Investigation of Tissue Specific Alternative Splicing of RNA Editing Related Family Proteins in *Arabidopsis thaliana*

Umme Qulsum

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Investigation of Tissue Specific Alternative Splicing of RNA Editing Related Family Proteins in *Arabidopsis thaliana*

Umme Qulsum

Supervisor: Professor Dr. Toshifumi Tsukahara

School of Materials Science

Japan Advanced Institute of Science and Technology

March, 2019

Abstract

Arabidopsis is most useful model plants in molecular biology. RNA editing is a post-transcriptional modification of genes that commonly occur in plant plastids and mitochondria. Alternative splicing is a post and co-transcriptional regulation of gene expression. Pentatricopeptide repeat (PPR) family proteins were recently found to be involved in RNA editing in plants. The aim of this study was to investigate the tissue-specific expression and alternative splicing of *PPR* family genes and their effects on PPR motif and functionality. Of the 27 *PPR* genes in *Arabidopsis thaliana*, I selected six *PPR* genes of the P subfamily that are likely alternatively spliced, which were confirmed by sequencing. Four of these genes show intron retention, and the two remaining genes have 3' alternative-splicing sites. Alternative-splicing events occurred in the coding regions of three genes and in the 3' UTRs of the three remaining genes. I also identified five previously unannotated alternatively spliced isoforms of these *PPR* genes, which were confirmed by PCR and sequencing. Among these, three contain 3' alternative-splicing sites, one contains a 5' alternative-splicing site, and the remaining gene contains a 3'-5' alternative-splicing site. The new isoforms of two genes affect protein, and three other alternative-splicing sites are located in 3' UTRs. These findings suggest that tissue-specific expression of different alternatively spliced transcripts occurs in Arabidopsis, even at different developmental stages.

Recently it has been revealed that, not only PPR family proteins but also other additional family proteins MORF/RIP, ORRM and OZ are involved in RNA editing. The aim of this study is to find out the tissue-specific expression and alternative splicing of *ZnF* family genes and their effect on protein and functionality. Out of 25 *ZnF* genes, I randomly selected seven which are probably alternatively spliced and most of the genes are located in protein coding region which is determined using Arabidopsis database. Among these, alternative splicing in 7 genes of *ZnF* family was confirmed by sequencing. Out of which five genes with intron retention, one gene with 3' alternative splice site and another one genes exon skipping were detected. Alternative splicing events were located in six genes in the coding region and one gene in 3' UTR region. Here I also reported three unannotated and new alternatively spliced isoforms from these *ZnF* genes that were confirmed by PCR and sequencing. Among these, one is with 3' alternative splice site and two with intron retention. New unannotated isoforms affecting protein in one gene and another one alternate splice located in 3' UTR region. This study suggests that tissue-specific expression of different alternatively spliced transcript happen even in different developmental stages.

RNA editing illustrated as any site-specific alteration in RNA sequences containing insertion or deletion and base substitution and has been broadly investigated in animals. In plant, RNA editing is a post-transcriptional modification of genes that commonly occur in plastids and mitochondria. In case of flowering plants, it is reported that not only PPR but also non-PPR proteins like MORF/RIP, ORRM and OZ partake in diverse RNA editing complex. Previously predicted 12 types RNA editing patterns may exist in the nuclear transcript, chloroplast and mitochondria in Arabidopsis. In the course of study of alternative splicing, tissue-specific RNA editing events were found in RNA editing related family genes. I collected samples of different tissues of different developmental stages from Arabidopsis. Such as seedling (whole plant) 4, 8, 12 days; 16, 21, 27 and 32 days old leaf, stem and root. Extraction of total RNA, cDNA synthesis and PCR were performed. After PCR, the targeted band was cut from PAGE then sequencing was performed. I found 9 types of RNA editing events these are C-to-U, U-to-C, A-to-I(G), A-to-C, A-to-U, G-to-A, G-to-C, U-to-A and U-to-G in targeted genes. Most of the editing events in seedling and leaf and less in stem tissues. Extensive editing U-to-C (60%) was detected in seedling 12 days, A-to-I(G) (54%) in leaf 21 days. This is the first experimental report that RNA editing could be regulated in tissue and development specific manner. During plant development, RNA editing machinery may play important role in proteins diversity and functionality thus ultimately affecting plant physiology.

Keywords: RNA editing, Alternative splicing, PPR, Zinc-finger motif, RNA editing events.

CONTENTS

Chapter I

General Introduction

1.1 Post transcriptional modification

1.1.1 Transcription

Genetic information is encoded in DNA in the form of nucleotide sequence. Chromosomes are tightly packed DNA present inside nucleus. Whereas, translation process occurs in the cytosol. So, it is obvious that the genetic information from nucleus is transported to cytosol where protein is synthesized. Transcription is a process where genetic information encoded in DNA is then transcribed using RNA polymerase into RNA which is then transported into cytosol after required post transcriptional modification. In eukaryotes, transcription takes place through an enzymatic process which requires RNA polymerase II. This process involves three main stages initiation, elongation and termination (Alberts et al. 2007). During initiation phase, high proof reading RNA polymerase binds to specific DNA sequence region also called as promoter region where specific proteins called transcription factors are recruited and they unwind DNA to allow RNA polymerase to initiate transcription of dsDNA into ssRNA which is then followed by elongation phase. During elongation RNA polymerase progresses along DNA template in 3' to 5' direction and synthesis of complementary RNA molecule in 5' to 3' direction takes place. Termination occurs when RNA polymerase reaches termination sequence and recruits Rho factors. After termination RNA polymerase falls off the DNA, mRNA is released. Inside cells, transcription is highly regulated. The expression of transcripts is differentially regulated with a differentiation in a tissue specific manner (Shinozaki et al. 1999).

The pre-mRNA undergoes certain post transcriptional modifications through which it becomes functional mRNA, this process is termed as mRNA maturation and involves

three major events namely 5'-capping, 3' cleavage/polyadenylation and RNA splicing. 5' and 3' end modifications of RNA play an important role to prevent degradation of mature RNA by various RNases. These post transcriptional modifications makes the mature RNA different from other kinds of RNAs. These post transcriptional modifications occur simultaneously as the pre mRNA is being synthesized inside the nucleus.

1.1.2 5' capping of the RNA

Shortly after DNA transcription is initiated, a protective cap RNA 7-methylguanosine / m7G cap is enzymatically added to the 5' end of the nascent pre mRNA. During this process Guanylyltransferase adds a guanine residue to the 5' end of RNA and methyl transferase adds a methyl group. The Capping of RNA occurs after RNA polymerase reaches a length of 20-25 nucleotides. The transcription and RNA capping occur simultaneously and therefore, they influence each other. 5' capping is catalyzed by recruiting a dimeric capping enzyme Guanylyltransferase and its activation by binding to Ser5- phosphorylated carboxyl-terminal domain (CTD) of RNA polymerase II. Capping enzyme does not associate with other RNA polymerases like RNA polymerase I and III, but is highly specific to RNA transcripts synthesized by RNA polymerase II. Capping helps mRNA protection against RNase or exoribonuclease and enhances translation, splicing, export into cytoplasm.

1.1.3 3' polyadenylation

Polyadenylation also termed as A-tailing is an important mechanism in eukaryotic mRNA, but not in histone mRNA. The poly (A) tail which is the terminal sequence is added to the end of RNA at the 3' end is actually not coded from the DNA. During polyadenylation about 150-200 adenylate nucleotide residues are added by poly (A)

polymerase. This is after the pre-mRNA is cleaved at a defined site. Presence of specific poly (A) signal sequence AAUAAA at near the 3' end of mRNA recruits different specific transcription factors which initiates cleavage of mRNA. The cleavage of the nascent mRNA is catalyzed by the enzyme CPSF (cleavage polyadenylation specificity factor) which binds to poly (A) signal. Other factors such as CstF (Cleavage stimulation Factor) and CFI and CFII (Cleavage Factors I and II) are also involved in this process (Lutz et al. 2008). The poly (A) polymerase adds first 12A residues to 3'-OH slowly due to the loose attachment of the enzyme. Once the binding of poly (A) binding protein II /PABPII to the small poly A tail, it increases the binding capacity of the poly (A) polymerase and rapidly catalyzes the addition of approximately 155-200 A`s to the 3' end of mRNA. Afterthat poly (A) addition is down because of low binding efficiency of poly (A) polymerase.

Polyadenylation is one of the important post transcriptional modification for exportation of mRNA from nucleus to the cytoplasm through nuclear pores. Poly (A) tail acts as a carrier signal and binds to poly (A) binding protein which facilitates in the export of mRNA from nucleus to cytosol. It also protects the mRNA from degradation. Moreover, poly (A) binding protein bind and recruits several proteins associated with translation (Siddiqui et al. 2007).

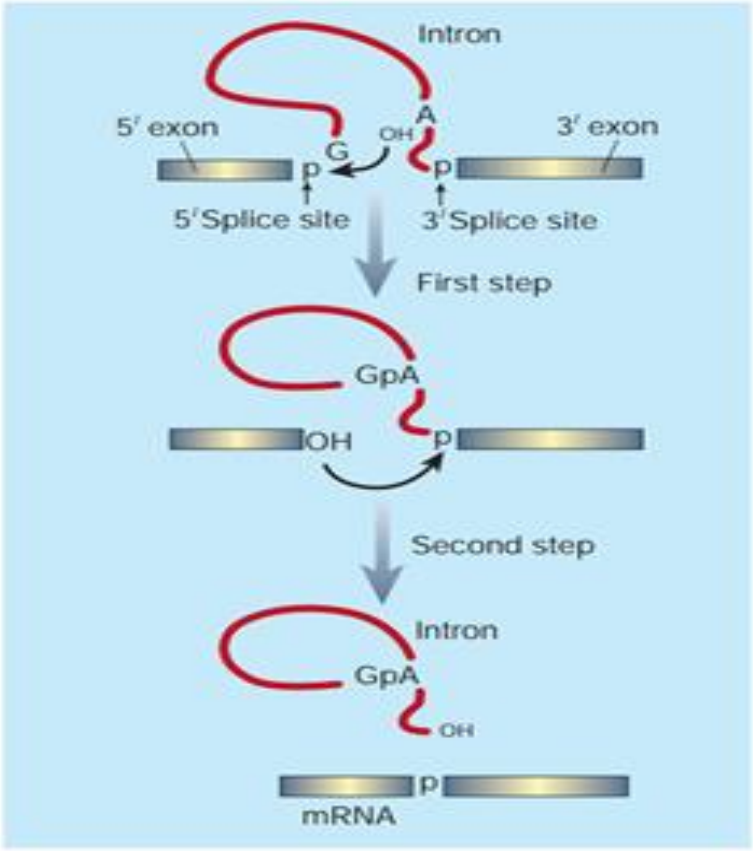


Figure 1. Mechanism of RNA splicing. In step One, the 2′ OH group on the adenosine which is the branch point of the splicing event attacks as a nucleophile on the 5' phosphodiester bond (p) which is the splice site. This results in a new bond phosphodiester formation (GpA). In step two, 3′ OH group at the end of exon attacks the free 5′ phosphate group making the removal of intronic region in the form of lariat and joins the exons together and thus mRNA is formed (Andy et al.2001).

1.1.4 Pre-mRNA splicing

Splicing of pre-mRNA is defined as the process of excision off the non-coding region called as Intron and association of the coding region called as Exon to form a continuous sequence called mature RNA. The boundaries of mRNA are decided by the 5' and 3' splice sites on the upstream and downstream end respectively (Zhang et al. 2007). Ribonucleoprotein (RNP) complex also referred as spliceosome excises the introns and joins exons to form a mature RNA (Will & Lu¨ hrmann 2011). The spliceosome is an

assembly of U1, U2, U4, U5 and U6 snRNPs and a large number of auxiliary proteins (Will et al. 2005). Splicing of RNA comprises two steps. In the first step, nucleotide A which is the branch point present at the 3' splice site chemically attacks at the 5' site to excise the intron (Will & Lu¨ hrmann 2011). In step 2, the 3'-OH of the first exon attacks the 5' site of start of the exon number two and thereby cleaves the intron and joins the exons. This reaction is actually termed as a lariat formation. After the execution of splicing, the mRNA transportation is initiated by various other enzymes to the cytosol.

1.1.5 Coupling transcription, mRNA splicing and mRNA export

As a summary, different post transcriptional modifications such as, capping, polyadenylation and splicing takes place in nucleus and then the mature mRNA transportation to the cytoplasm is done for the initiation of translation. The transcription mechanism and processing of gene expression is a coupling process which requires high proofreading. For very efficient gene expression, the coupling of RNA Polymerase II transcription and pre-mRNA splicing is essential because the nascent pre-mRNAs are protected from nuclear degradation by the splicing machinery. This is due to the association and interaction of complexes with nucleases and thereby inhibiting the nucleases. Therefore, both the processes transcription and pre-mRNA splicing plays an important role in secure and efficient transfer of the transcription complex to the splicing machinery. Also, the exportation of the mRNA is also coupled with splicing events (Reed 2003).

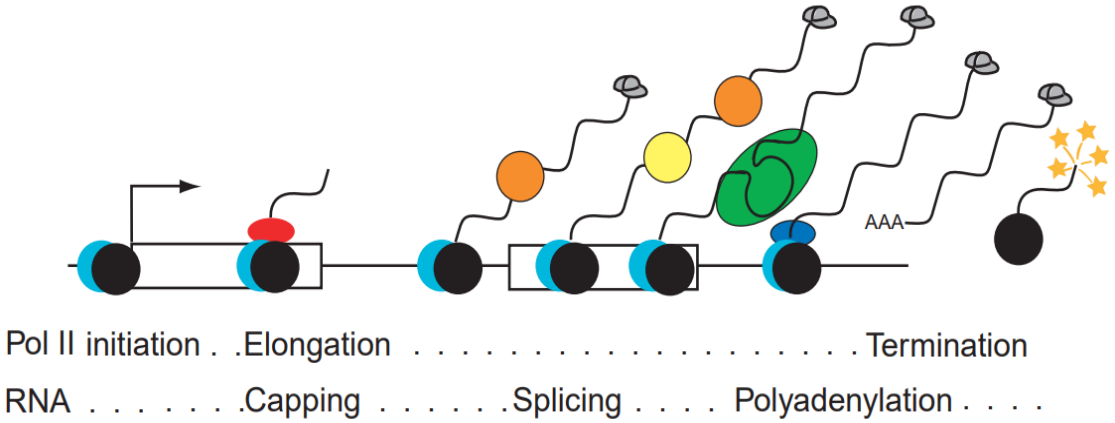


Figure 2. Transcription and pre-mRNA processing of RNA. Pictorial representation of coupled events transcription and also pre-mRNA processing together. Here polymerase II (black ball) along with the transcription assembly (blue) begins the transcription process by binding itself to the promoter sequence which is indicated by forward arrow. During the elongation phase, terminating and release of RNA from the template occurs by the polyadenylation signals. There are actually several polyadenylation factors. The polyadenylation factors like CPSF and CstF directly interact and form a complex with Pol II and as shown (dark blue ball and black one). Capping of RNA is done by the enzymes called as capping enzymes represented by red colored oval. They bind to Pol II as the transcription event begins and continues to be elongation phase and then terminates by falling off. 5′ cap of mRNA is represented by a picture of cap. The splicing event is co-transcriptional mechanism and the splicing enzymes bind to 5’ and 3’ splice sites and are hypothetically represented by Orange and yellow circles respectively. The spliceosome machinery is represented by a green oval within the transcription machinery. During termination process, Pol II actually falls off the DNA template then undergoes recycling and also the fragments that are left over from preRNA will undergo degradation. Mature RNA is then released or isolated from the transcription mechanism which then undergoes transportation to cytosol (Neugebauer 2002).

1.2 Alternative splicing

1.2.1 General characteristic of alternative splicing

In eukaryotes, during splicing the non-coding regions of RNA sequence called as introns are excised and protein coding regions also called as exons are joined together. Alternative splicing/differential splicing is a post and co-transcriptional process during which exons are exponentially spliced and joined together generating various mRNA

transcript from a single gene (Barbazuk et al. 2008). Alternative splicing allows diversity of the particular genes to express multiple mRNA that eventually translated into proteins (Figure 3). Thus, the proteome was diversified by alternative splicing (Shang et al.2017). Alternative splicing has been involved in many biological processes, which results in various diseases in mammals (Faial, 2015) and regulation of abiotic stress responses in plants (Filichkin et al. 2015). For a long period of time, it was assumed that a single gene produces single protein which means approximately 200000 proteins are produced from 200000 genes. However, this concept has been altered after the completion of full genome sequencing which exhibits that approximately 23000 genes involves in the human genome (Sachidanandam et al. 2001). These findings suggest that large number of transcripts are produced from small numbers of genes which is possible through the alternative splicing event. In plants, more than 60% of intron containing genes undergo alternative splicing (Syed et al. 2012).

Research by using techniques like expressed sequence tags and recent high-throughput transcriptome sequencing using a computational based analyses is a better way to understand the alternative splicing events in plants. By using these they have examined how frequent the alternative splicing events take place in different species [42% and 33% of intron-containing genes in Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*), respectively]. Among different types of alternative splicing events recently, it has been explored that alternative splicing can reveal differences in their protein stability, enzymatic activity, intracellular localization, post-translational modifications, and protein-protein interactions.

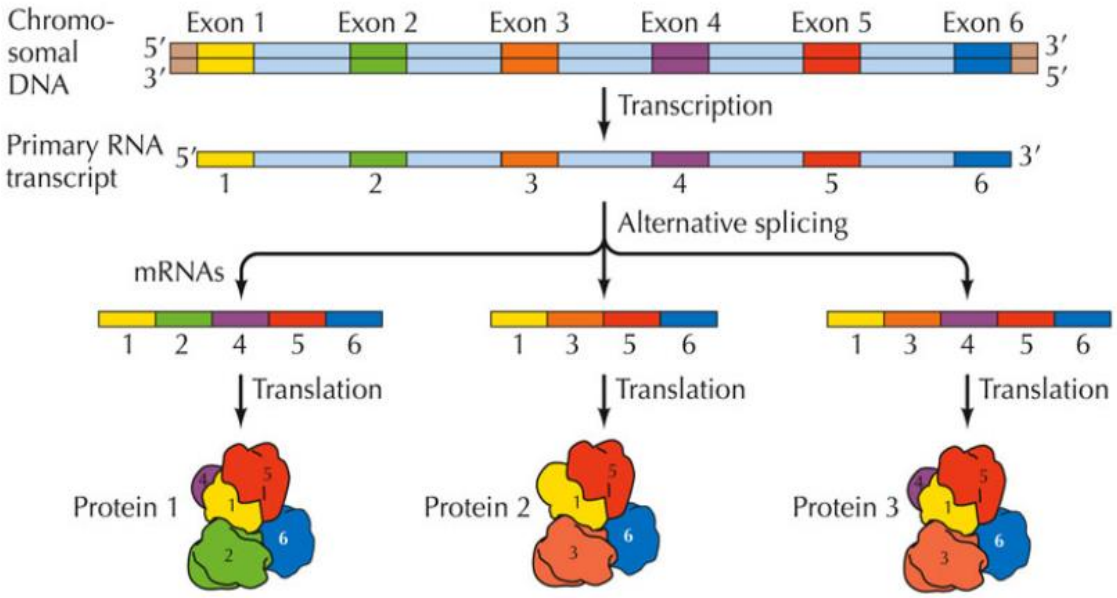


Figure 3. Schematic representation of alternative splicing. Due to the alternative splicing of nascent RNA different transcripts are generated from a single gene and different proteins are formed (Isoforms). In case of protein 1, all exons are included whereas, exon 4 and exon 2 are skipped in proteins 2 and 3 respectively (Copper et al. 2013).

1.2.2 Types of alternative splicing in plants

Alternative splicing events are divided into four subgroups in plant kingdom. Several different types of alternative splicing take place in eukaryotes resulting from the use of alternative acceptor and donor sites (Carvalho et al. 2013). In metazoan systems a phenomenon called as exon skipping is predominant (Kim et al. 2007), but not very common in Arabidopsis (<5%). Intron retention is the major type of alternative splicing in root tissues (>41%) (Li et al. 2016). Previous findings followed by alternative acceptor (<26%) and alternative donor (>12%) splicing variants (Filichkin et al. 2010). Eventually, other fewer frequent complex alternative splicing events that give rise to different alternative transcripts which include mutually exclusive exons, alternative promoters and multiple polyadenylation sites (Figure 4).

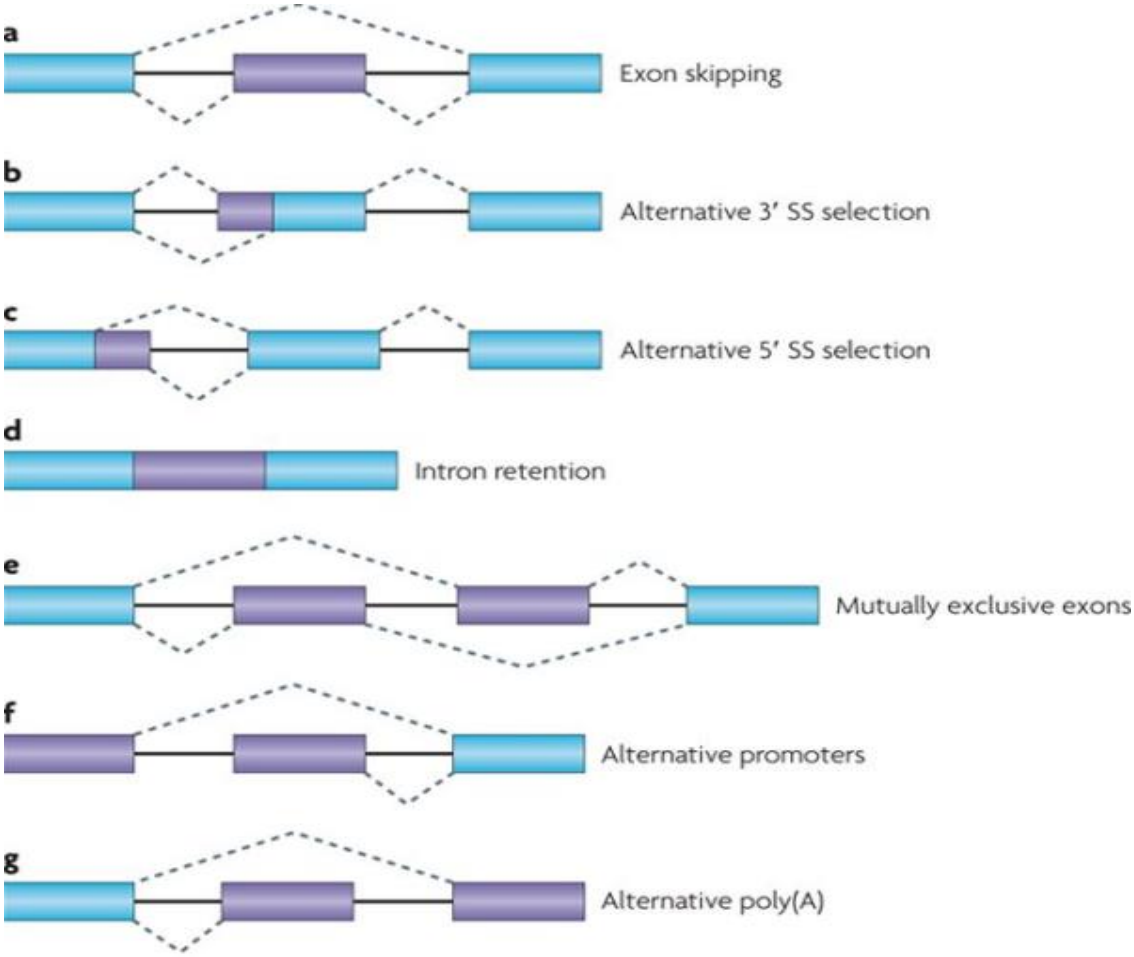


Figure 4. Patterns of alternative splicing events. In general representation of alternative splicing (**a**) Exon skipping (**b**) alternative 5' splice site selection (**c**) alternative 3' splice site selection (**d**) intron retention (**e**) mutually exclusive exons (**f**) alternative promoter and (**g**) alternative polyadenylation (Keren et al. 2010).

In *Arabidopsis thaliana*, among the alternative splicing event intron retention 40% whereas in human it is less than 5% (Figure 5). It indicates that intron retention activities much higher in plants.

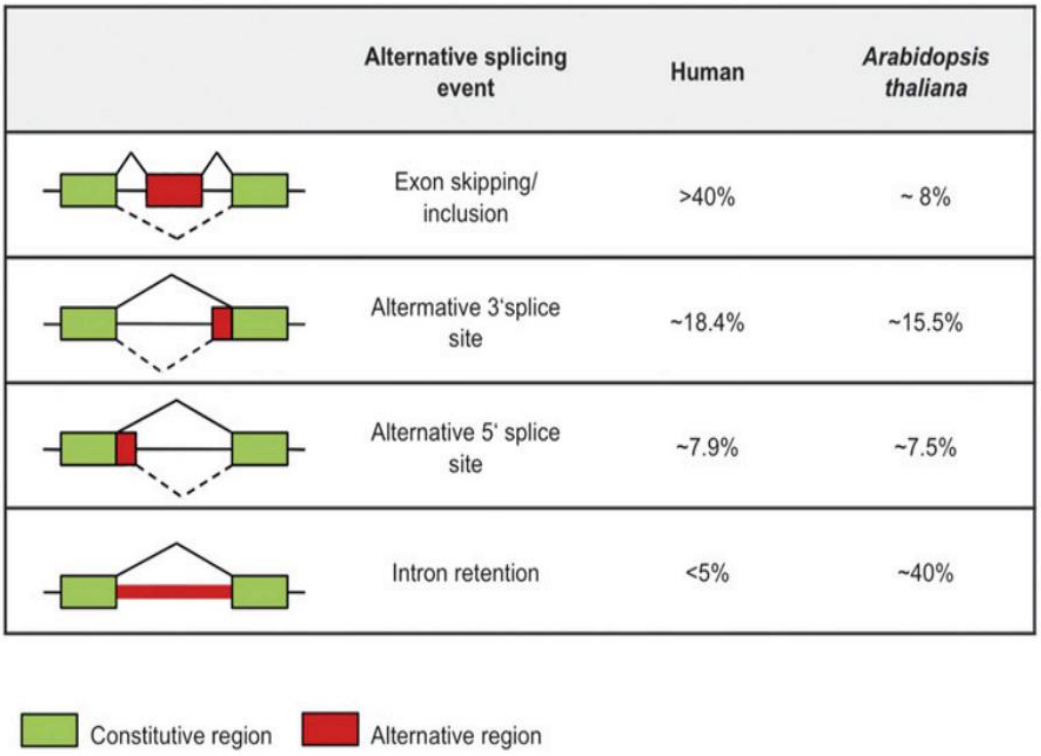


Figure 5. Comparison of alternative splicing events. A general view of alternative splicing and the common types of alternative splicing events in plants and human (Reddy et al. 2013). The Plant Cell: American Society of Plant Biologists.

