ames 29513

Table S3. Species that were included in dataset for *accD cis*-element and flanking sequence analyses.

Species	Family	Order
Lycopersicon esculentum	Solanaceae	Solanales
Solanum lycopersicum	Solanaceae	Solanales
Solanum phureja	Solanaceae	Solanales
Solanum tuberosum	Solanaceae	Solanales
Solanum bulbocastanum	Solanaceae	Solanales
Atropa belladonna	Solanaceae	Solanales
Nicotiana tomentosiformis	Solanaceae	Solanales
Nicotiana sylveris	Solanaceae	Solanales
Nicotiana tabacum	Solanaceae	Solanales
Coffea arabica	Rubiaceae	Gentianales
Ehretia acuminata	Boraginaceae	-
Olea europae	Oleaceae	Lamiales
Forsythia europe	Oleaceae	Lamiales
Aucuba japonica	Garryaceae	Garryales
Ilex cornuta	Aquifoliaceae	Aquifoliales
Panax ginseng	Araliaceae	Apiales
Daucus carota	Apiaceae	Apiales
Artermisia annua	Asteraceae	Asterales
Guizotia abyssinica	Asteraceae	Asterales
Parthenium argentatum	Asteraceae	Asterales
Helianthus annuus	Asteraceae	Asterales
Camellia oleifera	Theaceae	Ericales
Cornus florida	Cornaceae	Cornales
Berberidopsis corallina	Berberidopsidaceae	Berberidopsidales
Spinacia oleracea	Chenopodiaceae	Caryophyllales
Ximenia americana	Olacaceae	Santalales
Strombosiopsis tetrandra	Olacaceae	Santalales
Strombosia pustulata	Olacaceae	Santalales
Guinerra manicata	Gunneraceae	Gunnerales
Trochodendron aralioides	Trochodendraceae	Trochodendrales
Platanus occidentalis	Platanaceae	Proteales
Meliosma aff. Cuneifolia	Sabiaceae	-

Buxus microphylla	Buxaceae	Buxales
Nandina domestica	Berberidaceae	Ranunculales
Ceratophyllum demersum	Ceratophyllaceae	Ceratophyllales
Calycanthus fertilis	Calycanthaceae	Laurales
Chloranthus spicatus	Chloranthaceae	Chloranthales
Phoenix dactylifera	Arecaceae	Arecales
Nuphar advena	Nymphaeaceae	Nymphaeales
Liriodendron tulipifera	Magnoliaceae	-
Drimys granadensis	Winteraceae	Canellales
Vitis vinifera	Vitaceae	Vitales
Heuchera sanguinea	Saxifragaceae	Saxifragales
Liquidambar styraciflua	Altingiaceae	Saxifragales
Eucalyptus globulus	Myrtaceae	Myrtales
Dillenia indica	Dilleniaceae	-
Gonystylus bancanus	Thymelaeaceae	Malvales
Gossypium hirsutum	Malvaceae	Malvales
Nothofagus truncata	Nothofagaceae	Fagales
Nothofagus gunii	Nothofagaceae	Fagales
Nothofagus alessandri	Nothofagaceae	Fagales
Jatropha curcas	Euphorbiaceae	Malpighiales
Ricinus communis	Euphorbiaceae	Malpighiales
Manihot esculenta	Euphorbiaceae	Malpighiales
Populus alba	Salicaceae	Malpighiales
Populus trichocarpa	Salicaceae	Malpighiales
Oxalis latifolia	Oxidaceae	Oxidales
Fragaria vesca	Rosaceae	Rosales
Lotus japonicus	Fabaceae	Fabales
Euonymus americanus	Celastraceae	Celastrales
Staphylea colchica	Staphyleaceae	Crossosomatales
Morus indica	Moraceae	-
Ficus sp. Moore	Moraceae	-
Citrus sinensis	Rutaceae	Sapindales
Carica papaya	Caricaceae	Brassicales
Aethionema grandiflorum	Brassicaceae	Brassicales
Aethionema cordifolium	Brassicaceae	Brassicales
Nasturtium officinale	Brassicaceae	Brassicales

Arabidopsis thaliana	Brassicaceae	Brassicales
Capsella bursa-pastoris	Brassicaceae	Brassicales
Olimarabidopsis pumila	Brassicaceae	Brassicales
Crucihimalaya wallichii	Brassicaceae	Brassicales
Lepidium virginicum	Brassicaceae	Brassicales
Barbarea verna	Brassicaceae	Brassicales
Draba nemorosa	Brassicaceae	Brassicales
Arabis hirsuta	Brassicaceae	Brassicales
Lobularia maritima	Brassicaceae	Brassicales
Brassica rapa	Brassicaceae	Brassicales
Brassica oleracea	Brassicaceae	Brassicales
Brassica juncea	Brassicaceae	Brassicales
Brassica napus	Brassicaceae	Brassicales