1.2.3 Alternative splicing regulation

Regulation of alternative splicing events is done by both cis and trans elements. *Cis*-acting elements are those sequences which are present in the pre-mRNA and trans elements are the cellular factor. Pre-mRNA splicing can be defined as a process in which the intervening sequences which are not a part of protein coding also called as introns are removed from the RNA precursor by alternative splicing mechanism. The relative selection of splice site depends on the frequency with which an exon is selected. Splicing elements plays a significant role in determining which are the consecutive and alternate exons although these elements are vastly changeable. Based on the position and

depending on the function of the *cis*-regulatory elements, they are broadly divided into four major categories, exonic and intronic termed as exonic and intronic enhancers (ESE and ISE) and silencers (ESS and ISS) Respectively (Liu et al 2013). ESE and ISE stimulate splice site selection whereas ESS and ISS repress the splice site selection. ESE and ISE actually bound by SR (Ser-Arg) protein family and trans acting factor which results in stimulation of splicing site selection. ISS and ESS are commonly bound by heterogeneous nuclear RNPs (hnRNPs) which inhibit the splicing site selection. So this evidence suggested that both SR proteins and hnRNPs can function mutually to suppress and promote changeable exons respectively (Garneau et al. 2005). In most of the cases, these kind of splicing factors have a combined effect including positive and negative regulations which play a role in control of alternative splicing (Hertel KJ 2008; Barash et al. 2010).

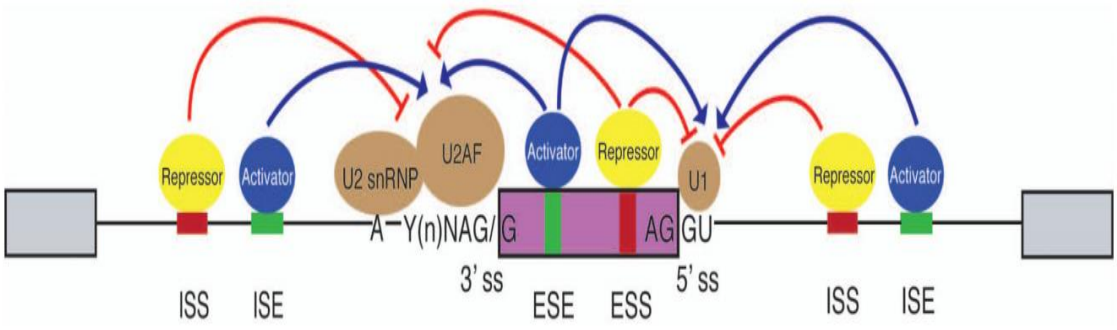


Figure 6. Schematic representation about alternative splicing regulation. There are three different center joining sequences which are perceived by the parts for spliceosome: U1 which binds itself to a 5' splice site (5'ss) What's more U2AF which ties itself to a 3' splice site (3'ss) that holds a GU an AG dinucleotide separately. U2 snRNP ties of the branch point, it is the place adenosine is shown. ESE and ESS represent exonic splicing enhancer and silencer, separately. ISE and ISS, furthermore indicate intronic splicing enhancer and silencer, individually. Splicing activators, furthermore repressors tie themselves to these *cis*-acting components for regulation (Liu et al 2013).

1.3 RNA editing in plants

1.3.1 Types of natural RNA editing events in plants

C-to-U editing

RNA editing can be called as a post-transcriptional modification which can be defined as any site-specific changes in RNA nucleotide sequences which may include insertion, deletion or base conversion of nucleotides and has been widely investigated in animals (Gott et al.2000; Maier et al. 1996; Shikanai et al.2006; Meng et al. 2010). In plants, the PTM/RNA editing is a molecular mechanism that specifically deaminates the C-to-U base substitution which is extensively expressed in mitochondria and chloroplast organs and many nuclear transcripts were also differentially edited (Meng et al. 2010). It was discovered in 1986 that the mechanism where uridines were actually inserted at different specific sites to restore or alter the protein coding sequence in the mitochondrial (kinetoplast) cytochrome *c oxidase II* (*coxII*) transcripts of *Trypanosoma brucei* to restore the function (Benne et al. 1986). In plant, RNA editing was first characterized as C-to-U conversion in mitochondrial transcripts in 1989 (Covello and Gray 1989; Gualberto et al. 1989; Hiesel et al. 1989). Most of the RNA editing sites C-to-U conversion are located mostly in the region for protein-coding which have an editing efficiently of about 90%–100% in green leaves. The greater part of the C-to-U substitution is chiefly in the protein-coding regions which safeguard the evolutionary codons. But the C-to-U substitution is the most usually observed type (Takenaka et al. 2013) and has been identified in the mitochondria organelles of diverse plants such as *Arabidopsis thaliana* (Giege and Brennicke 1999), *Oryza sativa* (Notsu et al. 2002), *Brassica napus* (Handa 2003), *Beta vulgaris* (Mower and Palmer 2006), *Vitis vinifera* (Picardi et al. 2010), *Nicotiana tabacum* (Grimes et al. 2014).

U-to-C editing

In case of some plant families, other than C-to-U editing the reverse U-to-C has also been reported in plants organelles (plastid and mitochondria) (Knie et al. 2016).

Consequently, RNA editing should fill in as a corrective mechanism at the post-transcriptional level for C-to-T (or T-to-C) changes, possibly acting as support to less favoured mutations (Maier et al.1996; Steinhauser et al. 1999). However, in *Arabidopsis thaliana* computational evidence of A-to-I RNA editing events in nucleus transcriptome suggests that they have been identified as homolog with ADAR (Du et al. 2009). Subsequently, this study would give a motivation for additionally research on mechanisms of RNA editing patterns of different nuclear transcripts across plant kingdom (Meng et al. 2010). In plants RNA editing is carried out usually in mitochondria and chloroplast as well as nucleus. In recent study they reported that some peculiar editing events were also found in plants. The 12 types of RNA editing events in Arabidopsis and Salvia*;* these events all are intra base substitutions (Meng et al. 2010; Wu et al.2015).

1.3.2 Site-specific RNA editing in plants

A set of different RNAs were subjected to site-specific RNA editing. For *AT1G29930.1* and *AT1G52400.1,* C-to-U and the U-to-C conversion both are highly enriched neighbouring the boundaries between the protein-coding region and the 3' untranslated region, which are also known as translation borders (Meng et al. 2010). Furthermore, the C-to-U and the reverse U-to-C editing sites come together, indicating that an amino-group, dissociated from C which further converts to U, could be merged with the surrounding U that afterwards converts to C. For *AT2G21660.1* and *AT2G21660.2,* A-to-I(G) RNA editing sites are also highly embellished neighbouring the

translation boundaries (Figure 7). Further investigation is needed for the biological approach of these site-specific editing events.

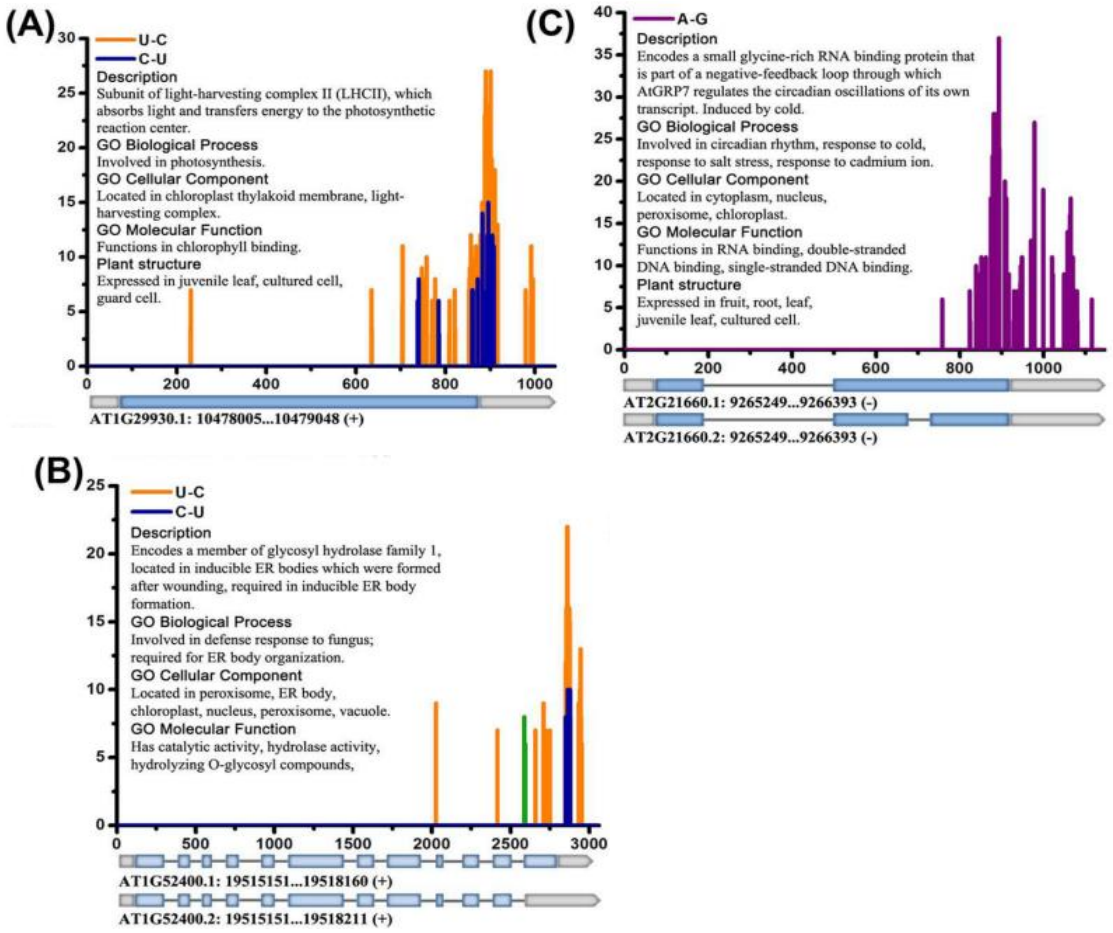


Figure 7. Site-specific RNA editing Mechanism of transcripts inside nucleus. **A-B.** Light blue colored boxes indicates exons, the gray colored boxes indicate the UTR regions and the lines indicate the introns. Orange colored bars indicate editing of RNA U-to-C nucleotides and blue colored bars indicate from C-to-U nucleotides of AT1G29930 and AT1G52400 respectively. **B**. Green colored bars indicate mRNA editing of C-to-U that is present in mRNA of AT1G52400.1. **C.** Purple colored bars indicate mRNA editing sites from A-to-I(G) nucleotides in the mRNAs of AT2G21660. The gene annotations and model ID's of **A**, **B**, and **C** are shown in figure. Here, X axis represents the transcript length; Y axis represents the number of individual and unique short-read sequences which represent a specific editing site (Meng et al. 2010).

1.4 RNA editing related family genes in plants

1.4.1 PPR proteins for RNA editing

Recently, it has been revealed that large plant-specific family of pentatricopeptide repeat (PPR) proteins may have been involved in RNA editing (Andrés-Colás et al. 2017). PPR proteins are defined by the presence of signature tandem repeats of a 35-amino acid PPR motif (Lurin et al. 2004). PPR family proteins contains RNA recognition code, these recognition codes are used for prediction of different and distinct RNA editing site. In Arabidopsis, large family members of PPR has approximately 450 members of which approximately 193 members has an E domain included while 87 others have both E and DYW domains in them. Whereas, remaining members belongs to P-subfamily (Lurin et al. 2004). In case of PPR proteins there is an additional C-terminal DYW domain significantly involved in RNA editing. DYW domains have Zinc binding motif (H*X*E and C*XX*C) which is essential for RNA editing in Arabidopsis (Hayes et al. 2013; Salone et al. 2007). In addition, DYW domain of PPR family proteins has an active site whose sequence is similar to cytidine deaminases which is an editing enzyme. It has a function of deamination of cytidine to uridine (Wagoner et al. 2015). Also, it has been shown that the E domain has an important role in RNA editing, especially the PG region (Hayes et al. 2013; Okuda et al. 2007; Okuda et al. 2009). In recent studies, it has been revealed that not only the PLS-subfamily but also the P-subfamily of PPR protein is involved in RNA editing in plants (Doniwa et al. 2010; Leu et al. 2016).

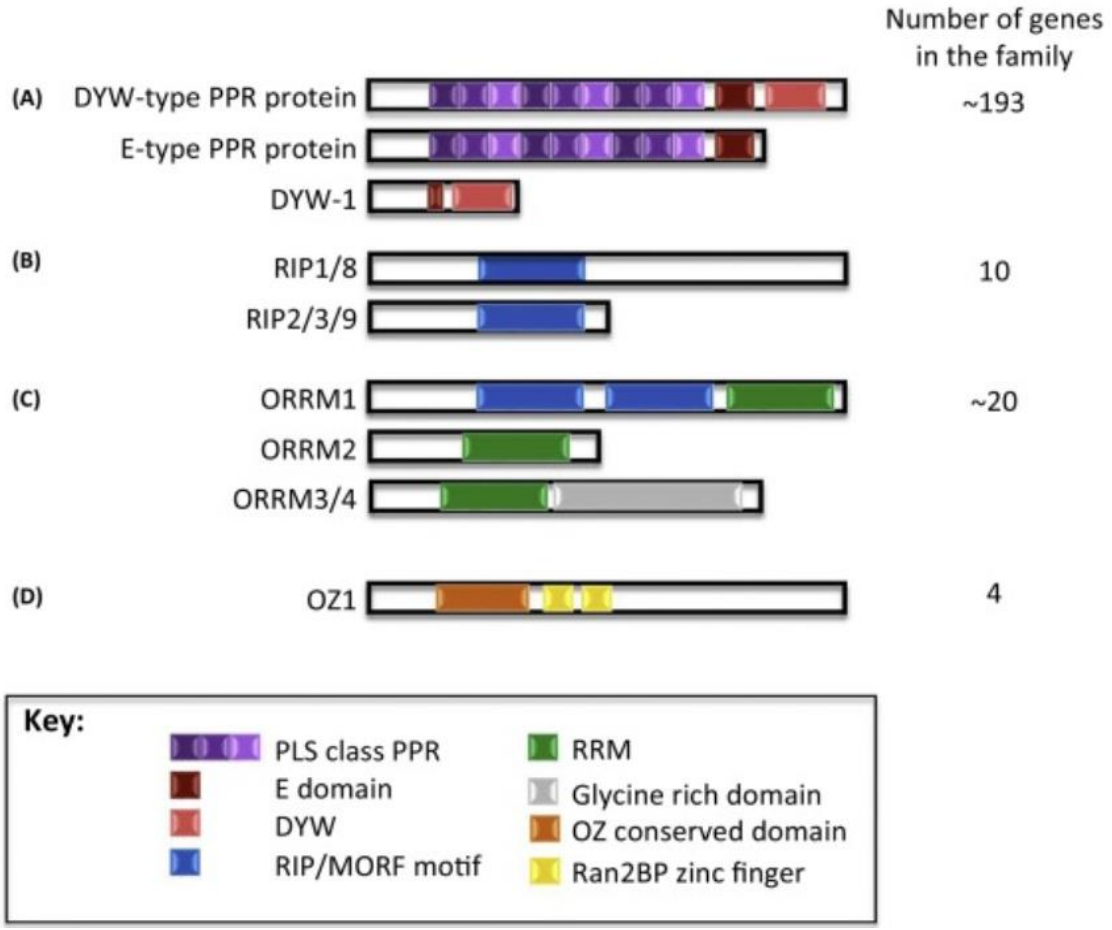


Figure 8. Pictorial Representation of the Essential RNA Editing Factors. **A.** There are around 193 different RNA editing factor proteins (pentatricopeptide repeat) in *Arabidopsis thaliana*. Pentatricopeptide proteins has a signature tandem PPR motifs and also including different specific domains such as C terminal additional E domain and a DYW domain. There are different PPR proteins E type and the DYW type. The E type proteins only differ in the DYW domain whereas the DYW-1 has a truncated E domain and complete removal of PLS class PPR domain. **B**. RIPs which are also called as Five RNA editing factor interacting proteins play a critical role in RNA editing. The RIPs have a N terminus conserved motif also called as RIP motif. RIP1 and RIP8 actually have a long extended C terminus whose function is unknown till date. **C**. Until today there are four organelle RNA recognition motif (ORRM) proteins which are known to have RNA editing mechanism. In figure, the ORRM1 consists of 2 RIP motifs indicated by blue and 1 RRM motif whereas ORRM2 has only RRM motif without the RIP motif. The ORRM3 and ORRM4 proteins have an RRM motif and are absent in RIP motif but have a glycine-rich region towards the N terminus. **D.** In total four different organelle zinc-finger (OZ) protein family members are known. But, only the OZ1 is the protein which is known to have function in RNA editing mechanism. The OZ1 protein is known to have an N terminus domain also called as OZ conserved domain and 2 zinc fingers called as Ran2BP zinc fingers (Sun et al. 2016).

1.4.2 Non-PPR proteins for RNA editing

Recent research, has identified that, non-PPR family proteins are also involved in RNA editing like RNA-Editing Factor Interacting Protein (RIP) family or Multiple Organelle RNA Editing Factor (MORF) family proteins (Takenaka et al. 2012; Bentolila et al. 2012), RNA-recognition motif (RRM)-containing proteins (Shi et al. 2015; Shi et al. 2016e; Shi et al. 2016), Protoporphyrinogen IX oxidase 1 (PPO1) (Zhang et al. 2014) and Organelle zinc-finger 1 (OZ1) (Sun et al. 2015). In Arabidopsis, there are ten members which belongs to RIP/MORF family proteins out of which five members play a vital role as editing factors for mitochondria/plastids RNA editing (Sun et al. 2016). While the remaining members have slight effect on RNA editing. Most RIP/MORF family proteins communicate with PPR factor and may affect the deaminase activity which is related in plant editosome (Sun et al. 2016). ORRMs influence many editing events in a site-specific manner. ORRMs can associate with RIP/MORF factors. They are involved in homo- or heterodimeric interactions (Shi et al. 2015; Shi et al. 2016). Alternatively, other than binding to RNA, some RRMs bind together with the different proteins which are involved in editosome formation (Sun et al. 2016). In plant, common editing event C-to-U exhibited to be zinc ion dependent because it is necessity for Zn2+ to be in the active center of having cytidine deaminase activity (Takenaka et al. 2007; Vasudevan et al. 2013). The additional factor (PPO1) vital enzyme for tetra pyrrole metabolism played an unexpected role in plastid RNA editing (Zhang et al. 2014). PPO1 does not interact with the PPR factors directly, but binds to the plastid RIPs/MORFs (Sun et al. 2016).

Table 1. List of Identified Non-PPR Editing Factors

Protein Family	Protein	Alias	Mature Protein Size (kDa)[a]	Subcellular Localization	% Affected Sites of All Chloroplast Sites	% Affected Sites of All Mitochondrial Sites
RIP/MORF	RIP1	MORF8	37	Dual	22%	77%
	RIP2	MORF2; DAL	19	Chloroplast	100%	NA[c]
	RIP3	MORF3	23	Mitochondrion	NA	26%
	RIP8	MORF1	42	Mitochondrion	NA	19%
	RIP9	MORF9; DAG	20	Chloroplast	97%	NA
ORRM	ORRM1		36	Chloroplast	62%	NA
	ORRM2		15	Mitochondrion	NA	6%[b]
	ORRM3	GR-RBP3; At-mRBP2b	30	Mitochondrion	NA	19%
	ORRM4	GR-RBP5; At-mRBP2a	29	Mitochondrion	NA	44%
OZ	OZ1	VAR3	82	Chloroplast	81%	NA
	PPO1	PPOX1	56	Chloroplast	50%	NA
	OCP3		32	Chloroplast	12%	NA
	CP31		26	Chloroplast	38%	NA

[a]protein is termed as mature protein if the protein is without the transit peptide of mitochondria/chloroplast. Table represents the sizes of different proteins and the information is based on the protein annotation in the plant genome database (http://ppdb.tc.cornell.edu). [b]Results are obtained by performing silencing experiments and not by null mutants. [c]NA indicate not applicable (Sun et al. 2016).

1.5 Mechanism of RNA editing in plants

In plants, RNA editosome complex has most of the editing sites for RNA not only take place in PPR site-specific recognition factor but also in the non-PPR factors (Sun et al. 2016). Recently it has been identified that four family members (PPRs, MORFs/RIPs, ORRMs and OZs) directly or indirectly are involved in plant RNA editosome (Okuda et

al. 2009; Doniwa et al. 2010; Chateigner-Boutin et al. 2013; Schallenberg-Rüdinger et al. 2013; Leu et al. 2016). Plant RNA editosome is one kind of diverse phenomenon, for example, there would be amino acid impairing due to U-to-C mutation which ultimately affects the protein function. Therefore, needing a solid choice for organelles that can conquer the inadequacy through RNA editing. Presumably, an existing PPR protein may show feeble interaction with a *cis* component close to another U-to-C mutation (Sun et al. 2016). To increase the RNA editing efficiency, it may be possible that the PPR protein or a *cis*-regulatory element may be recruited (Sun et al. 2016). More specific choice of proteins could increase the strength of interactions and this could lead to protection of organelles from a detrimental mutation. Generally, either chloroplasts or mitochondria, or both would be an effective target to the PPR and non PPR editing factors (Sun et al. 2016). This analogy would give us an assumption that this two organelles are involved in RNA editing. This is mostly proved by the experimental data obtained from Arabidopsis (Sun et al. 2016).

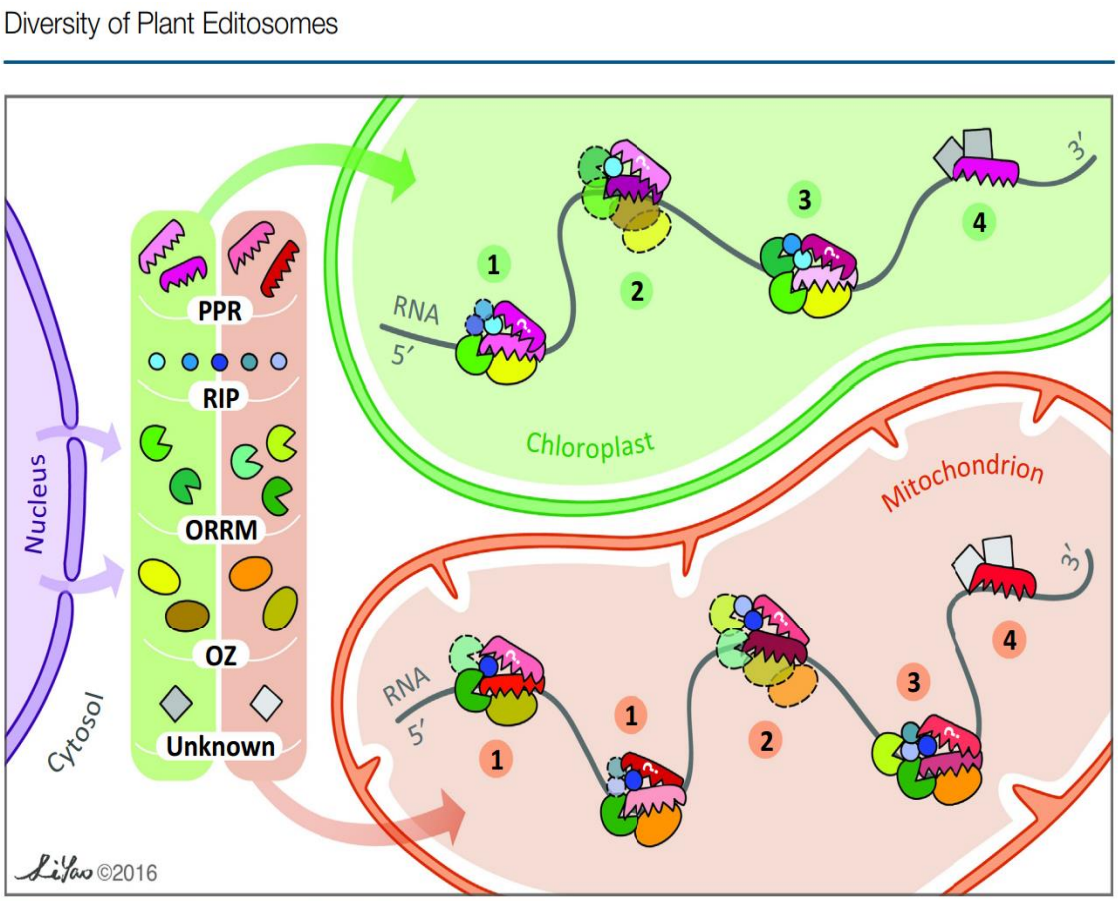


Figure 9. Diversity of plant RNA editosome. The Translation mechanism is initiated inside the nucleus i.e DNA to mRNA then, the PTM's occurs which helps in the migration of mRNA to cytosol for protein synthesis which results in translocation of protein into each organelle for function. The Figure indicates the diversity of plant editosome, in which different shapes indicate different family members whereas different color pattern indicates different members within a family of proteins. In figure there are certain unknown proteins which play a role in the RNA editing. The question marks indicate that the following PPR (Pentatricopeptide) family member **(1)** might play an important role in RNA editosome. The other non-PPR proteins like [organelle RNA recognition motif (ORRM) proteins, RNA editing factor interacting proteins (RIPs) and organelle zinc-finger (OZ) proteins] have a partial role in RNA editing mechanisms. Completely filled circles (Major Factors) play a vital role in RNA editing and are irreplaceable. Whereas, the dotted circles (dispensable factor) may replace them and have a reduced activity. Any Mutation in the major factor would completely alter the editing mechanism but, mutation in the minor factor would slightly create error in the RNA editing **(2)**. Different non-PPR family members also play an important role in RNA editing. Loss of a dispensable factor would not affect the editing much because it is replaced by another family member **(3)**. Editosome complex assembly requires multiple non-PPR copies, any disturbance in the assembly would affect the function severely **(4)**. There are several unknown proteins which may or may not play a role in editosome editing mechanism. In mitochondrial

editosomes, the presence of organelle zinc-finger (OZ) proteins is currently hypothetical (Sun et al. 2016).

1.6 Objectives of this study

Tsukahara Laboratory in JAIST mainly focusing on alternatively splicing and RNA editing. Animal RNA editing enzymes are well studied in-terms of their isoforms and activity. Recently Azad et al. has developed site-directed RNA editing with the help of MS2 system for the purpose of genetic restoration and gene therapy. In this case they engineered the deaminase domain (DD) of Adenosine Deaminase Acting in RNA (ADARs) to target specific adenosine (A) to convert it as inosine (I) that act as (G) guanosine during translation. Plant RNA editing system is more complex with different family genes and large in number. For example, U-to-C RNA editing events are frequent in plant but not found in animals. Therefore, it is obvious that variety of powerful machinery may be involved in plant RNA editing system that can be harnessed and utilized for site-directed RNA editing for genetic restoration and gene therapy.

Alternative splicing is a substantially important genetic event by which multiple transcripts are produced from a single gene. So far several family genes are identified which alternatively spliced but my research is to find out tissue-specific alternative splicing in RNA editing related family genes in Arabidopsis as it is not studied extensively. In the course of study, I also found tissue-specific RNA editing events which ultimately affect secondary structure of mRNA. Considering the above purpose the objectives of the present study are-

1. Tissue-specific alternative splicing of RNA editing related family genes (PPRs) in *Arabidopsis thaliana*
2. Identification of tissue-specific alternative splicing of zinc finger (ZnF) family genes in *Arabidopsis thaliana*
3. Analysis the tissue-specific RNA editing events in *Arabidopsis thaliana*

1.7 References

Alberts B et al (2007) Molecular Biology of the cell, 5th edition, page 336

Andrés-Colás N, Zhu Q, Takenaka M, et al (2017) Multiple PPR protein interactions are involved in the RNA editing system in *Arabidopsis* mitochondria and plastids. Proc Natl Acad Sci 114:201705815 . doi: 10.1073/pnas.1705815114

Andy N (2001): Molecular biology: RNA enzymes for RNA splicing. Nature 413, 695-696

Azad MTA, Bhakta S, Tsukahara T. (2017) Site-directed RNA editing by adenosine deaminase acting on RNA (ADAR1) for correction of the genetic code in gene therapy. Nature Publishing Group, Gene Therapy 2017 doi:10.1038/gt.2017.90

Barash Y., Calarco JA., Gao W, et al (2010) Deciphering the splicing code. Nature. 2010; 465:53–59

Bentolila S., Heller WP., Sun T., Babina AM, et al (2012) RIP1, a member of an Arabidopsis protein family, interacts with the protein RARE1 and broadly affects RNA editing. Proc Natl Acad Sci U S A. 2012 May 29; 109(22): E1453–E1461. doi:

10.1073/pnas.1121465109

Barbazuk WB, Fu Y & McGinnis KM. (2008) Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Res. 2008. 18:1381-1392. http:// www.genome.org/cgi/doi/10.1101/gr.053678.106

Carvalho RF, Feijão CV & Duque P. (2013) On the physiological significance of alternative splicing events in higher plants. Protoplasma (2013) 250:639–650 doi 10.1007/s00709-012-0448-9

Cooper GM and Hausman RE. (2013) A molecular approach, 6th edition, p186. The cell

Doniwa Y, Ueda M, Ueta M, et al (2010) The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. Gene 454:39–46 . doi: 10.1016/j.gene.2010.01.008

Du P., Chen Y & Li Y. (2009) Computational evidence of A-to-I RNA editing in nucleus transcriptome of *Arabidopsis thaliana*. Frontiers of Electrical and Electronic Engineering in China; December 2009, 4:349

Faial T. (2015) Rna splicing in common disease. *Nature Genetics* volume 47, page105 (2015). https://doi.org/10.1038/ng.3206

Filichkin SA., Cumbie JS., Dharmawadhana JP., Jaiswal P, et al (2015) Environmental Stresses Modulate Abundance and Timing of Alternatively Spliced Circadian Transcripts in Arabidopsis. Mol. Plant. 8, 207–227. https://doi.org/10.1016/j.molp. 2014.10.011

Filichkin SA., Priest HD., Givan SA., Shen R, et al (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. 2010. 20: 45-58.

doi:10.1101/gr.093302.109

Garneau D., Revil T, Fisette JF, et al (2005) Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator Bcl-x. J Biol Chem. 2005; 280:22641–22650

Hayes ML, Giang K, Berhane B & Mulligan RM. (2013) Identification of Two Pentatricopeptide Repeat Genes Required for RNA Editing and Zinc Binding by C-terminal Cytidine Deaminase-like Domains. J Biol Chem. 2013 Dec 20;288(51):36519-29. doi: 10.1074/jbc.M113.485755. Epub 2013 Nov 5

Hertel KJ. (2008) Combinatorial control of exon recognition. J Biol Chem. 2008; 283:1211– 1215

Keren H, Lev-Maor G & Ast G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* volume11, pages345–355 (2010)

Kim E., Magen A & Ast G. (2006) Different levels of alternative splicing among eukaryotes. Nucleic Acids Research, 2007, Vol. 35, No. 1 125–131 doi:10.1093/nar/gkl924

Knie N., Grewe F., Fischer S & Knoop V. (2016) Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. BMC Evolutionary Biology (2016) 16:134 DOI 10.1186/s12862-016-0707-z

Knoop V., Lurin C. (2007) A hypothesis on the identification of the editing enzyme in

plant organelles. FEBS Lett. 581, 4132–4138

Leu KC, Hsieh MH, Wang HJ, et al (2016) Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing. RNA Biol. 13:593–604

Li S, Yamada M, Han X, et al (2016) High-resolution expression map of the arabidopsis root reveals alternative splicing and lincRNA regulation. Dev Cell 39:508–522

Liu S and Cheng C (2013) Alternative RNA splicing and cancer. Wiley Interdiscip Rev RNA. 2013; 4(5): 547–566. doi:10.1002/wrna.1178

Lurin C, Andrés C, Aubourg S, et al (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089–2103

Lutz CS (2008) Alternative Polyadenylation: A Twist on mRNA 3' End Formation. ACS Chemical Biology 3, 609-17

Neugebauer KM (2002) On the importance of being co-transcriptional. Journal of cell science, 115, 3865-3871

Okuda K, Chateigner-Boutin A-L, Nakamura T, et al (2009) Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. Plant Cell 21:146–156 . doi: 10.1105/tpc.108.064667

Okuda K, Myouga F, Motohashi R, et al (2007) Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. Proc Natl Acad Sci U S A. 2007 May 8; 104(19): 8178–8183. doi: 10.1073/pnas.0700865104

Reddy, A.S.N., Marquez, Y., Kalyna, M & Barta, A (2013) Complexity of the alternative splicing landscape in plants. Plant Cell. doi/10.1105/tpc.113.117523

Reed R (2003) Coupling transcription, splicing and mRNA export. Current opinion in cell biology, 15, 326-331

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409, 928-933

Salone V., Rüdinger M., Polsakiewicz M., Hoffmann B, et al (2007) A hypothesis on the identification of the editing enzyme in plant organelles. FEBS Lett. 581, 4132–4138

Shang X., Cao Y & Ma L (2017) Alternative Splicing in Plant Genes: A Means of Regulating the Environmental Fitness of Plants. Int. J. Mol. Sci. 2017, 18, 432; doi: 10.3390/ijms18020432

Shi X., Bentolila S & Hanson MR (2016) Organelle RNA recognition motif-containing (ORRM) proteins are plastid and mitochondrial editing factors in Arabidopsis. Plant Plant Signal Behav. 2016 May 3; 11(5):e1167299. doi: 10.1080/15592324.2016.1167299

Shi X., Germain A., Hanson MR & Bentolila S (2016) RNA Recognition Motif-Containing Protein ORRM4 Broadly Affects Mitochondrial RNA Editing and Impacts Plant Development and Flowering. www.plantphysiol.org/cgi/doi/10.1104/pp.15.01280

Shi X., Hanson MR & Bentolila S (2015) Two RNA recognition motif-containing proteins are plant mitochondrial editing factors. Nucleic Acids Research, 2015, Vol.

43, No. 7; doi: 10.1093/nar/gkv245

Shinozaki A., Arahata K & Tsukahara T (1999) Changes in pre-mRNA splicing factors during neural differentiation in P19 embryonal carcinoma cells. The international journal of biochemistry and cell biology, 31, 1279-1287

Siddiqui N., Mangus DA., Chang TC, Palermino JM, et al (2007) Poly(A) Nuclease Interacts with the C-terminal Domain of Polyadenylate-binding protein Domain from Poly(A)-binding Protein. Journal of Biological Chemistry 282: 25067-75

Sun, T., Bentolila, S., & Hanson, M. R (2016) The Unexpected Diversity of Plant Organelle RNA Editosomes. Trends in Plant Science, *21*(11), 962–973. https://doi.org/10.1016/j.tplants.2016.07.005

Sun, T., Shi, X., Friso, G., Van Wijk, K., Bentolila, S., & Hanson, M. R (2015) A Zinc Finger Motif-Containing Protein Is Essential for Chloroplast RNA Editing. *PLoS Genetics*, *11*(3), 1–23. https://doi.org/10.1371/journal.pgen.1005028

Syed NH, Kalyna M, Marquez Y, et al (2012) Alternative splicing in plants - coming of age. Trends Plant Sci 17:616–623 . doi: 10.1016/j.tplants.2012.06.001

Takenaka M., Verbitskiy D., van der Merwe JA., Zehrmann A, et al (2007) In vitro RNA editing in plant mitochondria does not require added energy. FEBS Letters 581 (2007) 2743–2747; doi: 10.1016/j.febslet.2007.05.025

Takenaka M., Zehrmann A., Verbitskiy D., Kugelmann M., Härtel B & Brennicke A (2012) Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. PNAS March 27, 2012. 109 (13) 5104-5109; https://doi.org/10.1073/pnas.1202452109

Vasudevan AA., Smits SH., Höppner A., Häussinger D, et al (2016) Structural features of antiviral DNA cytidine deaminases. Biol Chem. 2013 Nov; 394(11):1357-70. doi: 10.1515/hsz-2013-0165

Will C.L. and Luhrmann R (2005) In The RNA World. (eds Gesteland, RF., Cech, TR. And Atkins, JF.) (Cold Spring Harbor Press, USA)

Will C.L. and Luhrmann R (2011) Spliceosome Structure and Function. Cold Spring Harb Perspect Biol; doi: 10.1101/cshperspect. a003707

Wagoner JA, Sun T, Lin L, Hanson MR (2015) Cytidine Deaminase Motifs within the DYW Domain of Two Pentatricopeptide Repeat-containing Proteins Are Required for Site-specific Chloroplast RNA Editing. THE JOURNAL OF BIOLOGICAL CHEMISTRY. 2015. doi: 10.1074/jbc.M114.622084

Zhang C., Hastings M.L., Krainer A.R & Zhang M.Q (2007) Dual-specificity splice sites function alternatively as 5' and 3' splice sites. www.pnas.orgcgidoi10.1073pnas.0703773104

Zhang F., Tang W., Hedtke B., Zhong L, et al (2014) Tetrapyrrole biosynthetic enzyme protoporphyrinogen IX oxidase 1 is required for plastid RNA editing. PNAS February 4, 2014. 111 (5) 2023-2028; https://doi.org/10.1073/pnas.1316183111

Chapter II

Tissue-Specific Alternative Splicing of Pentatricopeptide Repeat (PPR) Family Genes in *Arabidopsis thaliana*

2.1 Introduction

Alternative splicing is a bimolecular process in which multiple mRNAs are generated from the same gene through the selection of different splicing sites. This key process occurs in living organisms during development and is regulated in response to environmental factors (Graveley 2001; Lareau et al. 2004; Kelemen et al. 2013). Alternative splicing is a crucial phenomenon in plants. Approximately 15 years ago, it was thought that 1.2% of genes undergo alternative splicing in *Arabidopsis thaliana*, but due to improvements in next generation sequencing technology, we now know that, surprisingly, up to 60% of multiexon-containing genes undergo alternative splicing (Syed et al. 2012). Core circadian clock genes are also controlled by alternative splicing in Arabidopsis under different environmental conditions (James et al. 2012; Syed et al. 2012). The functional consequences and effects of alternative splicing on plant phenotypes are an important focus of study (Syed et al. 2012). Indeed, transcriptome analysis has shown that >95% of human genes undergo alternative splicing, greatly contributing to the huge diversity of proteins (Pan et al. 2008).

Arabidopsis contains approximately 450 pentatricopeptide repeat (PPR) proteins (Lurin et al. 2004). PPR proteins were first identified almost two decades ago (Small and Peeters 2000). This important protein family is involved in a wide variety of cellular processes in plants, such as RNA editing (Okuda et al. 2006; Andrés-Colás et al. 2017; Bayer-Császár et al. 2017), RNA stabilization (Choquet 2009; Prikryl et al. 2011), post-transcriptional RNA maturation (Delannoy et al. 2007; Williams-Carrier et al. 2008), seed development (Gutierrez-Marcos et al. 2007), endonuclease activity (Zhou et al. 2017), and various phenotypic effects (Cushing et al. 2005). The PPR protein family comprises the P and PLS subfamilies. P proteins contain a 35 amino-acid classical

tandem repeat known as the P motif. PLS subfamily proteins consist of a P motif, an S (short) 31 amino-acid motif, and an L (long) 35–36 amino-acid motif. The PLS subfamily can be divided into the PLS, DYW, and E subclasses. DYW and E proteins contain an additional, conserved C-terminal motif with deaminase properties that plays a distinct role in RNA editing in plants (Okuda et al. 2009; Chateigner-Boutin et al. 2013; Schallenberg-Rüdinger et al. 2013; Leu et al. 2016). The DYW domain has C-to-U converting activity and shares similarity with the domain found in cytidine deaminase (Shikanai 2015). The P subfamily members are mainly involved in RNA splicing, stabilization, cleavage, and the activation or repression of translation (Barkan and Small 2014). PPR proteins primarily form complexes with other family members or sometimes with associated group members to form editosome complexes (Andrés-Colás et al. 2017). In addition to the PLS subfamily, the P subfamily of PPR proteins is involved in RNA editing in plants (Doniwa et al. 2010; Leu et al. 2016).

RNA editing efficiency varies in different tissues and developmental stages, such as non-green tissues in seedlings and mature plants, although most RNA editing sites are found in green tissues (Tseng et al. 2013). The PPR motif can be reprogrammed to target any RNA molecule (Coquille et al. 2014). Recent advancements in the use of synthetic PPR molecules have created interest in programmable targeting to RNA substrates. Most PPR proteins localize to the mitochondria, chloroplasts, or both organelles, whereas very few localize to the cytoplasm and nucleus (Colcombet et al. 2013). PPR proteins play an important role in regulating organellar gene expression. Tissue-specific alternative splicing affects protein phosphorylation, which ultimately alters protein stability, enzymatic activity, subcellular localization in mammalian tissues (Merkin et al. 2012). However, the role of tissue-specific alternative splicing in plants

is less well understood.

To date, most studies investigating alternative splicing have been performed using whole plants. However, I was interested in investigating this process in specific tissues and at particular developmental stages. The aim of the current study was to investigate tissue-specific alternative splicing of selected *PPR* genes in Arabidopsis seedling, leaf, stem, stipe, and root tissue. I also investigated the expression patterns of alternatively spliced transcripts in various Arabidopsis tissues on day 4, 8, 12, 16, 21, 27, and 32 of plant development. Finally, I investigate how alternative splicing affects protein diversity in this plant, shedding light on this important process during plant development.

2.2 Materials and Methods

2.2.1 Plant growth conditions and sample collection

Arabidopsis thaliana ecotype Colombia (Col-0) seeds were sown in paper pots containing a 1:2:1 mixture of horticultural perlite, peat moss, and vermiculite, covered with plastic wrap, and incubated for 3 to 4 days in the dark. The pots were transferred to a U-ING Green Farm hydroponic grow box (Japan) in a growth room at 22°C, relative humidity 45%, and a 16 h light/8 h dark cycle. After germination, the plants were watered every morning and evening and fertilized twice weekly. Seedlings (whole plants on days 4, 8, and 12) and 16-, 21-, 27-, and 32-day-old leaf, stipe, stem, and root tissue were collected for analysis.

2.2.2 RNA extraction and cDNA synthesis

RNA was extracted from the samples using a Qiagen Plant Mini kit (Germany) according to the manufacturer's instructions. The RNA was treated with DNase (RQ1

RNase free DNase; Promega, Madison, WI, USA) to digest contaminating genomic DNA. After DNase treatment, the samples were purified by phenol-chloroform and ethanol precipitation. The final purified RNA was quantified using a NanoDrop (Thermo Scientific). The cDNA was synthesized using reverse transcriptase (Superscript III, Invitrogen) with oligo dT primers, which was confirmed using primers for the Arabidopsis housekeeping gene GAPDH: forward primer: GTTGTCATCTCTGCCCCAAG, reverse primer: TGCAACTAGCGTTGGAAACA.

2.2.3 Selection of PPR candidate genes from the Arabidopsis genome and mRNA database

Candidate of alternatively spliced genes were selected from various databases as follows: The accession numbers of 425 *PPR* genes were obtained from https://www.ncbi.nlm.nih.gov/. The accession numbers were used as queries against the www.plantgdb.org/AtGDB/ database to obtain a genomic map of each gene. Based on these genomic maps, likely alternatively spliced genes (whole genome model) were selected using At-TAIR10. After identifying 27 genes that are alternatively spliced and expressed in different tissues from the Arabidopsis information resource (TAIR; https://www.arabidopsis.org/), the full-length genomic DNA, mRNA, and cDNA of each gene was identified as well. Finally, for each gene, the whole genome sequence, intron-exon sequences, CDS, transcript sequence, and deduced protein sequence were downloaded from http://atgenie.org and crosschecked.

2.2.4 Primer design

Primers were designed using Primer3 primer design software (bioinfo.ut.ee/primer3-0.4.0/primer3/), and BLAST analysis was performed using NCBI/Primer BLAST. When the primers failed to produce the desired product, another set of primers was designed and tested. The primers (in TE buffer at a concentration of 50 pmol/µl under salt-free conditions) were purchased from Eurofins (Japan). Each primer set was diluted to 10 pmol/µl with TE buffer (working concentration).

Table 1: Candidates studied genes and accession number of PPR family genes

Tentative gene Name	Gene ID/ Accession no.	Annotation	AT°C	Cycle	Forward	Reverse
PPR1	AT5G24060	Pentatricopeptide repeat	55	30	GAACACAGAG GACGGAGGAG	TTGACTTCCCA GTTGGCTTC
PPR2	AT5G27300	Pentatricopeptide repeat	58	30	GTTTTGATCCC ACCCATCAC	TGAACGTGTT GGGACTGAAG
PPR3	AT2G19280	Pentatricopeptide repeat	58	30	TGATCCCGAAC AGTCTCAAGT	CTCTCATCTTC CTCCGGCTA
PPR4	AT4G38150	Pentatricopeptide repeat	59	35	ATTTCCGTCTTT GGCTTGG	CAGTGCCTTTG GAGGATGAT
PPR5	AT1G30610	Pentatricopeptide repeat	58	30	GGATATGGAAA GCCGTGGTA	GCCAGTGTAA GTCACCACGA
PPR6	AT1G05670	Pentatricopeptide repeat	58	30	AAGAACTGATG CTGGTGCAA	TGGAGAGATG ATGTGGCTCTT

All genes were confirmed by PCR and sequencing. The table indicates the annealing temperature (AT°C), thermal cycling conditions (number of cycles), and primer sequences used for PCR.

2.2.5 PCR and polyacrylamide gel image analysis

PCR was performed using 30–35 cycles at a denaturation temperature of 94°C and an elongation temperature of 72°C (Table 1); the annealing temperature varied depending on the primer. The PCR products were subjected to electrophoresis on a ~6% polyacrylamide gel and stained with SYBR Green Dye (Lonza, Rockland, ME, USA). An equal amount of PCR product was subjected to polyacrylamide gel electrophoresis in a gel containing ethidium bromide. Each gel image was photographed using different

exposure times to obtain high-quality images for analysis. The gel images were analyzed using LAS3000 software (Fujifilm, Tokyo, Japan). The experiments were conducted three times.

2.2.6 Transcript sequencing

The PCR products were sequenced on an Applied Biosystems 3130xl Genetic Analyzer (Foster City, CA, USA). The desired bands from PAGE were excised, transferred to a disposable pellet pestle/tissue grinder (Kimble®, Capitol Scientific, Inc., TX, USA, catalog no. 749520-0090), and incubated in a -80°C freezer for 1 h. The frozen gel piece was ground well with a pestle. Approximately 10 µl of 0.1× TE was added to the gel powder, followed by additional grinding. After discarding the pestle, the tube was vortexed for 10 minutes and centrifuged at full speed at 4°C for 20 minutes in a tabletop centrifuge. The supernatant was transferred to another tube, and 3 µl of sample was used for sequencing with a BigDye Terminator V3.1 Sequencing Standard kit (Applied Biosystems, Austin, TX, USA). However, in some samples where the PCR product was unable to sequence, in that case the PCR product was amplified by TA cloning using the pGEMT-Easy vector system (Promega). The sequencing results using the reverse primers were reverse complemented using the online software, http://www.bioinformatics.org/sms2/reference.html (Stothard 2000). All sequencing results were aligned with the Arabidopsis genome via BLAST searches (Kent 2002). The experiments were conducted three times.

2.2.7 Analyzing the tissue-specific expression patterns of the isoforms of different *PPR* genes

The PCR products were subjected to 6% polyacrylamide gel electrophoreses in 1× TBE buffer at 200 volts for 20 minutes. The gel was stained with SYBR Green Dye (Lonza, Rockland, ME, USA) for 20 minutes in 1× TBE buffer in a constant rotating shaker. The gels were imaged using an LAS3000 Imaging System (Fujifilm, Tokyo, Japan). Image J (NIH, MD, USA) software was used for densitometry analysis of the bands for comparative expression analysis of the different alternatively spliced transcripts, which were confirmed by sequencing.

2.2.8 Prediction of regulatory sequences involved in splicing

RegRNA (Regulatory RNA Motifs and Elements Finder) software (http://regrna.mbc.nctu.edu.tw) was used to identify common regulatory sequences that enhance or suppress alternative splicing. RegRNA is an integrated web-based server used to predict sequence homology and structural homology using an input RNA sequence (Stamm et al. 2006; Chang et al. 2013).

2.3 Results

I identified 27 alternatively spliced genes of PPR family proteins from an Arabidopsis database (TAIR; https://www.arabidopsis.org/) that were differentially expressed in different tissues. Of these genes, I have selected six genes due to their functional importance for further analysis by semi-quantitative PCR and sequencing. I have represented the alternatively spliced isoforms of a gene in alphabetic order a, b, c, d.

2.3.1 Tissue-specific expression patterns of the isoforms of different *PPR* genes

In *PPR1* harboring a 3' alternative-splicing site, a 39 nucleotide (nt) sequence was added to exon 3, creating *PPR1c* (Figure 1A and C). This 3' alternative-splicing site was identified in 12-day-old seedlings, in leaves on days 16, 27, and 32, in stipes on days 16, 21, 27, and 32, in stems on days 16, 21, 27, and 32, and in roots on days 16, 21, 27, and 32 (Figure 1A). I detected another new, unannotated splice isoform containing a 3' alternatively spliced site resulting in an additional 15 nt sequence in exon 3, creating *PPR1b*, which was expressed in leaves on day 21 (Figure 1A and C). *PPR1c* expression gradually increased in stipes but gradually decreased in stems and roots; however, this isoform was highly expressed in all of these tissues on day 27 (Figure 1B).

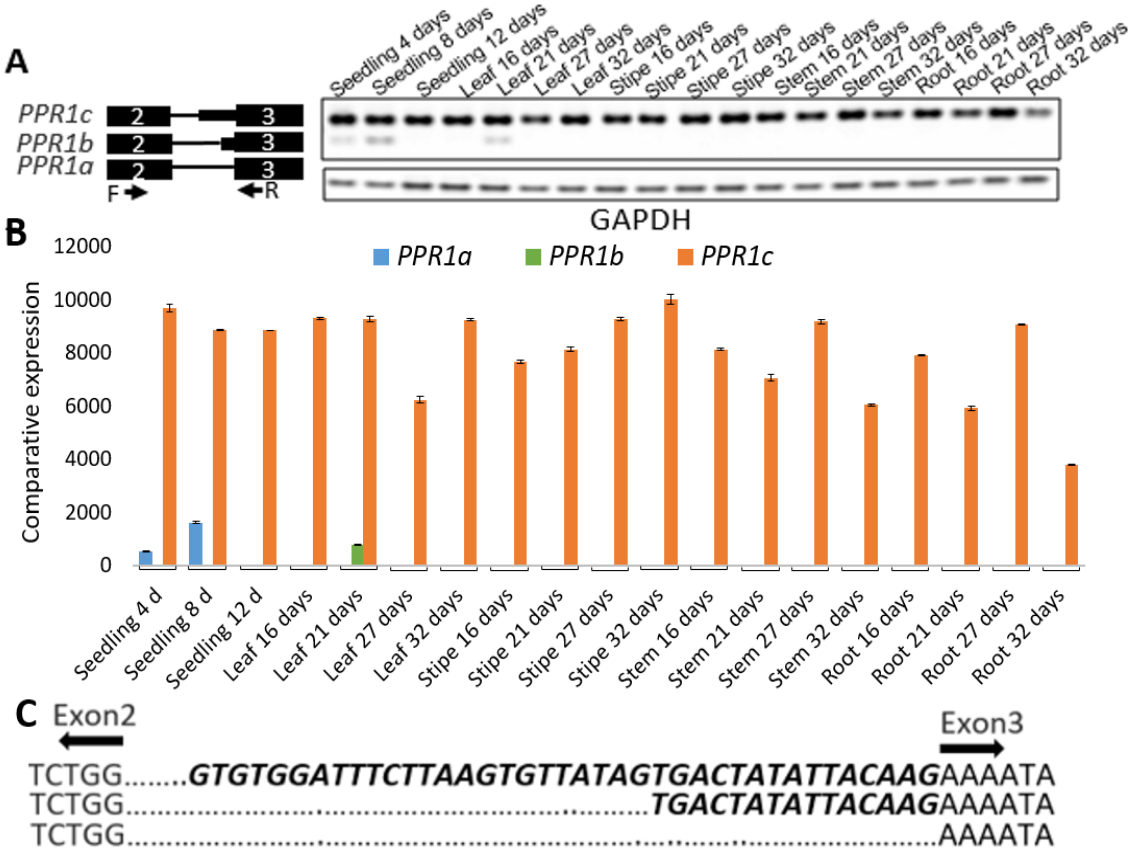


Figure 1. PCR and sequence analysis of PPR1 (accession no. AT5G24060). **A.** PCR analysis of seedling, leaf, stipe, stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of three alternatively spliced isoforms in different tissues. **C.** Genomic sequence of PPR1 from exon 2 to exon 3. Arrow indicates exon 2 and 3 boundary, and dots indicate intron sequences. Sequences in bold and italics indicate a 3′ alternative-splicing site.

In *PPR2* harboring a 3' alternatively spliced site, a 26 nt sequence was added to exon 2, creating *PPR2b*, a previously unannotated splicing product (Figure 2A and C). The 3' alternative-splicing site was activated in leaves on day 32 and in stems on days 16, 21, 27, and 32, but not in leaves on day 16 (Figure 2A). In leaves, *PPR2b* was expressed at gradually decreasing levels on days 21, 27, and 32 (Figure 2B). In stems, *PPR2a* was not expressed, but *PPR2b* was highly expressed. The expression of *PPR2b* in stems gradually increased on days 16, 21, and 27 but decreased on day 32 (Figure 2B).

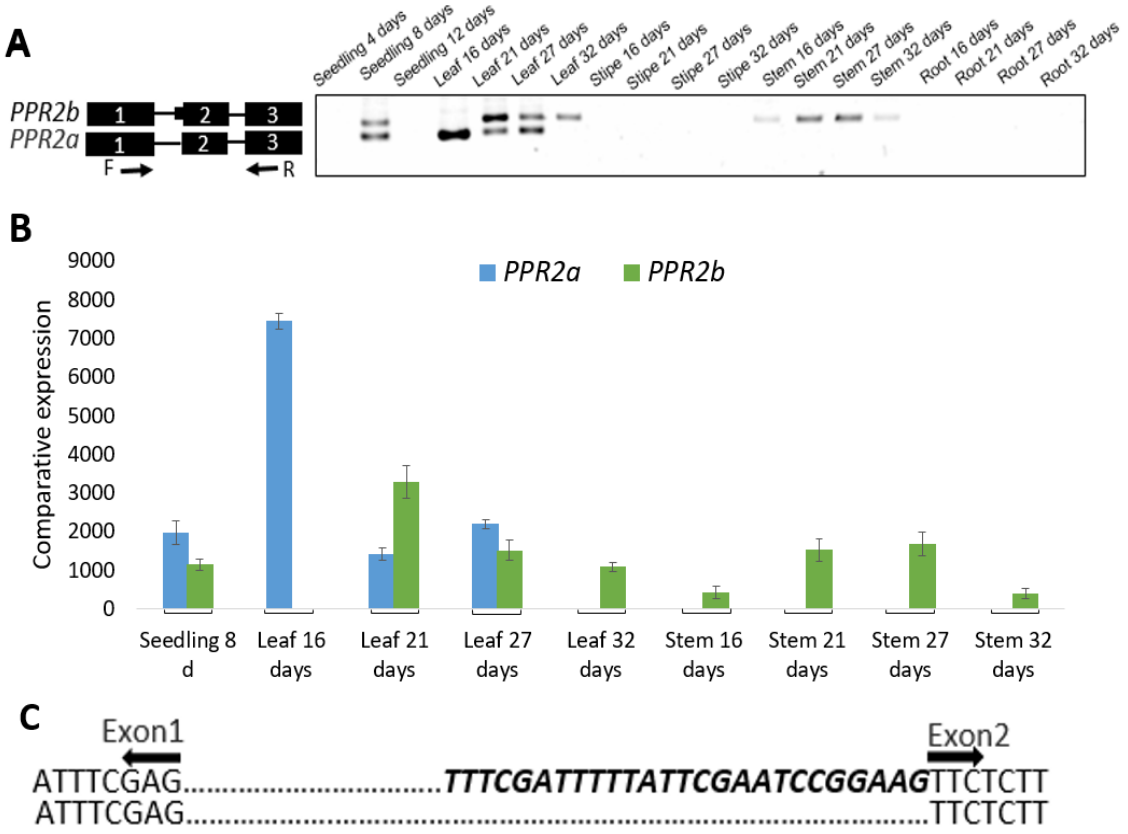


Figure 2. PCR and sequence analysis of PPR2 (accession no. AT5G27300). **A.** PCR analysis of seedling, leaf, stipe, stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of two splice isoforms in different tissues. **C.** Genomic sequence of PPR2 from exon 1 to exon 2. Arrow indicates the exon 1 and 2 boundary, and dots indicate intron sequences. Sequences in bold and italics indicate a 3′ alternative-splicing site.

In *PPR3*, a 73 bp intron sequence was retained within exons 3 and 4, creating *PPR3b* (Figure 3A and C). This isoform was not expressed in seedlings on days 4 and 12, leaves on day 27, stipes on days 16, 21, and 32, stems on days 21 and 32, or roots on days 16, 21, and 32 (Figure 3A). In leaves, *PPR3a* expression gradually increased on days 16, 21, 27, and 32, with the expression level on day 32 almost twice that on day 16. In stipes, the expression of this splice isoform gradually decreased on days 16, 21, 27, and 32 (Figure 3B); this expression pattern is opposite to that detected in leaves.

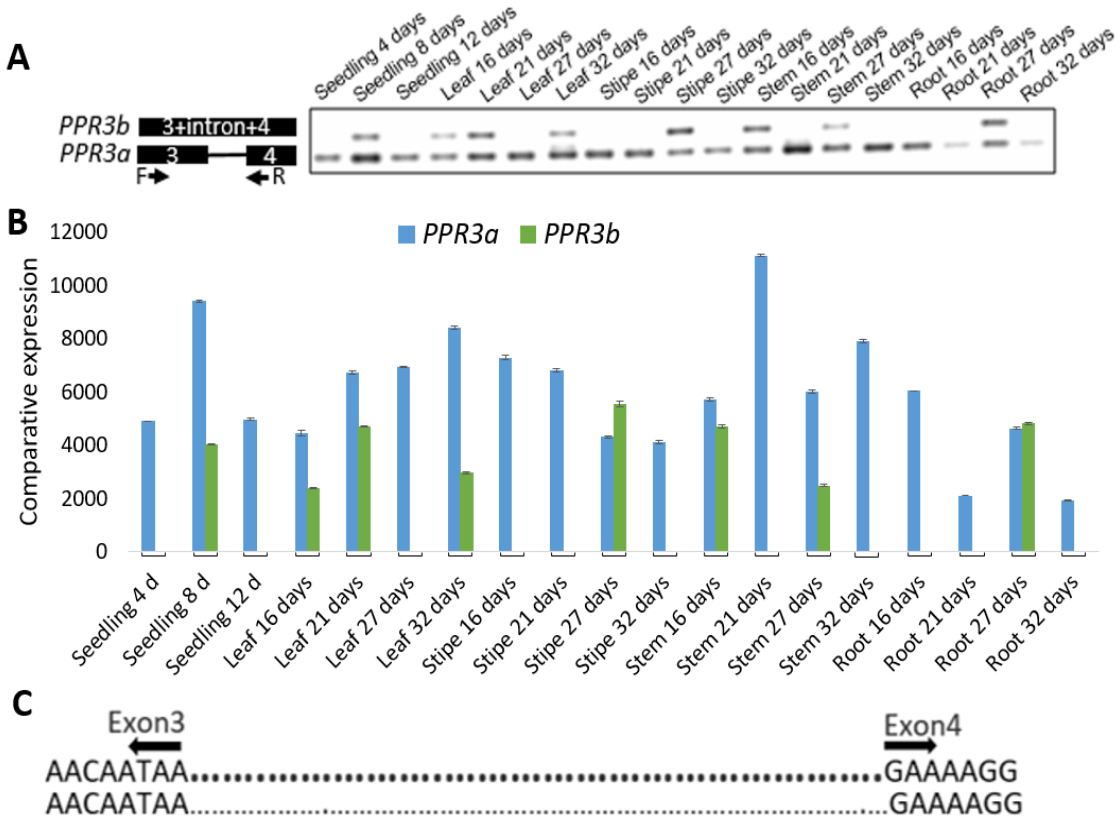


Figure 3. PCR and sequence analysis of PPR3 (accession no. AT2G19280). **A.** PCR analysis of seedling, leaf, stipe, stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of the two splice isoforms in different tissues. **C.** Genomic sequence of PPR3 from exon 3 to exon 4. Arrow indicates the exon 3 and 4 boundary, and dots indicate intron sequences. Bold dot indicate intron retained.

In *PPR4*, a 325 bp intron sequence was retained between exons 1 and 2, producing *PPR4d* (Figure 4A and C). This splice isoform was expressed in seedlings on days 4 and 12, in leaves on days 21, 27, and 32, in stipes on days 21 and 32, and in roots on days 16, 27, and 32 (Figure 4A). I detected two new, unannotated splice isoforms, including one with an additional 5' 80 nt sequence in exon 1, creating *PPR4b*, and another with a 201 bp 5'-3' sequence within exons 1 and 2, creating *PPR4c*, as confirmed by Sanger sequencing (Figure 4A and C).

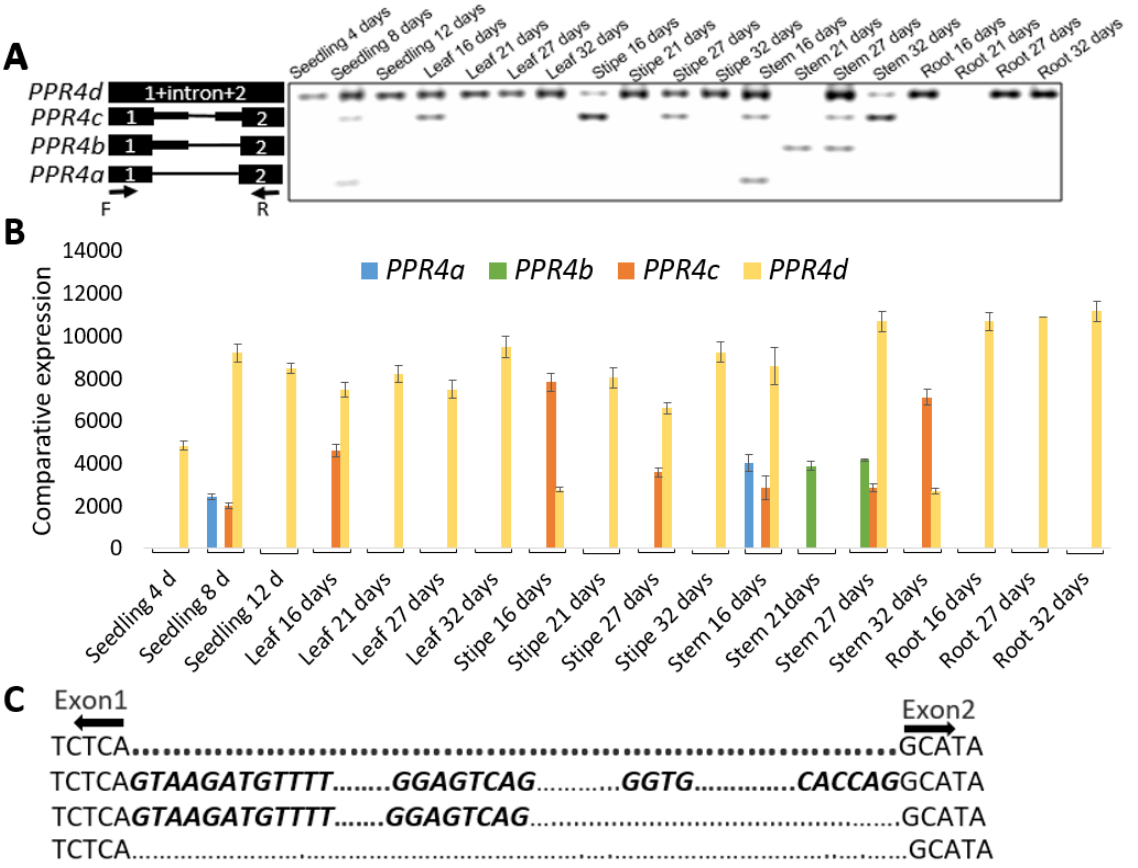


Figure 4. PCR and sequence analysis of PPR4 (accession no. AT4G38150). **A.** PCR analysis of seedling, leaf, stipe, stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of the splice isoforms in different tissues. **C.** Genomic sequence of PPR4 from exon 1 to exon 2. Arrow indicates the exon 1 and 2 boundary, and dots indicate intron sequences. Sequences in bold and italics and dot indicate a 3′-5′ alternative-splicing site. Bold dot indicate intron retained.

In *PPR5*, an 84 bp intron sequence was retained between exons 7 and 8, creating *PPR5b* (Figure 5A and C). This splice isoform was not activated in seedlings on day 4, leaves on days 16 and 27, stipes on days 21 and 32, stems on day 21, or roots on days 21, 27, and 32 (Figure 5A). In leaves, the expression of this splice isoform gradually decreased on days 16, 21, 27 but 32 days again increased (Figure 5B); this expression pattern is opposite to that detected in stipes.

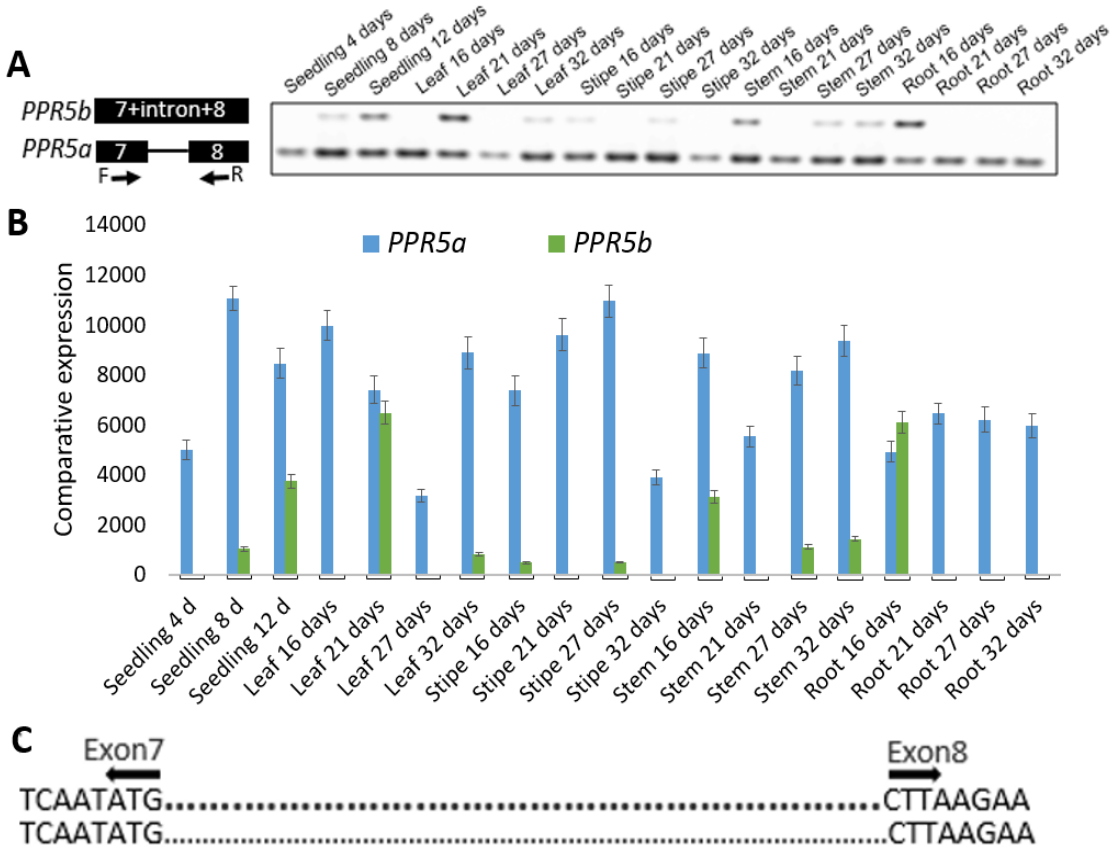
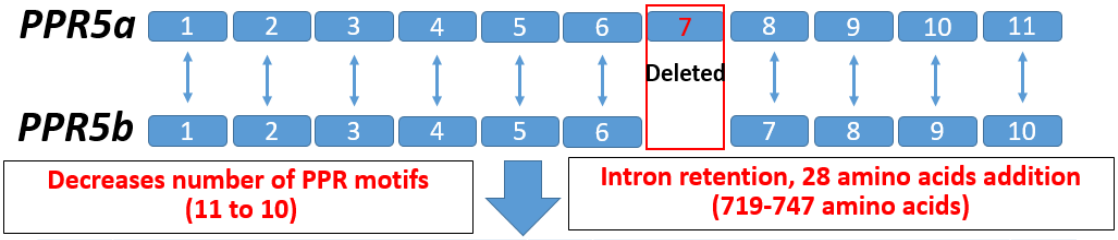


Figure 5. PCR and sequence analysis of PPR5 (accession no. AT1G30610). **A.** Polyacrylamide gel electrophoresis of PCR product amplified from seedling, leaf, stipe, stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of the two splice isoforms in different tissues. **C.** Genomic sequence of PPR5 from exon 7 to exon 8. Arrow indicates the exon 7 and 8 boundary, and dots indicate intron sequences. Bold dot indicate intron retained.

The *PPR5a* (shorter isoform) contain 11 PPR motifs, but due to alternative splicing (intron retention) the number of PPR motifs decrease from 11 to 10 (Figure 6). 7th PPR motif deleted from PPR5a and creates PPR5b longer isoform (Figure 6). 7th PPR motif contain 35 amino-acid sequences; AALYYDLARCLCSAGRCNEGLNMLKKICRVAN KPL (Figure 6).



PPR motif	Shorter isoform	A.A.	Longer isoform	A.A.
1st	DYTVMRLIHFLGKLGNWRRVLQVIEWLQRQDRYKI	35	DYTVMRLIHFLGKLGNWRRVLQVIEWLQRQDRYKI	35
2nd	IRIIYTTALNVLGKSRRPVEALNVFHAMLLQ	31	IRIIYTTALNVLGKSRRPVEALNVFHAMLLQ	31
3rd	DMVAYRSIAVTLGQAGHIKELFYVIDTMRSP	31	DMVAYRSIAVTLGQAGHIKELFYVIDTMRSP	31
4th	DVVVYNAVLNACVQRKQWEGAFWVLQQLKQRGQKP	35	DVVVYNAVLNACVQRKQWEGAFWVLQQLKQRGQKP	35
5th	SPVTYGLIMEVMLACEKYNLVHEFFRKMQKS	31	SPVTYGLIMEVMLACEKYNLVHEFFRKMQKS	31
6th	NALAYRVLVNTLWKEGKSDEAVHTVEDMESRGIVG	35	NALAYRVLVNTLWKEGKSDEAVHTVEDMESRGIVG	35
7th	AALYYDLARCLCSAGRCNEGLNMLKKICRVANKPL	35	LVVTYTGLIQACVDSGNIKNAAYIFDQMKKV	31
8th	LVVTYTGLIQACVDSGNIKNAAYIFDQMKKV	31	NLVTCNIMLKAYLQGGLFEEARELFQKMSEDGNHI	35
9th	NLVTCNIMLKAYLQGGLFEEARELFQKMSEDGNHI	35	DTYTFNTMLDTCAEQEKWDDFGYAYREMLRHGYHF	35
10th	DTYTFNTMLDTCAEQEKWDDFGYAYREMLRHGYHF	35	NAKRHLRMVLEASRAGKEEVMEATWEHMRRSNRIP	35
11th	NAKRHLRMVLEASRAGKEEVMEATWEHMRRSNRIP	35		

Figure 6. PPR5 (accession no. AT1G30610) affected PPR motifs by alternative splicing.

In *PPR6*, a 106 bp intron sequence was retained in exon 3, creating *PPR6c* (Figure 7A and C). This isoform was not activated in seedlings on days 4, 8, and 12, leaves on days 16 and 32, stipes on days 16, 21, and 27, stems on days 16, 21, 27, and 32, or roots on days 16, 21, 27, and 32 (Figure 7A).

Finally, I detected another unannotated alternative-splicing event: a 3' alternative-splicing site resulting in a 9 nt deletion in exon 3, which was detected in stems on day 16, as confirmed by Sanger sequencing (Figure 7A and C). *PPR6b*, containing a 3' alternative-splicing site, is expressed in all tissues. On the other hand, a splice variant with intron retention between exons 2 and 3 was expressed only in leaves on days 21 and 27 and in stipes on day 32 (Figure 7B).

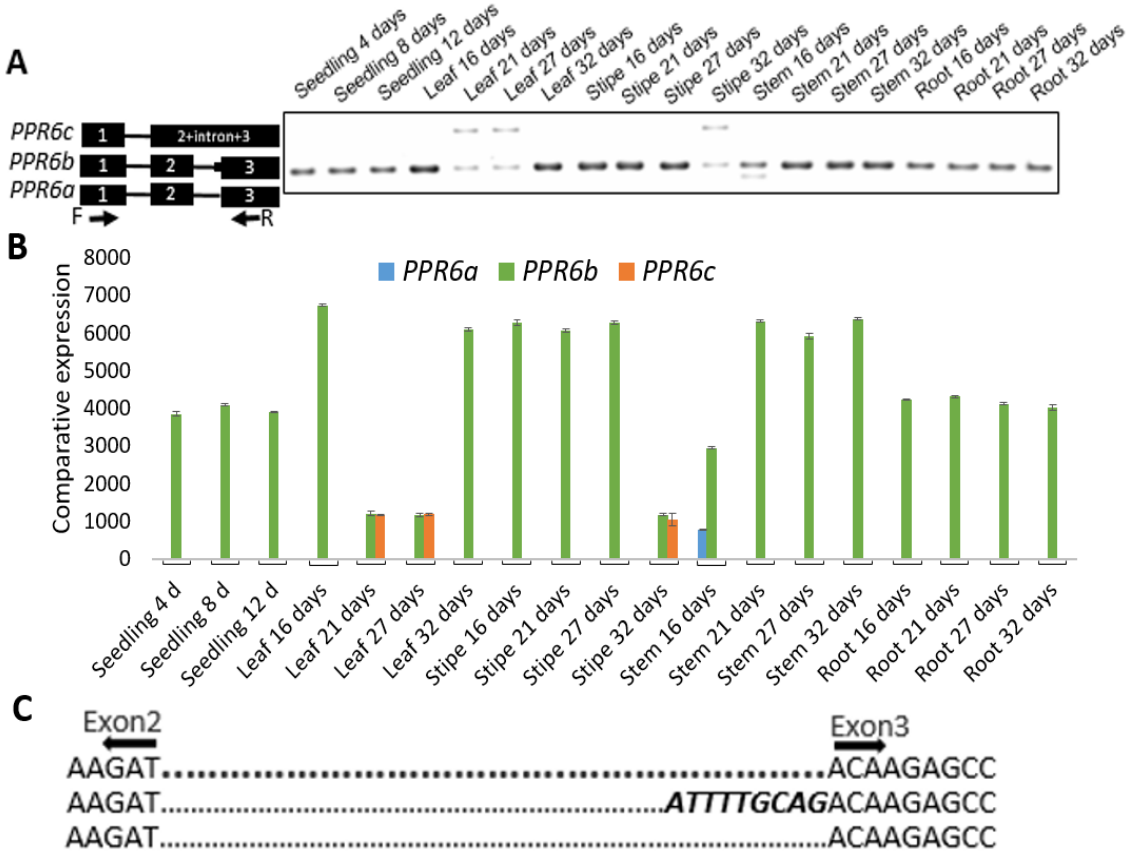


Figure 7. PCR and sequence analysis of PPR6 (accession no. AT1G05670). **A.** PCR analysis of seedling, leaf, stipe stem, and root tissue. Fragment size estimated from the gel in base pairs (bp). **B.** Comparative expression analysis of splice isoforms in different tissues. **C.** Genomic sequence of PPR6 from exon 2 to exon 3. Arrows indicate the exon 2 and 3 boundaries, and dots indicate intron sequences. Sequences in bold and italics indicate a 3′ alternative-splicing site. Bold dot indicate intron retained.

2.3.2 Determining the effect of alternative splicing on protein

I detected alternative-splicing events in the coding regions of PPR1, PPR2, and PPR5.

Table 2: Summary of alternative-splicing events affecting PPR protein

Tentative gene Name	Alternative-splicing event	Effect on	
		Nucleotides	Amino acids
PPR1a	Reference isoform shorter	-	-
PPR1b	New 3' alternative-splicing site	15 nucleotide addition	5 amino-acid addition
PPR1c	Reference isoform 3' alternative-splicing site	39 nucleotide addition	13 amino-acid addition
PPR2a	Reference isoform shorter	-	-
PPR2b	New 3' alternative-splicing site	26 nucleotide addition	Frameshift
PPR5a	Reference isoform shorter	-	-
PPR5b	Reference isoform intron retention	84 nucleotide addition	28 amino-acid addition

This table indicates that alternative splicing affects the proteins encoded by the selected genes, *PPR1*, *PPR2*, and *PPR5*.

PPR1 has two reference isoforms, PPR1a and PPR1c; however, I detected an additional, unannotated isoform, PPR1b (Table 2; Figure 1A and C), as confirmed by PCR and sequencing. PPR1b harbors a new 3' alternative-splicing site leading to the addition of a 15 nt sequence (Figure 1A and C), with an in-frame addition of five amino acids. At the same time, in the reference isoform PPR1c, a 3' alternative-splicing site leads to the in-frame addition of 39 nt encoding 13 amino acids.

PPR1a amino-acid sequences

Met A V P G F F S S **Met** A L L R H C P V S N T E D G G G S F F H V
A P R R T F R P H L L N T S S G K Y L R R N R T Q A I A E Y L G S A
S D P K K P T G K S S Y H P S E D I R A Y V P E K N P G D S R L S
P P E T A R T I I E V N K K G T L **Met** L S G L L G I G V H E N I L W P
D I P Y V T D Q H G N I Y F Q V K E N E D I **Met** Q T V V T S D N N Y
V Q V I V G F D T **Met** E **Met** I K D **Met** E L S S P S G I G F G I E E I E
D G E S E V E D E N K G D E D E G E D K D D E E W V A V L E D G
D D E D N Y V S D S D E S L G D W A N L E T **Met** R Y C H P **Met** Y F
A R R **Met** A E V A S T D P V N W **Met** D Q P S A G L A I Q G L L S P
V I V E D H S D I Q K H I S G C I S T G T D K N K E R E N S E E I F E

G I G E N E S E I L H V E N S R N A I Q Y Y K L E I I R I Q L I T A Q G
H Q T E V E V E D V R K A Q P D V I A C A S D G I L T R L E E D G D
K L T E A L R S L C W R N N G I Q A E E V K L I G I D S L G F D L R I
C S G **Met** Q I E T L R F A F S I R A T S E H N A E G Q L R E L L F A
S T P S K P Q K P K Q T N Q K E S **Stop**

PPR1a nucleotide and amino-acid sequences:

```
atggccgtccccgggttcttctcttccatggcgcttctccggcattgtcctgtatcgaac
 M  A  V  P  G  F  F  S  S  M  A  L  L  R  H  C  P  V  S  N
acagaggacggaggaggaagcttcttccatgtcgctcctcgccgaacttttcgtcctcac
 T  E  D  G  G  G  S  F  F  H  V  A  P  R  R  T  F  R  P  H
ctcctcaacactagttctggaaaatacctgcggaggaataggactcaagctatagcagag
 L  L  N  T  S  S  G  K  Y  L  R  R  N  R  T  Q  A  I  A  E
tatttgggttcagcttcagatcctaagaagccaactgggaagtcaagttatcatccttca
 Y  L  G  S  A  S  D  P  K  K  P  T  G  K  S  S  Y  H  P  S
gaagatatcagagcatatgtgccagagaagaatcctggagattctaggctttcacctcct
 E  D  I  R  A  Y  V  P  E  K  N  P  G  D  S  R  L  S  P  P
gaaactgctagaaccatcattgaggtgaacaaaaaaggaaccctgatgctctcaggttta
 E  T  A  R  T  I  I  E  V  N  K  K  G  T  L  M  L  S  G  L
cttggtattggagttcatgagaatattctctggccggatataccttatgtcacagatcag
 L  G  I  G  V  H  E  N  I  L  W  P  D  I  P  Y  V  T  D  Q
catggaaatatatatttcaagtaaaggaaaatgaggacataatgcagactgttgttaca
 H  G  N  I  Y  F  Q  V  K  E  N  E  D  I  M  Q  T  V  V  T
tctgataataattatgtgcaagtaatagtaggttttgatacgatggagatgatcaaggac
 S  D  N  N  Y  V  Q  V  I  V  G  F  D  T  M  E  M  I  K  D
atggagctgagcagtccatctggtattggttttgggattgaagaaatcgaagatggtgaa
 M  E  L  S  S  P  S  G  I  G  F  G  I  E  E  I  E  D  G  E
agtgaagttgaggatgaaaacaagggcgatgaggatgaaggagaagacaaagatgacgag
 S  E  V  E  D  E  N  K  G  D  E  D  E  G  E  D  K  D  D  E
gaatgggttgctgttctagaagatggagatgacgaggataactatgtctcagactctgat
 E  W  V  A  V  L  E  D  G  D  D  E  D  N  Y  V  S  D  S  D
gaatcacttggagactgggcaaacttggaaacgatgcgatattgccatcctatgtatttt
 E  S  L  G  D  W  A  N  L  E  T  M  R  Y  C  H  P  M  Y  F
gccaggaggatggccgaggttgcttcaactgatcctgtaaattggatggatcagccgtcc
 A  R  R  M  A  E  V  A  S  T  D  P  V  N  W  M  D  Q  P  S
gctggccttgctatccaagggctcttgagtcctgtcattgtggaagatcactcagatatc
 A  G  L  A  I  Q  G  L  L  S  P  V  I  V  E  D  H  S  D  I
caaaaacatatctctggttgtatatctacaggcactgacaaaaacaaggagagagaaaac
 Q  K  H  I  S  G  C  I  S  T  G  T  D  K  N  K  E  R  E  N
tcagaagagatatttgaaggcattggtgagaatgaatctgagatattgcatgttgaaaat
 S  E  E  I  F  E  G  I  G  E  N  E  S  E  I  L  H  V  E  N
tcgagaaatgctatacagtactacaagctggagataataagaatccagctcatcacggca
 S  R  N  A  I  Q  Y  Y  K  L  E  I  I  R  I  Q  L  I  T  A
caggggcatcagactgaagtggaagtggaagacgtcaggaaagcacaacctgatgttatt
 Q  G  H  Q  T  E  V  E  V  E  D  V  R  K  A  Q  P  D  V  I
gcttgtgcatcagatggtatactgactcgtctagaagaagatggcgataaactaactgaa
 A  C  A  S  D  G  I  L  T  R  L  E  E  D  G  D  K  L  T  E
gctctcagatctctgtgttggagaaataatgggattcaagcagaggaagtgaagctcatt
 A  L  R  S  L  C  W  R  N  N  G  I  Q  A  E  E  V  K  L  I
gggatagattcacttggtttcgaccttaggatttgctcaggaatgcaaattgagacatta
 G  I  D  S  L  G  F  D  L  R  I  C  S  G  M  Q  I  E  T  L
cggtttgcattctcgataagggcaacatcagaacacaacgctgagggacagttgagagaa
 R  F  A  F  S  I  R  A  T  S  E  H  N  A  E  G  Q  L  R  E
ttattattcgcttctacaccttccaagccgcagaaaccaaaacaaacgaatcaaaaggaa
```

```
 L  L  F  A  S  T  P  S  K  P  Q  K  P  K  Q  T  N  Q  K  E
tcttag
 S  -
```

PPR1b amino-acid sequences, yellow mark indicates spliced sequence.

Met A V P G F F S S **Met** A L L R H C P V S N T E D G G G S F F H V
A P R R T F R P H L L N T S S G D Y I T R K Y L R R N R T Q A I A E
Y L G S A S D P K K P T G K S S Y H P S E D I R A Y V P E K N P G D
S R L S P P E T A R T I I E V N K K G T L **Met** L S G L L G I G V H E N
I L W P D I P Y V T D Q H G N I Y F Q V K E N E D I **Met** Q T V V T S
D N N Y V Q V I V G F D T **Met** E **Met** I K D **Met** E L S S P S G I G F G
I E E I E D G E S E V E D E N K G D E D E G E D K D D E E W V A V
L E D G D D E D N Y V S D S D E S L G D W A N L E T **Met** R Y C H
P **Met** Y F A R R **Met** A E V A S T D P V N W **Met** D Q P S A G L A I Q
G L L S P V I V E D H S D I Q K H I S G C I S T G T D K N K E R E N
S E E I F E G I G E N E S E I L H V E N S R N A I Q Y Y K L E I I R I Q
L I T A Q G H Q T E V E V E D V R K A Q P D V I A C A S D G I L T R L
E E D G D K L T E A L R S L C W R N N G I Q A E E V K L I G I D S L
G F D L R I C S G **Met** Q I E T L R F A F S I R A T S E H N A E G Q L
R E L L F A S T P S K P Q K P K Q T N Q K E S **Stop**

PPR1b nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atggccgtccccgggttcttctcttccatggcgcttctccggcattgtcctgtatcgaac
 M  A  V  P  G  F  F  S  S  M  A  L  L  R  H  C  P  V  S  N
acagaggacggaggaggaagcttcttccatgtcgctcctcgccgaacttttcgtcctcac
 T  E  D  G  G  G  S  F  F  H  V  A  P  R  R  T  F  R  P  H
ctcctcaacactagttctggtgactatattacaagaaaatacctgcggaggaataggact
 L  L  N  T  S  S  G  D  Y  I  T  R  K  Y  L  R  R  N  R  T
caagctatagcagagtatttgggttcagcttcagatcctaagaagccaactgggaagtca
 Q  A  I  A  E  Y  L  G  S  A  S  D  P  K  K  P  T  G  K  S
agttatcatccttcagaagatatcagagcatatgtgccagagaagaatcctggagattct
 S  Y  H  P  S  E  D  I  R  A  Y  V  P  E  K  N  P  G  D  S
aggctttcacctcctgaaactgctagaaccatcattgaggtgaacaaaaaaggaaccctg
 R  L  S  P  P  E  T  A  R  T  I  I  E  V  N  K  K  G  T  L
atgctctcaggtttacttggtattggagttcatgagaatattctctggccggatatacct
 M  L  S  G  L  L  G  I  G  V  H  E  N  I  L  W  P  D  I  P
tatgtcacagatcagcatggaaatatatattttcaagtaaaggaaaatgaggacataatg
 Y  V  T  D  Q  H  G  N  I  Y  F  Q  V  K  E  N  E  D  I  M
cagactgttgttacatctgataataattatgtgcaagtaatagtaggttttgatacgatg
 Q  T  V  V  T  S  D  N  N  Y  V  Q  V  I  V  G  F  D  T  M
gagatgatcaaggacatggagctgagcagtccatctggtattggttttgggattgaagaa
 E  M  I  K  D  M  E  L  S  S  P  S  G  I  G  F  G  I  E  E
atcgaagatggtgaaagtgaagttgaggatgaaaacaagggcgatgaggatgaaggagaa
 I  E  D  G  E  S  E  V  E  D  E  N  K  G  D  E  D  E  G  E
gacaaagatgacgaggaatgggttgctgttctagaagatggagatgacgaggataactat
 D  K  D  D  E  E  W  V  A  V  L  E  D  G  D  D  E  D  N  Y
gtctcagactctgatgaatcacttggagactgggcaaacttggaaacgatgcgatattgc
```

```
 V  S  D  S  D  E  S  L  G  D  W  A  N  L  E  T  M  R  Y  C
catcctatgtattttgccaggaggatggccgaggttgcttcaactgatcctgtaaattgg
 H  P  M  Y  F  A  R  R  M  A  E  V  A  S  T  D  P  V  N  W
atggatcagccgtccgctggccttgctatccaagggctcttgagtcctgtcattgtggaa
 M  D  Q  P  S  A  G  L  A  I  Q  G  L  L  S  P  V  I  V  E
gatcactcagatatccaaaaacatatctctggttgtatatctacaggcactgacaaaaac
 D  H  S  D  I  Q  K  H  I  S  G  C  I  S  T  G  T  D  K  N
aaggagagagaaaactcagaagagatatttgaaggcattggtgagaatgaatctgagata
 K  E  R  E  N  S  E  E  I  F  E  G  I  G  E  N  E  S  E  I
ttgcatgttgaaaattcgagaaatgctatacagtactacaagctggagataataagaatc
 L  H  V  E  N  S  R  N  A  I  Q  Y  Y  K  L  E  I  I  R  I
cagctcatcacggcacaggggcatcagactgaagtggaagtggaagacgtcaggaaagca
 Q  L  I  T  A  Q  G  H  Q  T  E  V  E  V  E  D  V  R  K  A
caacctgatgttattgcttgtgcatcagatggtatactgactcgtctagaagaagatggc
 Q  P  D  V  I  A  C  A  S  D  G  I  L  T  R  L  E  E  D  G
gataaactaactgaagctctcagatctctgtgttggagaaataatgggattcaagcagag
 D  K  L  T  E  A  L  R  S  L  C  W  R  N  N  G  I  Q  A  E
gaagtgaagctcattgggatagattcacttggtttcgaccttaggatttgctcaggaatg
 E  V  K  L  I  G  I  D  S  L  G  F  D  L  R  I  C  S  G  M
caaattgagacattacggtttgcattctcgataagggcaacatcagaacacaacgctgag
 Q  I  E  T  L  R  F  A  F  S  I  R  A  T  S  E  H  N  A  E
ggacagttgagagaattattattcgcttctacaccttccaagccgcagaaaccaaaacaa
 G  Q  L  R  E  L  L  F  A  S  T  P  S  K  P  Q  K  P  K  Q
Acgaatcaaaaggaatcttag
 T  N  Q  K  E  S  -
```

PPR1c amino-acid sequences, yellow mark indicates spliced sequence.

Met A V P G F F S S **Met** A L L R H C P V S N T E D G G G S F F H V
A P R R T F R P H L L N T S S G C G F L K C Y S D Y I T R K Y L R R
N R T Q A I A E Y L G S A S D P K K P T G K S S Y H P S E D I R A Y
V P E K N P G D S R L S P P E T A R T I I E V N K K G T L **Met** L S G
L L G I G V H E N I L W P D I P Y V T D Q H G N I Y F Q V K E N E D
I **Met** Q T V V T S D N N Y V Q V I V G F D T **Met** E **Met** I K D **Met** E L
S S P S G I G F G I E E I E D G E S E V E D E N K G D E D E G E D
K D D E E W V A V L E D G D D E D N Y V S D S D E S L G D W A N
L E T **Met** R Y C H P **Met** Y F A R R **Met** A E V A S T D P V N W **Met** D
Q P S A G L A I Q G L L S P V I V E D H S D I Q K H I S G C I S T G T
D K N K E R E N S E E I F E G I G E N E S E I L H V E N S R N A I Q
Y Y K L E I I R I Q L I T A Q G H Q T E V E V E D V R K A Q P D V I A
C A S D G I L T R L E E D G D K L T E A L R S L C W R N N G I Q A E
E V K L I G I D S L G F D L R I C S G **Met** Q I E T L R F A F S I R A T
S E H N A E G Q L R E L L F A S T P S K P Q K P K Q T N Q K E
S **Stop**

PPR1c nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atggccgtccccgggttcttctcttccatggcgcttctccggcattgtcctgtatcgaac
 M  A  V  P  G  F  F  S  S  M  A  L  L  R  H  C  P  V  S  N
acagaggacggaggaggaagcttcttccatgtcgctcctcgccgaacttttcgtcctcac
 T  E  D  G  G  G  S  F  F  H  V  A  P  R  R  T  F  R  P  H
ctcctcaacactagttctgggtgtggatttcttaagtgttatagtgactatattacaaga
 L  L  N  T  S  S  G  C  G  F  L  K  C  Y  S  D  Y  I  T  R
aaatacctgcggaggaataggactcaagctatagcagagtatttgggttcagcttcagat
 K  Y  L  R  R  N  R  T  Q  A  I  A  E  Y  L  G  S  A  S  D
cctaagaagccaactgggaagtcaagttatcatccttcagaagatatcagagcatatgtg
 P  K  K  P  T  G  K  S  S  Y  H  P  S  E  D  I  R  A  Y  V
ccagagaagaatcctggagattctaggctttcacctcctgaaactgctagaaccatcatt
 P  E  K  N  P  G  D  S  R  L  S  P  P  E  T  A  R  T  I  I
gaggtgaacaaaaaaggaaccctgatgctctcaggtttacttggtattggagttcatgag
 E  V  N  K  K  G  T  L  M  L  S  G  L  L  G  I  G  V  H  E
aatattctctggccggatataccttatgtcacagatcagcatggaaatatatattttcaa
 N  I  L  W  P  D  I  P  Y  V  T  D  Q  H  G  N  I  Y  F  Q
gtaaaggaaaatgaggacataatgcagactgttgttacatctgataataattatgtgcaa
 V  K  E  N  E  D  I  M  Q  T  V  V  T  S  D  N  N  Y  V  Q
gtaatagtaggttttgatacgatggagatgatcaaggacatggagctgagcagtccatct
 V  I  V  G  F  D  T  M  E  M  I  K  D  M  E  L  S  S  P  S
ggtattggttttgggattgaagaaatcgaagatggtgaaagtgaagttgaggatgaaaac
 G  I  G  F  G  I  E  E  I  E  D  G  E  S  E  V  E  D  E  N
aagggcgatgaggatgaaggagaagacaaagatgacgaggaatgggttgctgttctagaa
 K  G  D  E  D  E  G  E  D  K  D  D  E  E  W  V  A  V  L  E
gatggagatgacgaggataactatgtctcagactctgatgaatcacttggagactgggca
 D  G  D  D  E  D  N  Y  V  S  D  S  D  E  S  L  G  D  W  A
aacttggaaacgatgcgatattgccatcctatgtattttgccaggaggatggccgaggtt
 N  L  E  T  M  R  Y  C  H  P  M  Y  F  A  R  R  M  A  E  V
gcttcaactgatcctgtaaattggatggatcagccgtccgctggccttgctatccaaggg
 A  S  T  D  P  V  N  W  M  D  Q  P  S  A  G  L  A  I  Q  G
ctcttgagtcctgtcattgtggaagatcactcagatatccaaaaacatatctctggttgt
 L  L  S  P  V  I  V  E  D  H  S  D  I  Q  K  H  I  S  G  C
atatctacaggcactgacaaaaacaaggagagagaaaactcagaagagatatttgaaggc
 I  S  T  G  T  D  K  N  K  E  R  E  N  S  E  E  I  F  E  G
attggtgagaatgaatctgagatattgcatgttgaaaattcgagaaatgctatacagtac
 I  G  E  N  E  S  E  I  L  H  V  E  N  S  R  N  A  I  Q  Y
tacaagctggagataataagaatccagctcatcacggcacaggggcatcagactgaagtg
 Y  K  L  E  I  I  R  I  Q  L  I  T  A  Q  G  H  Q  T  E  V
gaagtggaagacgtcaggaaagcacaacctgatgttattgcttgtgcatcagatggtata
 E  V  E  D  V  R  K  A  Q  P  D  V  I  A  C  A  S  D  G  I
ctgactcgtctagaagaagatggcgataaactaactgaagctctcagatctctgtgttgg
 L  T  R  L  E  E  D  G  D  K  L  T  E  A  L  R  S  L  C  W
agaaataatgggattcaagcagaggaagtgaagctcattgggatagattcacttggtttc
 R  N  N  G  I  Q  A  E  E  V  K  L  I  G  I  D  S  L  G  F
gaccttaggatttgctcaggaatgcaaattgagacattacggtttgcattctcgataagg
 D  L  R  I  C  S  G  M  Q  I  E  T  L  R  F  A  F  S  I  R
gcaacatcagaacacaacgctgagggacagttgagagaattattattcgcttctacacct
 A  T  S  E  H  N  A  E  G  Q  L  R  E  L  L  F  A  S  T  P
tccaagccgcagaaaccaaaacaaacgaatcaaaaggaatcttag
 S  K  P  Q  K  P  K  Q  T  N  Q  K  E  S  -
```

In PPR2, the reference isoform is PPR2a, and another new isoform, PPR2b (Table 2), harbors a 3' alternative-splicing site in exon 2, leading to the addition of 26 nt (Figure 2A and C). In PPR2, a 26 nt addition resulted in a frameshift.

PPR2a amino-acid sequences:

Met A A T R S T F L G S I F R T V K A R V L I P P I T R L S L P N P N
F T S S R F H T S S S L P Q S Q I E G S I S S S L L Q D L S I I H G V
I A I A S T S Q G R L A S V S**Met** C T Q Y S T S V P T R S L R R R I S
S R K K S S T K P I L N E S K F Q E T I S K L P P R F T P E E L A D A
I T L E E D P F L C F H L F N W A S Q Q P R F T H E N C S Y H I A I
R K L G A A K S G K L I R A V N I F R H **Met** V N S R N L E C R P
T **Met** R T Y H I L F K A L L G R G N N S F I N H L Y **Met** E T V R S L
F R Q **Met** V D S G I E P D V F A L N C L V K G R T I N T R E L L S
E **Met** K G K G F V P N G K S Y N S L V N A F A L S G E I D D A V K
C L W E **Met** I E N G R V V D F I S Y R T L V D E S C R K G K Y D E
A T R L L E **Met** L R E K Q L V D I D S D D K L K **Met** Y Q **Met** V I L V
L L F S S **Met** L P S V C D E S R Y **Met** I V R N V P A L G C G D D
L **Met** R L F **Met** T Y G E V E E C K P **Met** D A E D C A E F T D V Y W
I K F R L I T N A R K L D E S S F L G N R L Q I S Y A P E Y E S V S D
T K E K L E T R R K E V L A R L N P H K A K S T S Q V T K L A G P A
L T Q T D N F S P R R R E **Met** D Y Q F H R G N A P V T R V S S Y Q
E Y F A S S S **Met** N Q **Met** V K T V R E K L N K I E E S G N Q K R L
Q P S S Q T Q P D F K R T R V D N R R R I **Stop**

PPR2a nucleotide and amino-acid sequences:

```
atggctgcaactaggtctacatttctaggctccatcttcagaaccgttaaagcacgagtt
 M   A   A   T   R   S   T   F   L   G   S   I   F   R   T   V   K   A   R   V
ttgatcccacccatcacaagactctctctcccaaaccctaatttcacttcctctcgcttc
 L   I   P   P   I   T   R   L   S   L   P   N   P   N   F   T   S   S   R   F
cacacttcttcttctctgcctcagtctcaaatcgaaggttccatttcgagttctcttctg
 H   T   S   S   S   L   P   Q   S   Q   I   E   G   S   I   S   S   S   L   L
caagacctgagcatcatccatggtgtgattgcgattgcttccactagtcaaggtcgatta
 Q   D   L   S   I   I   H   G   V   I   A   I   A   S   T   S   Q   G   R   L
gcatctgtttccatgtgtacacaatactcaacttcagtcccaacacgttcacttagaagg
 A   S   V   S   M   C   T   Q   Y   S   T   S   V   P   T   R   S   L   R   R
agaatcagcagtagaaagaaatcgagtacaaaaccaatccttaatgaatcaaagtttcaa
 R   I   S   S   R   K   K   S   S   T   K   P   I   L   N   E   S   K   F   Q
gaaacgatatcaaagcttccaccaagattcacacctgaagaactagctgatgctataact
 E   T   I   S   K   L   P   P   R   F   T   P   E   E   L   A   D   A   I   T
cttgaagaagacccgtttctgtgtttccatctctttaactgggcatcgcaacagccgagg
 L   E   E   D   P   F   L   C   F   H   L   F   N   W   A   S   Q   Q   P   R
tttacgcatgagaattgctcttaccatatcgcgataaggaagctcggtgctgcgaaatct
 F   T   H   E   N   C   S   Y   H   I   A   I   R   K   L   G   A   A   K   S
gggaagttgatacgtgctgtgaatatatttagacatatggtgaatagtaggaacttggaa
```

```
 G  K  L  I  R  A  V  N  I  F  R  H  M  V  N  S  R  N  L  E
tgtagaccaacgatgagaacgtatcatattctcttcaaagcattgttgggtagaggtaac
 C  R  P  T  M  R  T  Y  H  I  L  F  K  A  L  L  G  R  G  N
aactctttcataaaccatctgtatatggagacagtaagatctttgtttcgacaaatggtg
 N  S  F  I  N  H  L  Y  M  E  T  V  R  S  L  F  R  Q  M  V
gatagtggtattgaaccagacgtatttgctttgaactgtttggttaaaggtaggaccatc
 D  S  G  I  E  P  D  V  F  A  L  N  C  L  V  K  G  R  T  I
aatactagagaattgcttagtgagatgaaaggaaaaggctttgttcctaatgggaaatct
 N  T  R  E  L  L  S  E  M  K  G  K  G  F  V  P  N  G  K  S
tataactctctggttaatgcttttgcgcttagtggtgagattgatgatgcggtgaaatgt
 Y  N  S  L  V  N  A  F  A  L  S  G  E  I  D  D  A  V  K  C
ttgtgggagatgattgagaatggtcgtgtggttgattttattagctatagaacacttgtt
 L  W  E  M  I  E  N  G  R  V  V  D  F  I  S  Y  R  T  L  V
gatgagagttgtaggaaagggaagtatgatgaagcgacgagattgttggagatgttgcga
 D  E  S  C  R  K  G  K  Y  D  E  A  T  R  L  L  E  M  L  R
gagaaacagcttgtggatatagattctgatgataagcttaagatgtaccaaatggtgatt
 E  K  Q  L  V  D  I  D  S  D  D  K  L  K  M  Y  Q  M  V  I
cttgttcttcttttctcctcaatgttgccttcagtttgcgatgaatcaagatacatgata
 L  V  L  L  F  S  S  M  L  P  S  V  C  D  E  S  R  Y  M  I
gtacgtaatgtgccagctttgggttgtggtgatgatctcatgagattattcatgacttat
 V  R  N  V  P  A  L  G  C  G  D  D  L  M  R  L  F  M  T  Y
ggagaagttgaagaatgtaaacctatggatgcagaagactgtgcggagttcacggatgtc
 G  E  V  E  E  C  K  P  M  D  A  E  D  C  A  E  F  T  D  V
tactggatcaagtttcgtctcatcactaatgctaggaagttggatgaatcaagtttttg
 Y  W  I  K  F  R  L  I  T  N  A  R  K  L  D  E  S  S  F  L
ggaaatcggctccaaatctcatatgctcctgaatacgagagcgtcagtgacacaaaggag
 G  N  R  L  Q  I  S  Y  A  P  E  Y  E  S  V  S  D  T  K  E
aagttagaaactagaaggaaagaagtgcttgcaagactgaacccccataaagcaaagagc
 K  L  E  T  R  R  K  E  V  L  A  R  L  N  P  H  K  A  K  S
acctcccaagttacaaaattggctggaccagctttgacccaaaccgacaatttttccccct
 T  S  Q  V  T  K  L  A  G  P  A  L  T  Q  T  D  N  F  S  P
cggcggagagagatggattaccaattccatagaggaaatgctcctgttactcgagtttca
 R  R  R  E  M  D  Y  Q  F  H  R  G  N  A  P  V  T  R  V  S
tcatatcaggagtattttgcgtcatcttcaatgaatcagatggttaaaactgtgagggag
 S  Y  Q  E  Y  F  A  S  S  S  M  N  Q  M  V  K  T  V  R  E
aaactcaataagattgaagaaagtggtaaccaaaagaggttacaaccaagcagccaaaca
 K  L  N  K  I  E  E  S  G  N  Q  K  R  L  Q  P  S  S  Q  T
caacctgacttcaagagaacccgagtcgataaccgaagaagaatataa
 Q  P  D  F  K  R  T  R  V  D  N  R  R  R  I  -
```

PPR2b amino-acid sequences:

G C N **Stop** V Y I S R L H L Q N R **Stop** S T S F D P T H H K T L S P
K P **Stop** F H F L S L P H F F F S A S V S N R R F H F E F R F L F E
S G S S L L Q D L S I I H G V I A I A S T S Q G R L A S V S **Met** C T Q
Y S T S V P T R S L R R R I S S R K K S S T K P I L N E S K F Q E T I
S K L P P R F T P E E L A D A I T L E E D P F L C F H L F N W A S Q
Q P R F T H E N C S Y H I A I R K L G A A K S G K L I R A V N I F R
H **Met** V N S R N L E C R P T **Met** R T Y H I L F K A L L G R G N N S F
I N H L Y **Met** E T V R S L F R Q **Met** V D S G I E P D V F A L N C L V
K G R T I N T R E L L S E **Met** K G K G F V P N G K S Y N S L V N A F
A L S G E I D D A V K C L W E **Met** I E N G R V V D F I S Y R T L V D

E S C R K G K Y D E A T R L L E **Met** L R E K Q L V D I D S D D K L
K **Met** Y Q **Met** V I L V L L F S S **Met** L P S V C D E S R Y **Met** I V R N
V P A L G C G D D L **Met** R L F **Met** T Y G E V E E C K P **Met** D A E D
C A E F T D V Y W I K F R L I T N A R K L D E S S F L G N R L Q I S Y
A P E Y E S V S D T K E K L E T R R K E V L A R L N P H K A K S T S
Q V T K L A G P A L T Q T D N F S P R R R E **Met** D Y Q F H R G N A
P V T R V S S Y Q E Y F A S S S **Met** N Q **Met** V K T V R E K L N K I
E E S G N Q K R L Q P S S Q T Q P D F K R T R V D N R R R I **Stop**

PPR2b nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atggctgcaactaggtctacatttctaggctccatcttcagaaccgttaaagcacgagtttt
   G  C  N  -  V  Y  I  S  R  L  H  L  Q  N  R  -  S  T  S  F
gatcccacccatcacaagactctctctcccaaaccctaatttcacttcctctcgcttcca
 D  P  T  H  H  K  T  L  S  P  K  P  -  F  H  F  L  S  L  P
cacttcttcttctctgcctcagtctcaaatcgaaggttccatttcgagtttcgattttta
 H  F  F  F  S  A  S  V  S  N  R  R  F  H  F  E  F  R  F  L
ttcgaatccggaagttctcttctgcaagacctgagcatcatccatggtgtgattgcgatt
 F  E  S  G  S  S  L  L  Q  D  L  S  I  I  H  G  V  I  A  I
gcttccactagtcaaggtcgattagcatctgtttccatgtgtacacaatactcaacttca
 A  S  T  S  Q  G  R  L  A  S  V  S  M  C  T  Q  Y  S  T  S
gtcccaacacgttcacttagaaggagaatcagcagtagaaagaaatcgagtacaaaacca
 V  P  T  R  S  L  R  R  R  I  S  S  R  K  K  S  S  T  K  P
atccttaatgaatcaaagtttcaagaaacgatatcaaagcttccaccaagattcacacct
 I  L  N  E  S  K  F  Q  E  T  I  S  K  L  P  P  R  F  T  P
gaagaactagctgatgctataactcttgaagaagacccgtttctgtgtttccatctcttt
 E  E  L  A  D  A  I  T  L  E  E  D  P  F  L  C  F  H  L  F
aactgggcatcgcaacagccgaggtttacgcatgagaattgctcttaccatatcgcgata
 N  W  A  S  Q  Q  P  R  F  T  H  E  N  C  S  Y  H  I  A  I
aggaagctcggtgctgcgaaatctgggaagttgatacgtgctgtgaatatatttagacat
 R  K  L  G  A  A  K  S  G  K  L  I  R  A  V  N  I  F  R  H
atggtgaatagtaggaacttggaatgtagaccaacgatgagaacgtatcatattctcttc
 M  V  N  S  R  N  L  E  C  R  P  T  M  R  T  Y  H  I  L  F
aaagcattgttgggtagaggtaacaactctttcataaaccatctgtatatggagacagta
 K  A  L  L  G  R  G  N  N  S  F  I  N  H  L  Y  M  E  T  V
agatctttgtttcgacaaatggtggatagtggtattgaaccagacgtatttgctttgaac
 R  S  L  F  R  Q  M  V  D  S  G  I  E  P  D  V  F  A  L  N
tgtttggttaaaggtaggaccatcaatactagagaattgcttagtgagatgaaaggaaaa
 C  L  V  K  G  R  T  I  N  T  R  E  L  L  S  E  M  K  G  K
ggctttgttcctaatgggaaatcttataactctctggttaatgcttttgcgcttagtggt
 G  F  V  P  N  G  K  S  Y  N  S  L  V  N  A  F  A  L  S  G
gagattgatgatgcggtgaaatgtttgtgggagatgattgagaatggtcgtgtggttgat
 E  I  D  D  A  V  K  C  L  W  E  M  I  E  N  G  R  V  V  D
tttattagctatagaacacttgttgatgagagttgtaggaaagggaagtatgatgaagcg
 F  I  S  Y  R  T  L  V  D  E  S  C  R  K  G  K  Y  D  E  A
acgagattgttggagatgttgcgagagaaacagcttgtggatatagattctgatgataag
 T  R  L  L  E  M  L  R  E  K  Q  L  V  D  I  D  S  D  D  K
cttaagatgtaccaaatggtgattcttgttcttcttttctcctcaatgttgccttcagtt
 L  K  M  Y  Q  M  V  I  L  V  L  L  F  S  S  M  L  P  S  V
tgcgatgaatcaagatacatgatagtacgtaatgtgccagctttgggttgtggtgatgat
 C  D  E  S  R  Y  M  I  V  R  N  V  P  A  L  G  C  G  D  D
ctcatgagattattcatgacttatggagaagttgaagaatgtaaacctatggatgcagaa
 L  M  R  L  F  M  T  Y  G  E  V  E  E  C  K  P  M  D  A  E
```

```
gactgtgcggagttcacggatgtctactggatcaagtttcgtctcatcactaatgctagg
 D  C  A  E  F  T  D  V  Y  W  I  K  F  R  L  I  T  N  A  R
aagttggatgaatcaagttttttgggaaatcggctccaaatctcatatgctcctgaatac
 K  L  D  E  S  S  F  L  G  N  R  L  Q  I  S  Y  A  P  E  Y
gagagcgtcagtgacacaaaggagaagttagaaactagaaggaaagaagtgcttgcaaga
 E  S  V  S  D  T  K  E  K  L  E  T  R  R  K  E  V  L  A  R
ctgaacccccataaagcaaagagcacctcccaagttacaaaattggctggaccagctttg
 L  N  P  H  K  A  K  S  T  S  Q  V  T  K  L  A  G  P  A  L
acccaaaccgacaatttttccccctcggcggagagagatggattaccaattccatagagga
 T  Q  T  D  N  F  S  P  R  R  R  E  M  D  Y  Q  F  H  R  G
aatgctcctgttactcgagtttcatcatatcaggagtattttgcgtcatcttcaatgaat
 N  A  P  V  T  R  V  S  S  Y  Q  E  Y  F  A  S  S  S  M  N
cagatggttaaaactgtgagggagaaactcaataagattgaagaaagtggtaaccaaaag
 Q  M  V  K  T  V  R  E  K  L  N  K  I  E  E  S  G  N  Q  K
aggttacaaccaagcagccaaacacaacctgacttcaagagaacccgagtcgataaccga
 R  L  Q  P  S  S  Q  T  Q  P  D  F  K  R  T  R  V  D  N  R
agaagaatataa
 R  R  I  -
```

In PPR5, an 84 nt addition to the reference isoform PPR5a, due to intron retention between exons 7 and 8 (Figure 5A and C) led to the formation of the longer isoform, PPR5b (Table 2), encoding an additional 28 amino-acid.

PPR5a amino-acid sequences:

Met A V T I S T N A F V N A S L L D E S R N S F W R P L F H Q P Y Y
N C R R V V R L N S R K L N S K V **Met** F C L N L N T K E V G L Q K
P G D K G F E F K P S F D Q Y L Q I **Met** E S V K T A R K K K K F D
R L K V E E D D G G G G N G D S V Y E V K D **Met** K I K S G E L K D
E T F R K R Y S R Q E I V S D K R N E R V F K R N G E I E N H R V
A T D L K W S K S G E S S V A L K L S K S G E S S V T V P E D E S
F R K R Y S K Q E Y H R S S D T S R G I E R G S R G D E L D L V V
E E R R V Q R I A K D A R W S K S R E S S V A V K W S N S G E S S
V T **Met** P K D E S F R R R Y S K Q E H H R S S D T S R G I A R G S
K G D E L E L V V E E R R V Q R I A K D V R W S K S D E S L V P V
S E D E S F R R G N P K Q E **Met** V R Y Q R V S D T S R G I E R G S
K G D G L D L L A E E R R I E R L A N E R H E I R S S K L S G T R R
I G A K R N D D D D D S L F A **Met** E T P A F R F S D E S S D I V D K
P A T S R V E **Met** E D R I E K L A K V L N G A D I N **Met** P E W Q F S
K A I R S A K I R Y T D Y T V **Met** R L I H F L G K L G N W R R V L Q
V I E W L Q R Q D R Y K S N K I R I I Y T T A L N V L G K S R R P V E
A L N V F H A **Met** L L Q I S S Y P D **Met** V A Y R S I A V T L G Q A G
H I K E L F Y V I D T **Met** R S P P K K K F K P T T L E K W D P R L E
P D V V V Y N A V L N A C V Q R K Q W E G A F W V L Q Q L K Q R

G Q K P S P V T Y G L I **Met** E V **Met** L A C E K Y N L V H E F F R
K **Met** Q K S S I P N A L A Y R V L V N T L W K E G K S D E A V H T V
E D **Met** E S R G I V G S A A L Y Y D L A R C L C S A G R C N E G L
N **Met** L K K I C R V A N K P L V V T Y T G L I Q A C V D S G N I K N A
A Y I F D Q **Met** K K V C S P N L V T C N I **Met** L K A Y L Q G G L F E
E A R E L F Q K **Met** S E D G N H I K N S S D F E S R V L P D T Y T F
N T **Met** L D T C A E Q E K W D D F G Y A Y R E **Met** L R H G Y H F N
A K R H L R **Met** V L E A S R A G K E E V **Met** E A T W E H **Met** R R S
N R I P P S P L I K E R F F R K L E K G D H I S A I S S L A D L N G K I
E E T E L R A F S T S A W S R V L S R F E Q D S V L R L **Met** D D V
N R R L G S R S E S S D S V L G N L L S S C K D Y L K T R T H N
L **Stop**

PPR5a nucleotide and amino-acid sequences:

```
atggcggtgacgatttcgacgaatgctttcgtaaatgcatcgcttttagatgaaagtcgg
 M  A  V  T  I  S  T  N  A  F  V  N  A  S  L  L  D  E  S  R
aattctttctggagaccattgtttcatcagccatactataattgccgacgagtcgttcgt
 N  S  F  W  R  P  L  F  H  Q  P  Y  Y  N  C  R  R  V  V  R
cttaattcgaggaaattgaattcaaaggtaatgttttgcttgaatttgaacacgaaggag
 L  N  S  R  K  L  N  S  K  V  M  F  C  L  N  L  N  T  K  E
gttggtttgcaaaaacccggtgataaaggttttgaattcaaacccagttttgatcagtac
 V  G  L  Q  K  P  G  D  K  G  F  E  F  K  P  S  F  D  Q  Y
ctgcaaatcatggaatcggttaaaacagcaaggaagaagaagaaattcgacagattgaaa
 L  Q  I  M  E  S  V  K  T  A  R  K  K  K  K  F  D  R  L  K
gttgaggaagatgatggtggaggtgggaatggtgatagtgtttatgaagtgaaagatatg
 V  E  E  D  D  G  G  G  G  N  G  D  S  V  Y  E  V  K  D  M
aagattaagagtggtgagctaaaagatgaaactttcaggaagagatactcaaggcaggag
 K  I  K  S  G  E  L  K  D  E  T  F  R  K  R  Y  S  R  Q  E
attgtaagtgataaacgtaatgagagagttttcaagaggaatggagaaattgaaaatcat
 I  V  S  D  K  R  N  E  R  V  F  K  R  N  G  E  I  E  N  H
agagtggctacagatttgaaatggagtaagagtggtgaatcttcagtggctctgaaattg
 R  V  A  T  D  L  K  W  S  K  S  G  E  S  S  V  A  L  K  L
agtaagagtggtgaatcttcagtgactgtgcctgaagatgagagtttcaggaaaaggtac
 S  K  S  G  E  S  S  V  T  V  P  E  D  E  S  F  R  K  R  Y
tctaagcaggagtatcaccgttcctctgatacatcgagagggattgaaagaggttcgaga
 S  K  Q  E  Y  H  R  S  S  D  T  S  R  G  I  E  R  G  S  R
ggtgatgaattggatcttgttgttgaagaaaggagagttcagagaatagccaaagatgca
 G  D  E  L  D  L  V  V  E  E  R  R  V  Q  R  I  A  K  D  A
agatggagtaaaagtcgtgaatcttcagtggctgtgaaatggagtaatagtggtgaatct
 R  W  S  K  S  R  E  S  S  V  A  V  K  W  S  N  S  G  E  S
tcagtgactatgcctaaagatgagagctttaggagaagatactctaagcaggagcatcac
 S  V  T  M  P  K  D  E  S  F  R  R  R  Y  S  K  Q  E  H  H
cgttcctctgatacatccagagggattgcaagaggttcaaaaggtgatgaattggagctt
 R  S  S  D  T  S  R  G  I  A  R  G  S  K  G  D  E  L  E  L
gttgttgaagaaaggagagttcagagaatagccaaagatgtaagatggagtaagagtgat
 V  V  E  E  R  R  V  Q  R  I  A  K  D  V  R  W  S  K  S  D
gaatctttagtgcctgtgtcagaagatgagagtttcaggagagggaatccgaagcaggag
 E  S  L  V  P  V  S  E  D  E  S  F  R  R  G  N  P  K  Q  E
atggtgaggtatcagcgtgtctctgatacatcgagagggattgagagaggttccaaagga
 M  V  R  Y  Q  R  V  S  D  T  S  R  G  I  E  R  G  S  K  G
gatggattggatcttcttgctgaagaaaggcggattgagagattagccaatgagaggcat
```

```
 D  G  L  D  L  L  A  E  E  R  R  I  E  R  L  A  N  E  R  H
gagataagaagtagcaaattgagtggaaccaggagaattggtgctaagagaaatgatgat
 E  I  R  S  S  K  L  S  G  T  R  R  I  G  A  K  R  N  D  D
gatgatgatagcttgtttgccatggaaactcctgcctttaggttttctgatgagtccagt
 D  D  D  S  L  F  A  M  E  T  P  A  F  R  F  S  D  E  S  S
gacatagtggacaagccagctacttcacgagtcgaaatggaagacagaatcgagaagtta
 D  I  V  D  K  P  A  T  S  R  V  E  M  E  D  R  I  E  K  L
gcaaaagtgttgaatggtgcagacatcaatatgcctgagtggcagttttccaaggcgatc
 A  K  V  L  N  G  A  D  I  N  M  P  E  W  Q  F  S  K  A  I
aggagtgcaaaaatcagatatacggattacacagtaatgagactgatccactttctaggg
 R  S  A  K  I  R  Y  T  D  Y  T  V  M  R  L  I  H  F  L  G
aaactaggaaactggagacgagttcttcaagtcattgagtggcttcaaaggcaagaccgt
 K  L  G  N  W  R  R  V  L  Q  V  I  E  W  L  Q  R  Q  D  R
tacaaatctaacaagataagaatcatctatacaactgcactaaatgttcttggtaaatca
 Y  K  S  N  K  I  R  I  I  Y  T  T  A  L  N  V  L  G  K  S
aggaggcctgtggaagctctcaatgtattccacgctatgctgttacaaatttcatcatat
 R  R  P  V  E  A  L  N  V  F  H  A  M  L  L  Q  I  S  S  Y
ccggatatggtagcataccgttcaattgcagtcacacttggacaagctgggcatatcaag
 P  D  M  V  A  Y  R  S  I  A  V  T  L  G  Q  A  G  H  I  K
gaactcttctatgtgattgacacaatgaggtctccacctaaaaagaagttcaagccaaca
 E  L  F  Y  V  I  D  T  M  R  S  P  P  K  K  K  F  K  P  T
acacttgaaaaatgggatccccggcttgaaccagatgttgttgtttacaatgcggtgctc
 T  L  E  K  W  D  P  R  L  E  P  D  V  V  V  Y  N  A  V  L
aacgcatgtgttcaacgaaagcaatgggaaggagcattctgggtattgcaacagttgaag
 N  A  C  V  Q  R  K  Q  W  E  G  A  F  W  V  L  Q  Q  L  K
caacgagggcaaaaaccttctcctgtaacctatggcctcatcatggaggtaatgttagca
 Q  R  G  Q  K  P  S  P  V  T  Y  G  L  I  M  E  V  M  L  A
tgtgagaagtacaatttagttcatgaattcttcaggaagatgcagaaatcttctatccct
 C  E  K  Y  N  L  V  H  E  F  F  R  K  M  Q  K  S  S  I  P
aatgctctagcatatagagttcttgttaatactctatggaaagaaggtaaaagcgacgag
 N  A  L  A  Y  R  V  L  V  N  T  L  W  K  E  G  K  S  D  E
gccgtacatacggttgaggatatggaaagccgtggtattgttggatcagctgctctttac
 A  V  H  T  V  E  D  M  E  S  R  G  I  V  G  S  A  A  L  Y
tacgaccttgctcgctgtctatgtagcgcaggaaggtgtaatgaagggctcaatatgctt
 Y  D  L  A  R  C  L  C  S  A  G  R  C  N  E  G  L  N  M  L
aagaagatatgtagagttgcaaataaacctctcgtggtgacttacactggcctgatccaa
 K  K  I  C  R  V  A  N  K  P  L  V  V  T  Y  T  G  L  I  Q
gcatgcgtcgactcgggaaacatcaagaatgcagcttacatcttcgatcagatgaagaag
 A  C  V  D  S  G  N  I  K  N  A  A  Y  I  F  D  Q  M  K  K
gtctgcagccctaacctagtcacttgcaacataatgctaaaagcttatctacaaggcgga
 V  C  S  P  N  L  V  T  C  N  I  M  L  K  A  Y  L  Q  G  G
ttgtttgaagaagcaagggaacttttccagaagatgtcagaagacggaaatcatataaaa
 L  F  E  E  A  R  E  L  F  Q  K  M  S  E  D  G  N  H  I  K
aacagctcggatttcgaatcaagagtattgccagacacgtacacgttcaacacgatgcta
 N  S  S  D  F  E  S  R  V  L  P  D  T  Y  T  F  N  T  M  L
gacacgtgtgctgaacaagaaaagtgggatgattttggttatgcgtatcgggagatgttg
 D  T  C  A  E  Q  E  K  W  D  D  F  G  Y  A  Y  R  E  M  L
cgtcatggataccatttcaatgcgaaacgccatctcagaatggtacttgaagctagcaga
 R  H  G  Y  H  F  N  A  K  R  H  L  R  M  V  L  E  A  S  R
gcaggaaaggaggaggtgatggaagcgacatgggaacacatgagacgaagtaatagaatt
 A  G  K  E  E  V  M  E  A  T  W  E  H  M  R  R  S  N  R  I
ccgccatcgcctctaatcaaagaaagattcttcaggaaactcgagaaaggcgatcatatt
 P  P  S  P  L  I  K  E  R  F  F  R  K  L  E  K  G  D  H  I
tcggctatatcatctcttgctgatcttaatggaaaaattgaggagactgagttacgagca
 S  A  I  S  S  L  A  D  L  N  G  K  I  E  E  T  E  L  R  A
ttttcaacttccgcatggtccagagtcttgtcccgatttgagcaagattcagttttgagg
 F  S  T  S  A  W  S  R  V  L  S  R  F  E  Q  D  S  V  L  R
ttaatggacgatgtgaacagacgcctaggttcgagaagtgagtcttcggattcggttttg
```

```
 L  M  D  D  V  N  R  R  L  G  S  R  S  E  S  S  D  S  V  L
gggaatctattgagttcttgtaaagattatctgaagaccagaacacataacttgtaa
 G  N  L  L  S  S  C  K  D  Y  L  K  T  R  T  H  N  L  -
```

PPR5b amino-acid sequences, yellow mark indicates spliced sequence.

Met A V T I S T N A F V N A S L L D E S R N S F W R P L F H Q P Y Y
N C R R V V R L N S R K L N S K V **Met** F C L N L N T K E V G L Q K
P G D K G F E F K P S F D Q Y L Q I **Met** E S V K T A R K K K K F D
R L K V E E D D G G G G N G D S V Y E V K D **Met** K I K S G E L K D
E T F R K R Y S R Q E I V S D K R N E R V F K R N G E I E N H R V
A T D L K W S K S G E S S V A L K L S K S G E S S V T V P E D E S
F R K R Y S K Q E Y H R S S D T S R G I E R G S R G D E L D L V V
E E R R V Q R I A K D A R W S K S R E S S V A V K W S N S G E S S
V T **Met** P K D E S F R R R Y S K Q E H H R S S D T S R G I A R G S
K G D E L E L V V E E R R V Q R I A K D V R W S K S D E S L V P V
S E D E S F R R G N P K Q E **Met** V R Y Q R V S D T S R G I E R G S
K G D G L D L L A E E R R I E R L A N E R H E I R S S K L S G T R R
I G A K R N D D D D D D S L F A **Met** E T P A F R F S D E S S D I V D K
P A T S R V E **Met** E D R I E K L A K V L N G A D I N **Met** P E W Q F S
K A I R S A K I R Y T D Y T V **Met** R L I H F L G K L G N W R R V L Q
V I E W L Q R Q D R Y K S N K I R I I Y T T A L N V L G K S R R P V E
A L N V F H A **Met** L L Q I S S Y P D **Met** V A Y R S I A V T L G Q A G
H I K E L F Y V I D T **Met** R S P P K K K F K P T T L E K W D P R L E
P D V V V Y N A V L N A C V Q R K Q W E G A F W V L Q Q L K Q R
G Q K P S P V T Y G L I **Met** E V **Met** L A C E K Y N L V H E F F R
K **Met** Q K S S I P N A L A Y R V L V N T L W K E G K S D E A V H T V
E D **Met** E S R G I V G S A A L Y Y D L A R C L C S A G R C N E G L
N **Met** V N F V N P V V L K L I E N L I Y K A D L V H T I Q F Q L K K I
C R V A N K P L V V T Y T G L I Q A C V D S G N I K N A A Y I F D
Q **Met** K K V C S P N L V T C N I **Met** L K A Y L Q G G L F E E A R E L
F Q K **Met** S E D G N H I K N S S D F E S R V L P D T Y T F N
T **Met** L D T C A E Q E K W D D F G Y A Y R E **Met** L R H G Y H F N A
K R H L R **Met** V L E A S R A G K E E V **Met** E A T W E H **Met** R R S N
R I P P S P L I K E R F F R K L E K G D H I S A I S S L A D L N G K I E
E T E L R A F S T S A W S R V L S R F E Q D S V L R L **Met** D D V N
R R L G S R S E S S D S V L G N L L S S C K D Y L K T R T H N
L **Stop**

PPR5b nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atggcggtgacgatttcgacgaatgctttcgtaaatgcatcgcttttagatgaaagtcgg
 M  A  V  T  I  S  T  N  A  F  V  N  A  S  L  L  D  E  S  R
aattctttctggagaccattgtttcatcagccatactataattgccgacgagtcgttcgt
 N  S  F  W  R  P  L  F  H  Q  P  Y  Y  N  C  R  R  V  V  R
cttaattcgaggaaattgaattcaaaggtaatgttttgcttgaatttgaacacgaaggag
 L  N  S  R  K  L  N  S  K  V  M  F  C  L  N  L  N  T  K  E
gttggtttgcaaaaacccggtgataaaggttttgaattcaaacccagttttgatcagtac
 V  G  L  Q  K  P  G  D  K  G  F  E  F  K  P  S  F  D  Q  Y
ctgcaaatcatggaatcggttaaaacagcaaggaagaagaagaaattcgacagattgaaa
 L  Q  I  M  E  S  V  K  T  A  R  K  K  K  K  F  D  R  L  K
gttgaggaagatgatggtggaggtgggaatggtgatagtgtttatgaagtgaaagatatg
 V  E  E  D  D  G  G  G  G  N  G  D  S  V  Y  E  V  K  D  M
aagattaagagtggtgagctaaaagatgaaactttcaggaagagatactcaaggcaggag
 K  I  K  S  G  E  L  K  D  E  T  F  R  K  R  Y  S  R  Q  E
attgtaagtgataaacgtaatgagagagttttcaagaggaatggagaaattgaaaatcat
 I  V  S  D  K  R  N  E  R  V  F  K  R  N  G  E  I  E  N  H
agagtggctacagatttgaaatggagtaagagtggtgaatcttcagtggctctgaaattg
 R  V  A  T  D  L  K  W  S  K  S  G  E  S  S  V  A  L  K  L
agtaagagtggtgaatcttcagtgactgtgcctgaagatgagagtttcaggaaaaggtac
 S  K  S  G  E  S  S  V  T  V  P  E  D  E  S  F  R  K  R  Y
tctaagcaggagtatcaccgttcctctgatacatcgagagggattgaaagaggttcgaga
 S  K  Q  E  Y  H  R  S  S  D  T  S  R  G  I  E  R  G  S  R
ggtgatgaattggatcttgttgttgaagaaaggagagttcagagaatagccaaagatgca
 G  D  E  L  D  L  V  V  E  E  R  R  V  Q  R  I  A  K  D  A
agatggagtaaaagtcgtgaatcttcagtggctgtgaaatggagtaatagtggtgaatct
 R  W  S  K  S  R  E  S  S  V  A  V  K  W  S  N  S  G  E  S
tcagtgactatgcctaaagatgagagctttaggagaagatactctaagcaggagcatcac
 S  V  T  M  P  K  D  E  S  F  R  R  R  Y  S  K  Q  E  H  H
cgttcctctgatacatccagagggattgcaagaggttcaaaaggtgatgaattggagctt
 R  S  S  D  T  S  R  G  I  A  R  G  S  K  G  D  E  L  E  L
gttgttgaagaaaggagagttcagagaatagccaaagatgtaagatggagtaagagtgat
 V  V  E  E  R  R  V  Q  R  I  A  K  D  V  R  W  S  K  S  D
gaatctttagtgcctgtgtcagaagatgagagtttcaggagagggaatccgaagcaggag
 E  S  L  V  P  V  S  E  D  E  S  F  R  R  G  N  P  K  Q  E
atggtgaggtatcagcgtgtctctgatacatcgagagggattgagagaggttccaaagga
 M  V  R  Y  Q  R  V  S  D  T  S  R  G  I  E  R  G  S  K  G
gatggattggatcttcttgctgaagaaaggcggattgagagattagccaatgagaggcat
 D  G  L  D  L  L  A  E  E  R  R  I  E  R  L  A  N  E  R  H
gagataagaagtagcaaattgagtggaaccaggagaattggtgctaagagaaatgatgat
 E  I  R  S  S  K  L  S  G  T  R  R  I  G  A  K  R  N  D  D
gatgatgatagcttgtttgccatggaaactcctgcctttaggttttctgatgagtccagt
 D  D  D  S  L  F  A  M  E  T  P  A  F  R  F  S  D  E  S  S
gacatagtggacaagccagctacttcacgagtcgaaatggaagacagaatcgagaagtta
 D  I  V  D  K  P  A  T  S  R  V  E  M  E  D  R  I  E  K  L
gcaaaagtgttgaatggtgcagacatcaatatgcctgagtggcagttttccaaggcgatc
 A  K  V  L  N  G  A  D  I  N  M  P  E  W  Q  F  S  K  A  I
aggagtgcaaaaatcagatatacggattacacagtaatgagactgatccactttctaggg
 R  S  A  K  I  R  Y  T  D  Y  T  V  M  R  L  I  H  F  L  G
aaactaggaaactggagacgagttcttcaagtcattgagtggcttcaaaggcaagaccgt
 K  L  G  N  W  R  R  V  L  Q  V  I  E  W  L  Q  R  Q  D  R
tacaaatctaacaagataagaatcatctatacaactgcactaaatgttcttggtaaatca
 Y  K  S  N  K  I  R  I  I  Y  T  T  A  L  N  V  L  G  K  S
aggaggcctgtggaagctctcaatgtattccacgctatgctgttacaaatttcatcatat
 R  R  P  V  E  A  L  N  V  F  H  A  M  L  L  Q  I  S  S  Y
ccggatatggtagcataccgttcaattgcagtcacacttggacaagctgggcatatcaag
```

```
 P  D  M  V  A  Y  R  S  I  A  V  T  L  G  Q  A  G  H  I  K
gaactcttctatgtgattgacacaatgaggtctccacctaaaaagaagttcaagccaaca
 E  L  F  Y  V  I  D  T  M  R  S  P  P  K  K  K  F  K  P  T
acacttgaaaaatgggatccccggcttgaaccagatgttgttgtttacaatgcggtgctc
 T  L  E  K  W  D  P  R  L  E  P  D  V  V  V  Y  N  A  V  L
aacgcatgtgttcaacgaaagcaatgggaaggagcattctgggtattgcaacagttgaag
 N  A  C  V  Q  R  K  Q  W  E  G  A  F  W  V  L  Q  Q  L  K
caacgagggcaaaaaccttctcctgtaacctatggcctcatcatggaggtaatgttagca
 Q  R  G  Q  K  P  S  P  V  T  Y  G  L  I  M  E  V  M  L  A
tgtgagaagtacaatttagttcatgaattcttcaggaagatgcagaaatcttctatccct
 C  E  K  Y  N  L  V  H  E  F  F  R  K  M  Q  K  S  S  I  P
aatgctctagcatatagagttcttgttaatactctatggaaagaaggtaaaagcgacgag
 N  A  L  A  Y  R  V  L  V  N  T  L  W  K  E  G  K  S  D  E
gccgtacatacggttgaggatatggaaagccgtggtattgttggatcagctgctctttac
 A  V  H  T  V  E  D  M  E  S  R  G  I  V  G  S  A  A  L  Y
tacgaccttgctcgctgtctatgtagcgcaggaaggtgtaatgaagggctcaatatggta
 Y  D  L  A  R  C  L  C  S  A  G  R  C  N  E  G  L  N  M  V
aattttgtaaatcctgtagtccttaagcttatcgagaatttgatttacaaagctgatctt
 N  F  V  N  P  V  V  L  K  L  I  E  N  L  I  Y  K  A  D  L
gttcataccatccaatttcagcttaagaagatatgtagagttgcaaataaacctctcgtg
 V  H  T  I  Q  F  Q  L  K  K  I  C  R  V  A  N  K  P  L  V
gtgacttacactggcctgatccaagcatgcgtcgactcgggaaacatcaagaatgcagct
 V  T  Y  T  G  L  I  Q  A  C  V  D  S  G  N  I  K  N  A  A
tacatcttcgatcagatgaagaaggtctgcagccctaacctagtcacttgcaacataatg
 Y  I  F  D  Q  M  K  K  V  C  S  P  N  L  V  T  C  N  I  M
ctaaaagcttatctacaaggcggattgtttgaagaagcaagggaacttttccagaagatg
 L  K  A  Y  L  Q  G  G  L  F  E  E  A  R  E  L  F  Q  K  M
tcagaagacggaaatcatataaaaaacagctcggatttcgaatcaagagtattgccagac
 S  E  D  G  N  H  I  K  N  S  S  D  F  E  S  R  V  L  P  D
acgtacacgttcaacacgatgctagacacgtgtgctgaacaagaaaagtgggatgatttt
 T  Y  T  F  N  T  M  L  D  T  C  A  E  Q  E  K  W  D  D  F
ggttatgcgtatcgggagatgttgcgtcatggataccatttcaatgcgaaacgccatctc
 G  Y  A  Y  R  E  M  L  R  H  G  Y  H  F  N  A  K  R  H  L
agaatggtacttgaagctagcagagcaggaaaggaggaggtgatggaagcgacatgggaa
 R  M  V  L  E  A  S  R  A  G  K  E  E  V  M  E  A  T  W  E
cacatgagacgaagtaatagaattccgccatcgcctctaatcaaagaaagattcttcagg
 H  M  R  R  S  N  R  I  P  P  S  P  L  I  K  E  R  F  F  R
aaactcgagaaaggcgatcatatttcggctatatcatctcttgctgatcttaatggaaaa
 K  L  E  K  G  D  H  I  S  A  I  S  S  L  A  D  L  N  G  K
attgaggagactgagttacgagcattttcaacttccgcatggtccagagtcttgtcccga
 I  E  E  T  E  L  R  A  F  S  T  S  A  W  S  R  V  L  S  R
tttgagcaagattcagttttgaggttaatggacgatgtgaacagacgcctaggttcgaga
 F  E  Q  D  S  V  L  R  L  M  D  D  V  N  R  R  L  G  S  R
agtgagtcttcggattcggttttggggaatctattgagttcttgtaaagattatctgaag
 S  E  S  S  D  S  V  L  G  N  L  L  S  S  C  K  D  Y  L  K
accagaacacataacttgtaa
 T  R  T  H  N  L  -
```

2.3.3 Predicted regulatory sequences involved in alternative splicing

Regulatory sequences allow transcription factors to bind directly or indirectly (with various associated factors) to their target RNAs (Narlikar and Ovcharenko 2009). Regulatory sequences regulate development and physiology by influencing gene expression. Any alteration in these elements can result in phenotypic diversity (Wittkopp and Kalay 2012). In the 3′ alternative-splicing site, I detected two *cis*-acting elements, i.e., intronic regulatory motifs ATGTTT and TTT, in *PPR1* and *PPR2*. In *PPR1a*, ATGTTT is located 148–154 nt downstream of the second intron, and another TTT sequence is repeated eight times upstream of exon 3. In *PPR2a*, ATGTTT is located 56–62 nt downstream of the first intron and is repeated five times upstream of exon 2. However, in *PPR4a*, the *cis*-acting element, ATGTTT, is present 6–12 nt upstream of the first intron and another TTT is repeated seven times upstream of exon 2 (Figure 8A). The transcriptional regulatory motif, TCAAT, is present once in *PPR1a* but twice in *PPR4a* (Figure 8B). Another reverse-complement transcriptional regulatory motif, ATTGA, is also present in *PPR2a* (Figure 8B). In the splice site in *PPR4a*, RNA sequencing data suggest that the first intron contains one donor site (GT) in the upstream region and two acceptor sites (AG) in the downstream region. The donor site is present at 81–82 nt, and the acceptor sites are found at 124–125 and 203–204 nt. However, in the current study, I detected the donor site at 81–82 nt but the acceptor site at 124–125 nt. In *PPR2a*, RNA sequencing data suggest that an alternative-splicing acceptor site is present at position 92–93 nt, affecting the 3′ end of the first intron; the same site was confirmed by my experimental results.

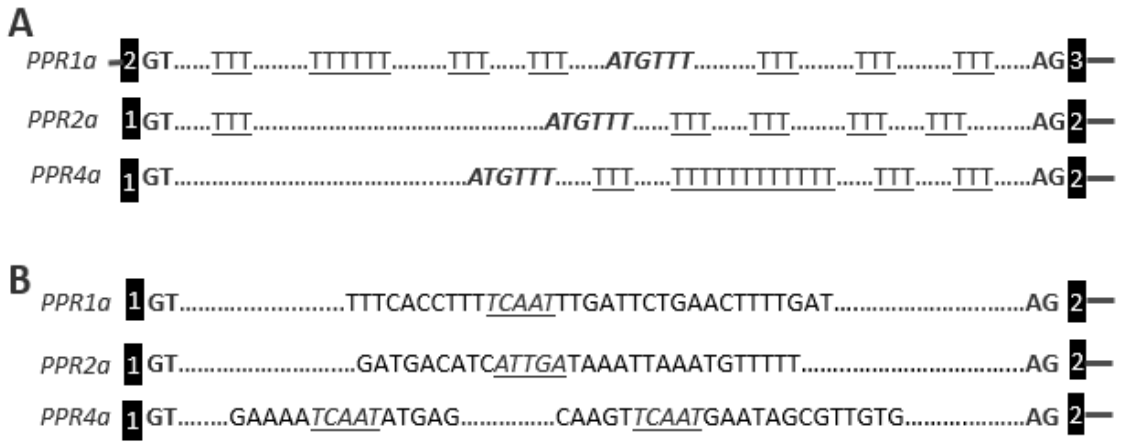


Figure 8. Predicted cis-regulatory elements affecting alternative splicing. **A**. Black boxes indicate exons, dots indicate intron sequences, and bold-italic sequences and underlined sequences indicate intronic regulatory motifs. **B**. Italics and underlined sequences indicate transcriptional regulatory motifs. **GT** and **AG** indicates **5'** and **3'** end of intron.

2.4 Discussion

Much is known about alternative splicing in humans and animals, but little is known about this process in plants. In plants, alternative splicing is a highly diversified process, which greatly affects transcript diversity; in Arabidopsis, even a single nucleotide exon has been reported (Guo and Liu 2015). In this study, I focused on alternative splicing of *PPR* genes. According to the http://www.uniprot.org database, *PPR1* binds with iron as a cofactor, but its biological function is unknown. *PPR2*, *PPR3*, *PPR4*, *PPR5*, and *PPR6* exhibit endonuclease activity and bind to RNA, and function in RNA modification. *PPR5* likely functions in plant embryo development (Cushing et al. 2005; Meinke et al. 2009). Another name for *PPR4* is CIHUATEOTL, a word from the Totonac language (Aguilar-Moreno 2007) meaning the goddess of fertility. This protein is expressed in the female gametophyte and is responsible for the initiation of gametogenesis (Greco et al. 2012). *PPR4* is predicted to be localized to the

mitochondria (Lurin et al. 2004).

The tissue-specific expression patterns of *PPR* genes differ due to alternative splicing. In the current study, I found that retained introns were highly expressed in all tissues than 3′ and 5′ alternative-splicing sites. In different cell of Arabidopsis root tissues, intron retention is more common than other types of alternative splicing (Li et al. 2016). However, it is interesting that in almost all type of tissues even in different aged condition, I found intron retention is the common phenomenon.

In the current study, I detected higher tissue-specific expression of *PPR2* and *PPR5* than of the other *PPR* genes. RNA editing in *ndhB-149*, *ndhB-1255*, and *ndhD-2* does not occur in roots or in lincomycin-treated seedlings (Tseng et al. 2013). In addition, *matK-640* is fully edited, and *accD-794*, *atpF-92*, *psbE-214*, *psbF-77*, *psbZ-50*, and *rps14-50* are completely or highly edited, in both green and non-green tissues (Tseng et al. 2013). Based on these findings and the current results, it is likely that alternative splicing of *PPR* genes affects tissue-specific RNA editing.

I also investigated the effects of alternative splicing on protein. PPR1c contains a 13 amino-acid sequence, CGFLKCYSDYITR (48–60). This amino-acid sequence is homologous to that of THCA synthase in *Cannabis sativa*. The major role of these residues is substrate binding rather than direct catalysis (Shoyama et al. 2012). Another homolog of this peptide is found in a juvenile hormone esterase-related protein in *Operophtera brumata*. This protein plays a major role in controlling growth and development (Kontogiannatos et al. 2013). In PPR1b, the addition of the amino-acid sequence DYITR (48–52). This peptide is homologous to non-ribosomal peptide synthases (NRPSs), which help produce natural products with antimicrobial and anticancer properties (Agrawal et al. 2017). The additional 26 nt sequence in PPR2b is

homologous to that in *Apis cerana* endothelin-converting enzyme 1 (ECE-1), as revealed by BLAST searches (https://www.ncbi.nlm.nih.gov). ECE-1 and ECE-2 can both cleave amyloid-β in mouse brain (Palmer et al. 2013). In PPR5 two isoforms are predicted that produce full length protein according to TAIR database. Intron retention in between exon 7 and 8 adds 84 nucleotides. The shorter isoform (PPR5a) is expressed in all tissues whereas the longer isoform (PPR5b) expressed in lower level in some seedling, leaf, stem and root tissues (Figure 5 A and B). In a recent study Huang et al. reported that only the shorter isoform was detectable in whole plant (Col-0) but they could not detect the longer isoform which is contrary to the data reported in TAIR (Huang et al. 2018). This may be due to study on whole plant extracted RNA as the highly expressed shorter isoform may outweigh the low level expression of the longer isoform. Side by side it may be our technique is more sensitive enough to visualize the lower level expression. However, intron retention decreases the PPR motif from 11 to 10. Huang et al. reported that shorter isoform having 11 PPR motif is functional (Huang et al. 2018). Each PPR motif recognizes a specific nucleotide in RNA. Therefore it clearly indicates that RNA recognition must be change due to alternatively splicing in PPR5. PPR5a is responsible for embryo development, responsible for plastid rp12 and trnK intron splicing as well (Huang et al. 2018). Function of the longer isoform need to be investigated whether it is related to plant immunity or stress tolerance and what is the impact of PPR motif difference to its cognate RNA, these answers will be important finding.

In retained introns, I found only TTT elements, but in *PPR4*, I found both ATGTTT and TTT. I identified a new 3′ and 5′ alternative-splicing site in this gene, suggesting that ATGTTT is likely involved in alternative splicing of the 3′ and 5′ alternative-splicing

site. Indeed, these two sequences are involved in alternative splicing in the APP gene in human (Shibata et al. 1996). I found two genes containing alternative transcriptional regulatory motifs. A transcriptional regulatory motif is present in the 3′ UTR of *PPR4*; this motif may inhibit the translation or increase the degradation of this transcript (Boeva 2016). Alternative splicing in 3′ UTRs or any mutation in these sites may alter mRNA recognition sites, thereby greatly affecting protein expression (Chin et al. 2008; Ramsingh et al. 2010).

In this study, I found that alternative splicing in *PPR* family genes are highly diversified due to alternative splicing. Protein-protein interaction pathways can also be altered due to tissue-specific alternative splicing (Ellis et al. 2012).

Therefore, it is obvious that due to alterative splicing our studied PPR proteins might be greatly affected with altered even opposite function as these proteins bind to its target RNA with an algorithmic manner (Okuda et al. 2006; Nakamura et al. 2012). Indeed, in the nuclear transcripts of AT1G29930.1 and AT1G52400.1 from Arabidopsis, C-to-U and U-to-C RNA editing occurs in the translation borders (Meng et al. 2010). These deamination and amination reactions occur in highly adjacent sites, suggesting that the deamination reaction serves as the donor of amino groups for the amination reaction (Meng et al. 2010), although the amination frequency is higher. Another factor might also function as an amino-group donor.

Taken together, these findings indicate that alternative-splicing events in PPR transcripts have strong effects on plant physiology in Arabidopsis. More investigations of the localizations of alternatively spliced transcripts and proteins, as well as site-specific and tissue-specific RNA editing, are needed to further understand their effects on growth and development. PPRs are modular proteins that are highly reprogrammable.

Therefore, it would be interesting to investigate how these types of natural editing events in transcripts affect substrate recognition and plant physiology. Additional experiments are needed to understand the precise effects of alternative-splicing and editing events on the highly programmable PPR protein family.

2.5 References

Agrawal S, Acharya D, Adholeya A, et al (2017) Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential. Front Pharmacol 8:1–26

Aguilar-Moreno M (2007) Handbook of life in the Aztec world. Oxford: Oxford University Press

Andrés-Colás N, Zhu Q, Takenaka M, et al (2017) Multiple PPR protein interactions are involved in the RNA editing system in Arabidopsis mitochondria and plastids. Proc Natl Acad Sci 114: 8883–8888

Barkan A, Small I (2014) Pentatricopeptide repeat proteins in plants. Annu Rev Plant Biol. 65:415-442

Bayer-Császár E, Haag S, Jörg A, et al (2017) The conserved domain in MORF proteins has distinct affinities to the PPR and E elements in PPR RNA editing factors. Biochim Biophys Acta - Gene Regul Mech 1860:813–828

Boeva V (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. Front Genet 7:1-15

Chang T, Huang H, Hsu JB, et al (2013) An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC Bioinformatics 14:1-8

Chateigner-Boutin AL, Colas Des Francs-Small C, Fujii S, et al (2013) The E domains of pentatricopeptide repeat proteins from different organelles are not functionally equivalent for RNA editing. Plant J. 74:935-945

Chin LJ, Ratner E, Leng S, et al (2008) A SNP in a let-7 microRNA complementary site in the KRAS 3′ untranslated region increases non-small cell lung cancer risk. Cancer Res 68:8535–8540

Choquet Y (2009) 5′ and 3′ ends of chloroplast transcripts can both be stabilised by protein “caps”: a new model for polycistronic RNA maturation. EMBO J 28:1989–1990

Colcombet J, Lopez-Obando M, Heurtevin L, et al (2013) Systematic study of subcellular localization of Arabidopsis PPR proteins confirms a massive targeting to organelles. RNA Biol.10:9, 1557–1575

Coquille S, Filipovska A, Chia T, et al (2014) An artificial PPR scaffold for programmable RNA recognition. Nat Commun 5:1–9

Cushing DA, Forsthoefel NR, Gestaut DR, Vernon DM (2005) Arabidopsis emb175 and other ppr knockout mutants reveal essential roles for pentatricopeptide repeat (PPR) proteins in plant embryogenesis. Planta 221:424–436

Delannoy E, Stanley WA, Bond CS, Small ID (2007) Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. Biochem Soc Trans 35:1643–1647

Doniwa Y, Ueda M, Ueta M, et al (2010) The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. Gene 454:39–46

Ellis JD, Barrios-Rodiles M, Çolak R, et al (2012) Tissue-Specific alternative splicing

remodels protein-protein interaction networks. Mol Cell 46:884–892

Graveley BR (2001) Alternative splicing: Increasing diversity in the proteomic world. Trends Genet 17:100–107

Greco M, Chiappetta A, Bruno L, Bitonti MB (2012) In Posidonia oceanica cadmium induces changes in DNA methylation and chromatin patterning. J Exp Bot 63:695–709

Guo L, Liu CM (2015) A single-nucleotide exon found in Arabidopsis. Sci Rep 5:1–5

Gutierrez-Marcos JF, Dal Pra M, Giulini A, et al (2007) empty pericarp4 encodes a mitochondrion-targeted pentatricopeptide repeat protein necessary for seed development and plant growth in maize. Plant Cell Online 19:196–210

Huang, W., Zhu, Y., Wu, W., Li, X., Zhang, D., Yin, P., & Huang, J. (2018). The Pentatricopeptide Repeat Protein SOT5/EMB2279 Is Required for Plastid *rpl2* and *trnK* Intron Splicing. Plant Physiology, 177(2), 684–697. https://doi.org/10.1104/pp.18.00406

James AB, Syed NH, Bordage S, et al (2012) Alternative Splicing Mediates Responses of the Arabidopsis Circadian Clock to Temperature Changes. Plant Cell 24:961–981

Kelemen O, Convertini P, Zhang Z, et al (2013) Function of alternative splicing. Gene 514:1–30

Kent WJ (2002) BLAT -The BLAST-Like alignment tool. Genome Res 12:656–664

Kontogiannatos D, Swevers L, Maenaka K, et al (2013) Functional characterization of a juvenile hormone esterase related gene in the moth sesamia nonagrioides through RNA interference. PLoS One. 8:1-15

Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of

alternative splicing. Curr Opin Struct Biol 14:273–282

Leu KC, Hsieh MH, Wang HJ, et al (2016) Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing. RNA Biol. 13:593–604

Li S, Yamada M, Han X, et al (2016) High-resolution expression map of the arabidopsis root reveals alternative splicing and lincRNA regulation. Dev Cell 39:508–522

Lurin C, Andrés C, Aubourg S, et al (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089–2103

Meinke D, Sweeney C, Muralla R (2009) Integrating the genetic and physical maps of *Arabidopsis thaliana*: Identification of mapped alleles of cloned essential (EMB) genes. PLoS One 4:1-11

Meng Y, Chen D, Jin Y, et al (2010) RNA editing of nuclear transcripts in *Arabidopsis thaliana.* BMC Genomics. 11:1-7

Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science (80) 338:1593–1599

Nakamura T, Yagi Y, Kobayashi K (2012) Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. Plant Cell Physiol 53:1171–1179

Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes Briefings Funct Genomics Proteomics 8:215–230

Okuda K, Chateigner-Boutin A-L, Nakamura T, et al (2009) Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. Plant Cell 21:146–156

Okuda K, Nakamura T, Sugita M, et al (2006) A pentatricopeptide repeat protein is a site recognition factor in chloroplast RNA editing. J Biol Chem 281:37661–37667

Palmer JC, Tayler HM, Love S (2013) Endothelin-converting enzyme-1 activity, endothelin-1 production, and free radical-dependent vasoconstriction in Alzheimer's disease. J Alzheimer's Dis 36:577–587

Pan Q, Shai O, Lee LJ, et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415

Prikryl J, Rojas M, Schuster G, Barkan A (2011) Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. Proc Natl Acad Sci 108:415–420

Ramsingh G, Koboldt DC, Trissal M, et al (2010) Complete characterization of the microRNAome in a patient with acute myeloid leukemia. Blood 116:5316–5327

Schallenberg-Rüdinger M, Lenz H, Polsakiewicz M, et al (2013) A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. RNA Biol 10:1549–1556

Shibata A, Hattori M, Suda H, Sakaki Y (1996) GENE Identification of cis-acting elements involved in an alternative splicing of the amyloid precursor protein (APP) gene. Gene 175:203–208

Shikanai T (2015) RNA editing in plants: Machinery and flexibility of site recognition. Biochim Biophys Acta 1847:779–785

Shoyama Y, Tamada T, Kurihara K, et al (2012) Structure and function of Δ1-tetrahydrocannabinolic acid (THCA) synthase, the enzyme controlling the psychoactivity of Cannabis sativa. J Mol Biol 423:96–105

Small ID, Peeters N (2000) The PPR motif – a TPR-related motif prevalent in plant

organellar proteins. Trends Biochem Sci. 25:46-47

Stamm S, Riethoven JM, Stamm S, et al (2006) ASD : A bioinformatics resource on alternative splicing ASD : a bioinformatics resource on alternative splicing 34: D46-D55

Stothard P (2000) The sequence manipulation suite: javascript programs for analyzing and formatting protein and DNA sequences. Bio Tech 28:1102–1104

Syed NH, Kalyna M, Marquez Y, et al (2012) Alternative splicing in plants - coming of age. Trends Plant Sci 17:616–623

Tseng C-C, Lee C-J, Chung Y-T, et al (2013) Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant J 82:375–392

Williams-Carrier R, Kroeger T, Barkan A (2008) Sequence-specific binding of a chloroplast pentatricopeptide repeat protein to its native group II intron ligand. Rna 14:1930–1941

Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13:59–69

Zhou W, Lu Q, Li Q, et al (2017) PPR-SMR protein SOT1 has RNA endonuclease activity. Proc Natl Acad Sci 114:E1554–E1563

Chapter III

Identification of Tissue-Specific Alternative Splicing of Zinc Finger (ZnF) Family Genes in *Arabidopsis thaliana*

3.1 Introduction

Alternative splicing is a bimolecular process that involves the generation of multiple mRNAs from a single gene sequence. This is done by differentially targeting splice sites. It is a vital process in the development of living organisms and is influenced by environmental factors (Graveley 2001; Lareau et al. 2004; Kelemen et al. 2013). Alternative splicing is a key occurrence in plants. About a decade and a half ago, it was documented that 1.2% of Arabidopsis genes exhibited instances of alternative splicing. However, with the advent of next-generation sequencing technologies, it came to light that up to 60% of genes that contained multiple exons exhibit alternative splicing (Syed et al. 2012). It has also been discovered that the primary genes involved in the circadian clock of Arabidopsis are heavily influenced by alternative splicing in varying environmental conditions (James et al. 2012; Syed et al. 2012). Prior studies on alternative splicing have been conducted using whole plant samples, but, in this study, we have chosen to take the tissue-specific approach which is to be observed at different stages of plant development. The ramifications on functionality as well as the phenotypic impact of alternative splicing are crucial and study into this is ongoing (Syed et al. 2012). In human cells, over 90% of transcribed genes undergo alternative splicing, which results in the production of an immense variety of diverse proteins (Pan et al. 2008).

In plants, the conversion of C-to-U as affected by cytidine deaminase has been shown to be dependent on zinc ions. This is possibly due to the requirement for Zn^{2+} in the active center of cytidine deaminase (Vasudevan et al. 2013). In numerous plant ancestries, the ubiquitous OZ (Zinc Finger) family proteins may be discerned, including in *Selaginella*, which may suggest that it plays an evolutionarily conserved role in the plant editosome (Sun et al. 2015; Sun et al. 2016). Recent advancements made in the field of plant RNA

editing have begun to focus on the non-PPR editing factors, which highlight an unpredicted diversity in the plant editosome (Sun et al. 2016). Amongst the most predominant motifs in eukaryotic proteins may be found the Zinc Finger (ZnF). They are characterized following the number as well as order of the histidine and cysteine residues present in the amino acid chain, as these interact with zinc ions to maintain the tertiary structure of the proteins. Analogous to the function at the molecular level, it was discovered that Zinc Finger Proteins (ZFPs) play a role in a wide variety of different biological processes including the regulation of transcription, modification of chromatin and degradation of proteins by ubiquitin (Appelhagen et al. 2010; Laity et al. 2001). The RanBP2 Zinc Finger family protein is a potential chloroplast RNA editing factor (Sun et al. 2015). The efficiency of editing RNAs has been seen to vary in different tissues and at different stages of development e.g. in non-green tissues and seedlings while a majority of editing sites were found in green seedlings and green tissues (Tseng et al. 2013). The greatest number of proteins in the OZ (Zinc Finger) family are essential in RNA editing and also play a role in other aspects of RNA metabolism in plant organelles (Sun et al. 2015). Correspondence has been identified amongst the Zinc Finger proteins based on their zinc finger motifs. The structures of the Zinc Finger proteins (ZnF) are defined by their evolutionarily conserved zinc finger motifs. A number of Zinc Finger proteins are used in multiplexed autoantibody assays for the minimally invasive detection of colorectal cancer (O'Reilly et al. 2015).

Tissue-specific alternative splicing may alter the stability of the protein, activity of the enzyme, localization at the sub-cellular level as well as other features in mammalian tissues (Merkin et al. 2012). Tissue-specific alternative splicing in plants, however, is as yet not well understood.

The aim of this study is to identify instances of alternative splicing of nominated *ZnF* gene sequences in various plant tissues including the leaf, stem, stipe and root as well as the whole seedling. The expression patterns of the alternatively spliced mRNAs shall be compared and investigated using different Arabidopsis tissues at varying developmental stages i.e. at 4, 8, 12, 16, 21, 27 and 32 days. Tissue-specific alternative splicing influences the resulting functional protein diversity in Arabidopsis shall also be investigated.

3.2 Materials and Methods

3.2.1 *Arabidopsis thaliana* sample collection and growth conditions

Paper pots were gently filled with a mixture of horticultural perlite, peat moss and vermiculite in a compositional ratio of 1:2:1. The Arabidopsis seeds were then sowed in the paper pots, and the pots covered using cling film, before being kept under dark conditions for three to four days. Afterwards, they were transferred to the Green Farm U.ING (made in Japan) kept in a growth room, under the following conditions: fixed growth room temperature 22°C, relative humidity 45%, 16 hours' exposure to light and 8 hours in darkness. After germination, watering was carried out twice daily (every morning and evening), fertilization two times per week, and different kinds of samples collected. The following samples were collected: seedling (whole plants at 4, 8 and 12 days) and 16, 21, 27, 32 days old leaf, stipe, stem and root.

3.2.2 Plant RNA extraction and cDNA synthesis

I have extracted RNA from those sample by using Qiagen plant mini kit (Germany) according to the manufacturer's instructions. The extracted RNA was then treated with

DNase (RQ1 RNase free DNase; Promega, Madison, WI, USA) to digest the contaminating genomic DNA. The samples were then purified by phenol-chloroform and ethanol precipitation. The NanoDrop™ (Thermo Scientific) was finally used to quantify the RNA obtained. Using this purified RNA, cDNA was then synthesized by reverse transcriptase (Superscript III, Invitrogen) with oligo dT primer. cDNA synthesis was confirmed using a housekeeping primer for *Arabidopsis thaliana*; GAPDH forward primer: GTTGTCATCTCTGCCCCAAG, reverse primer: TGCAACTAGCGTTGGAAACA.

3.2.3 Selection of prospective alternatively spliced *ZnF* gene sequences from the Arabidopsis database

Candidate genes for probable alternative splicing were identified by combing through the NCBI gene database (https://www.ncbi.nlm.nih.gov/gene/). The accession numbers pertaining to each *ZnF* gene sequence identified were sourced from the NCBI database and input into the *Arabidopsis thaliana* Plant Genome Database (http://www.plantgdb.org/AtGDB/) with the aim of establishing the genomic map of each gene. Using this information, probable alternatively spliced genes were selected (whole genome model) by At- TAIR10. The sequences expressed in different tissues obtained from The Arabidopsis Information Resource (TAIR) https://www.arabidopsis.org/ enabled the identification of 25 genes that were alternatively spliced. Information on the full-length genomic DNA, mRNA and cDNA was obtained as well. Ultimately, the genomic sequence of each gene, as well as information relating to intron and exon sequences, coding sequences, transcript and resulting amino acid sequence were obtained

from The Arabidopsis Genome Integrative Explorer (http://atgenie.org/) and cross-checked.

Table 1: Accession numbers of candidate ZnF family genes

Tentative gene Name	Gene ID/ Accession no.	Annotation	AT°C	Cycle	Forward	Reverse
ZnF1	AT2G47850	Zinc finger C-x8-C-x5-C-x3-H type family protein	58	30	CATGAGCCAC GTTCCACTAA	GGCGCCTTGC ATATAAGAAC
ZnF2	AT1G75340	Zinc finger C-x8-C-x5-C-x3-H type family protein	55	35	CCCTTCTCCAG CATTCTCTG	GTTGACGCTG AAAGCATTTG
ZnF3	AT3G61850, DAG1	Dof-type zinc finger DNA-binding family protein	57	30	CCAGTCGGAG GTAGCTCAAG	GGCATTGTTG GAAACCCTAA
ZnF4	AT3G51950	Zinc finger CCCH-type family protein	55	35	CAACCTGTCG TGTTGGGATA	TGTGTGGTCC ATCATCACAT
ZnF5	AT1G06040, STO	B-box zinc finger family protein	58	35	CAGCAGCAAC AACAACCTTC	CACCGAAGAA TCCCATGTCT
ZnF6	AT1G29800	RING/FYVE/PHD-type zinc finger family protein	57	30	TCGCTTGAAG GAAGCATTTT	CCAACGAGAT GAAGCGAATC
ZnF7	AT4G06634	Zinc finger (C2H2 type) family protein	55	35	CTTTCGAGAG ACGACCCATC	TTCCCGCAAC CATCATAACT

Techniques like PCR and Sequencing were used to confirm genes. The following table indicates annealing temperature, number of thermal cycles and primer sequences.

3.2.4 Design of Primers

Primers were designed using the Primer3 primer design software（*primer3 software bioinfo.ut.ee/primer3-0.4.0/primer3/*）and primer blast was performed by NCBI primer blast. In case of failure or inappropriate outcome with the first designed primer set, another set of primers were designed. Primers were purchased from (Eurofins, Japan); stored in TE buffer at a concentration of 50 pmol/µl in salt-free condition. The primer set was diluted to 10 pmol/µl with TE buffer as working concentration.

3.2.5 Polymerase Chain Reaction and analysis of polyacrylamide gel image

The number of thermal cycles used was 30-35. The denaturation temperature was set at 94°C and elongation temperature of 72°C. The annealing temperatures were set according to each primer as suggested by the primer 3 software after primer design (Table 1). A 6% polyacrylamide gel was used for electrophoresis which was stained with the SYBR® Green dye (Lonza, Rockland, ME, USA). Equal amounts of PCR product were loaded onto the polyacrylamide gel and electrophoresis carried out. Each gel image was photographed using different exposure times in order to obtain high-quality images for analysis. Gel images were analyzed using the LAS3000 (Fujifilm, Tokyo, Japan).

3.2.6 Sequencing of the transcript

PCR products were sequenced in Applied Biosystems (Applied Biosystems 3130xl Genetic Analyzer, Foster City, CA, USA). The desired bands from PAGE were excised from the gel and kept in disposable pellet pestle/Tissue grinder tubes (Kimble®, Capitol Scientific, Inc. TX 78758, USA) before being frozen at -80°C for 1 hour. After that, the frozen gel pieces were well crushed using the pestle. About 10 µl 0.1X TE was added to the crushed powder before being subjected to further grinding. The pestle was then discarded and the tubes vortexed for 10 minutes. Following this, the tubes were centrifuged at full speed at 4°C for 20 minutes in a tabletop centrifuge. The supernatant was transferred into another tube and 3 µl was used as sequencing sample. The Big dye Terminator V3.1 sequencing standard kit (Austin, TX, USA catalogue no. 4336935) was used for sequencing. Sequence results with the reverse primer were reverse complemented using the Sequence Manipulation Suite online software, http://www.bioinformatics.org/sms2/reference.html (Stothard 2000). The entire set of

sequencing results were aligned with those of the *Arabidopsis thaliana* genome by way of BLAST (Kent 2002).

3.2.7 Determination of tissue-specific expression of the isoforms of different *ZnF* genes

The products obtained from PCR were loaded onto a 6% polyacrylamide gel and run in the 1XTBE buffer at 200 volts for 20 minutes. SYBR® Green dye (Lonza, Rockland, ME, USA) was then used to stain the gel in the presence of 1XTBE buffer and placed on a rocker for 20 minutes. After this, the gel was placed in the LAS3000 (Fujifilm, Tokyo, Japan) and its image captured. Finally, the image obtained was analysed using the Image J (NIH, Maryland, USA) software whereby densitometry analysis was done on each of the resulting bands in order to compare the expression levels of the transcripts that exhibit alternative splicing. These results were then confirmed by sequencing.

3.3 Results

In this study, I identified 25 alternatively spliced genes of ZnF family proteins from Arabidopsis database TAIR (https://www.arabidopsis.org/) that were differentially expressed in tissues and development specific aspect. From these genes, I have selected seven genes that were located in coding region (6) and UTR (1) due to their functional importance for further testing by semi-quantitative PCR and Sanger sequencing. I have represented the alternatively spliced isoforms of a gene in alphabetic order a, b, c, d.

3.3.1 Determinations of tissue-specific expression of the isoforms of different *ZnF* genes

In *ZnF1*, 3' alternative spliced site, 78 nt added with exon 5 creating *ZnF1b* (Figure 1A and C). This 3' alternative spliced site was selected in seedling 4 days, leaf 16, 21, 27 days and stipe 32 days (Figure 1A). In root *ZnF1a*, expression gradually increases but *ZnF1b* expression gradually decrease, 32 days of expression again increase (Figure 1B).

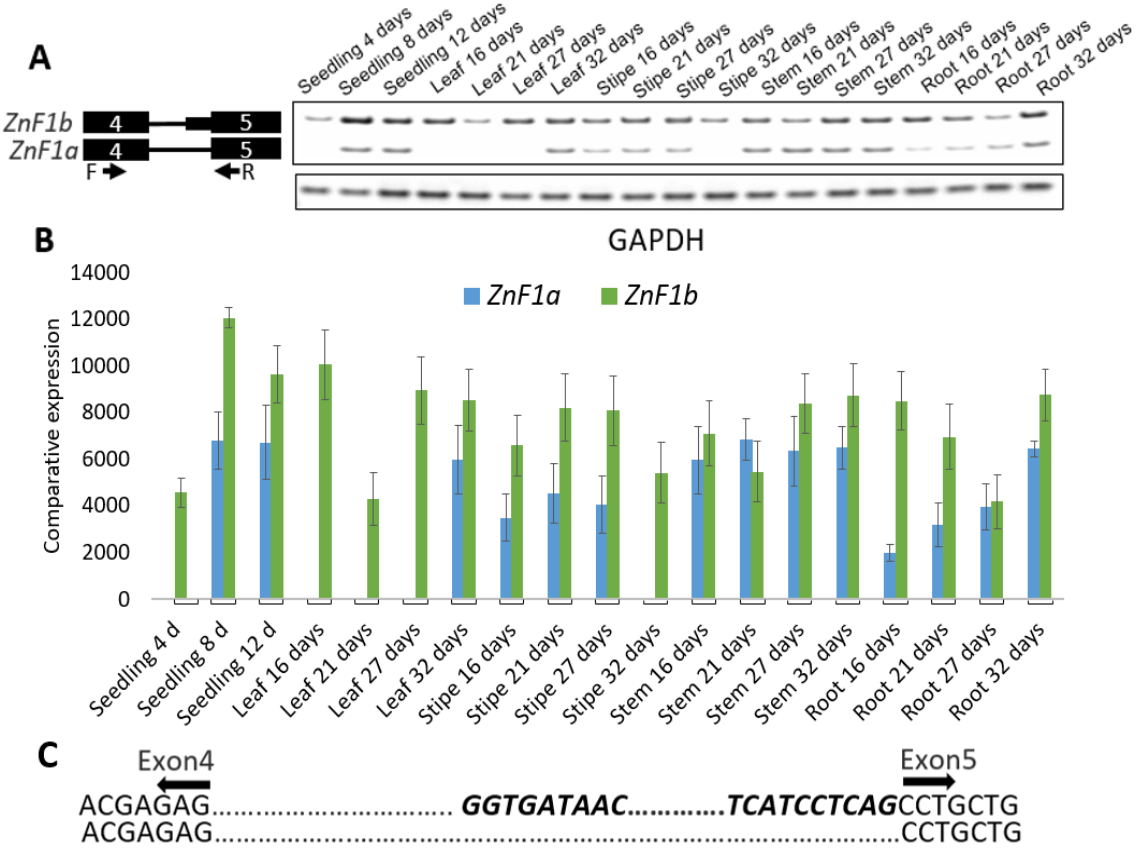


Figure 1. *ZnF1* accession no. AT2G47850 **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in two spliced isoforms. **C.** The genomic sequence of *ZnF1* from exon 4 to exon 5. Arrow indicates exon 4 and exon 5 boundaries, dot indicates intron sequences. Bold and italic sequences and dot indicate 3' alternate splice site.

In *ZnF2*, 78 and 93 nt containing exon 7 and 8 respectively skipped and creating *ZnF2a* (Figure 2A and C). I selected two new unannotated spliced isoforms, one 100 nt intron retained between exon 7 and 8 creating *ZnF2c* and another one 106 nt intron retained between exon 8 and 9 creating *ZnF2d* were confirmed by Sanger sequencing (Figure 2A and C). Unannotated spliced product intron retention *ZnF2c* was activated in leaf 32 days and stipe 16 days and *ZnF2d* was activated in stem 16, 27 days and root 16 days (Figure 2A). In *ZnF2b* expressed well in all tissues (Figure 2B). Expression of *ZnF2d* in stem 16 and 27 days gradually decreases whereas in 21 and 32 days did not express (Figure 2B).

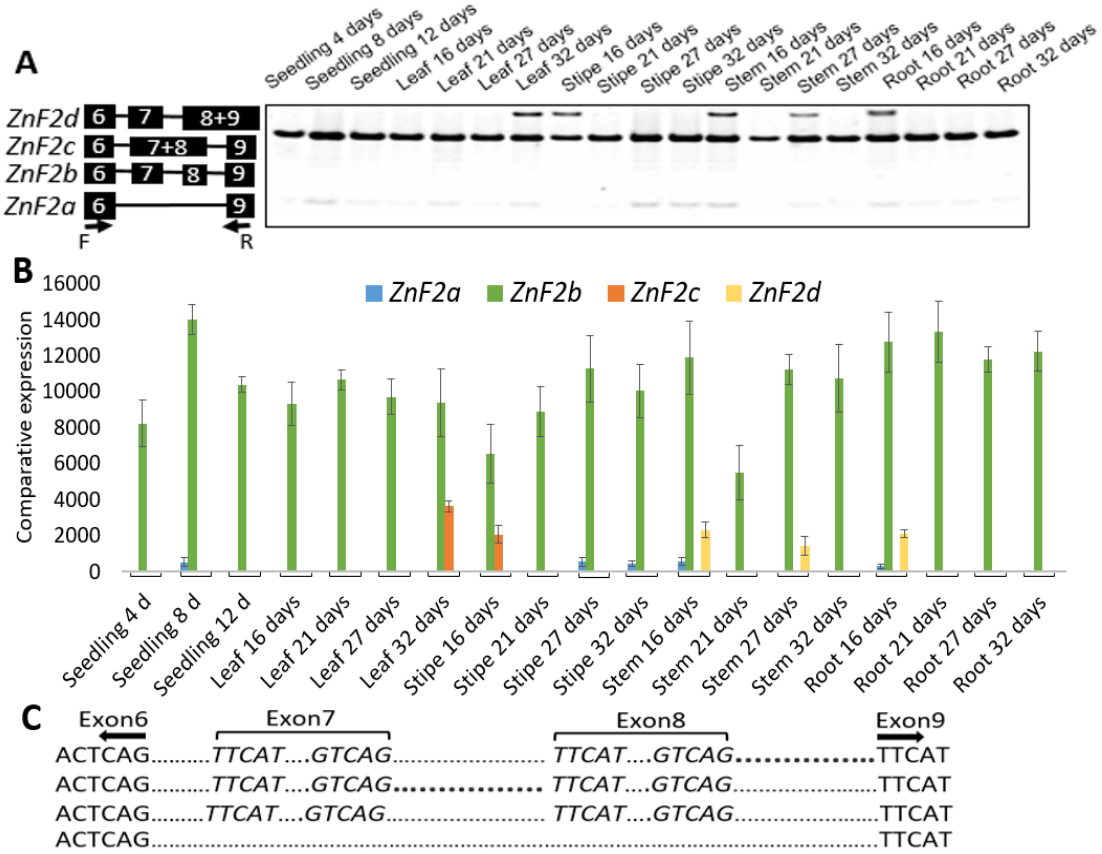


Figure 2. *ZnF2* accession no.AT1G75340. **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in four spliced isoforms. **C.** The genomic sequence of *ZnF2* from exon 6 to exon 9. Arrow indicates exon 6 and exon 9 boundaries whereas dot indicates intron sequences. Italic sequences and dot indicate exon 7 and exon 8. Bold dot indicate intron retained.

Regarding intron retention, in *ZnF3*, the 84 nt intron was retained within exon 2 and 3 creating *ZnF3b* (Figure 3A and C). Intron retention was not activated in seedling 4 days, leaf 16,21,27,32 days, stipe 21 days and stem 21 days (Figure 3A). In leaf *ZnF3a* expression gradually decreased in 16, 21, 27, 32 days (Figure 3B). In seedling 4 days *ZnF3b* did not express but in 8 and 12 days expression gradually decreased, *ZnF3a* expression gradually increased in 4,8,12 days (Figure 3B).

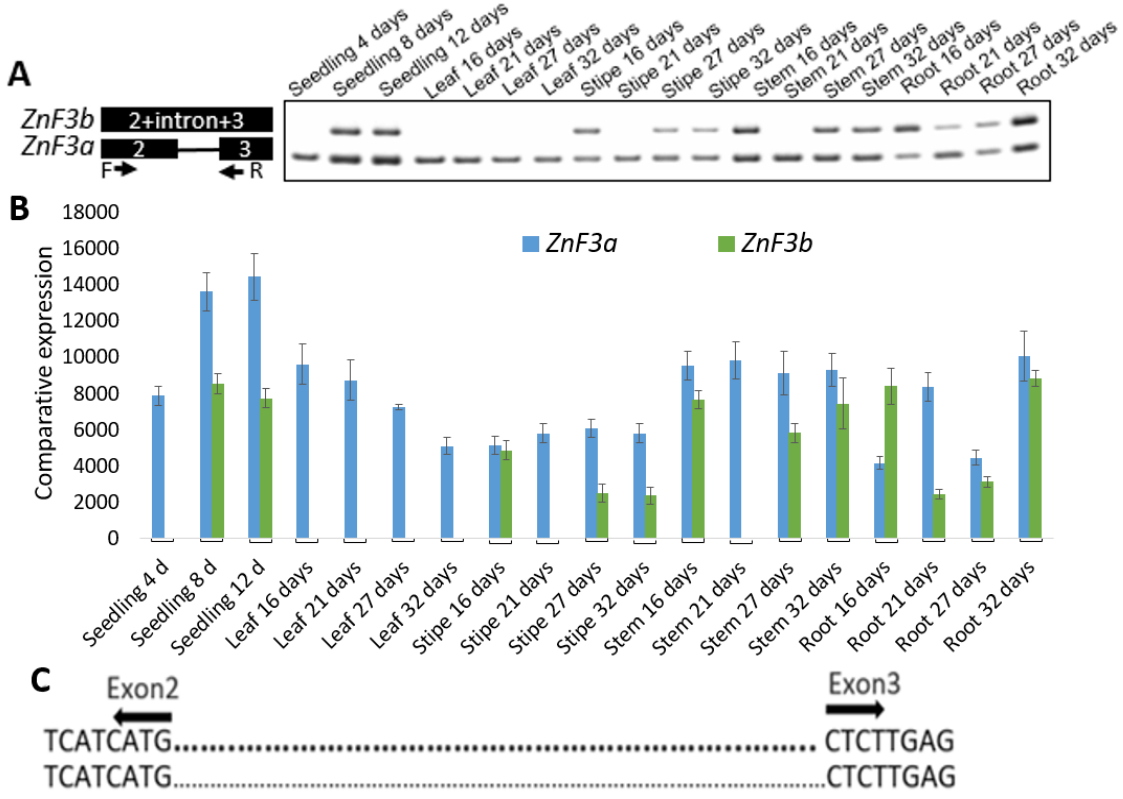


Figure 3. *ZnF3* accession no. AT3G61850, DAG1. **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in spliced isoforms. **C.** The genomic sequence of *ZnF3* from exon 2 to exon 3. Arrow indicates exon 2 and exon 3 boundaries, dot indicates intron sequences. Bold dot indicate intron retained.

In *ZnF4*, the 88 nt intron was retained between exon 9 and 10 producing the amplified product of *ZnF4c*, new unannotated spliced isoforms 3' alternative spliced site, 32 nt added with exon 10 creating *ZnF4b* (Figure 4A and C). This 3' alternative spliced site was

selected in seedling 4 days according to our designed PCR primers (Figure 4A). Intron retention was not activated in seedling 12 days and root 16 days (Figure 4A). In root *ZnF4c* in 16 days was not expressed but 21, 27 and 32 days expression gradually increased (Figure 4B).

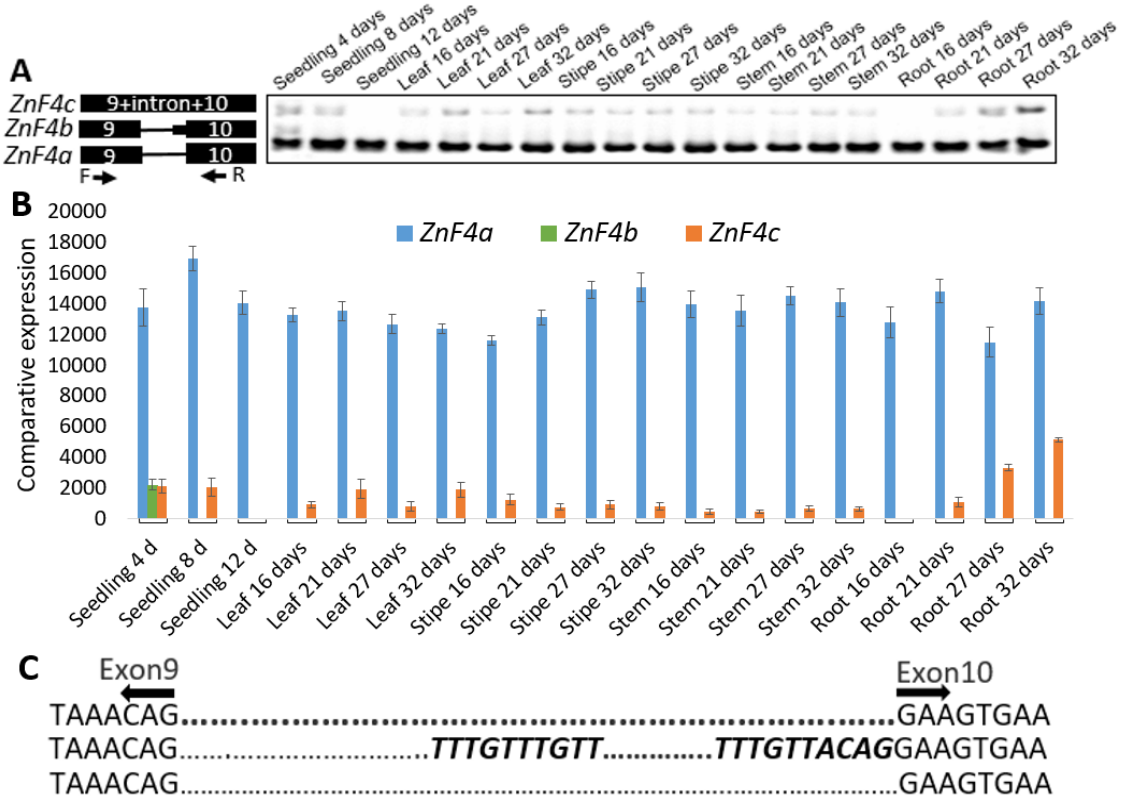


Figure 4. *ZnF4*/AT3G51950 **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression in different tissues of two spliced isoforms. **C.** The genomic sequence of *ZnF4* from exon 9 to exon 10. Arrow indicates exon 9 and exon 10 boundaries, dot indicates intron sequences. Bold and italic sequences and dot indicate 3' alternate splice site. Bold dot indicate intron retained.

In *ZnF5*, between exon 2 and 3, 82 nt intron was retained creating *ZnF5b*, transcript according to the mentioned PCR primer (Figure 5A and C). Intron retention was not activated in seedling 4 days, stipe 16, 27, 32 days, stem 21,27,32 days and root 27, 32 days (Figure 5A).

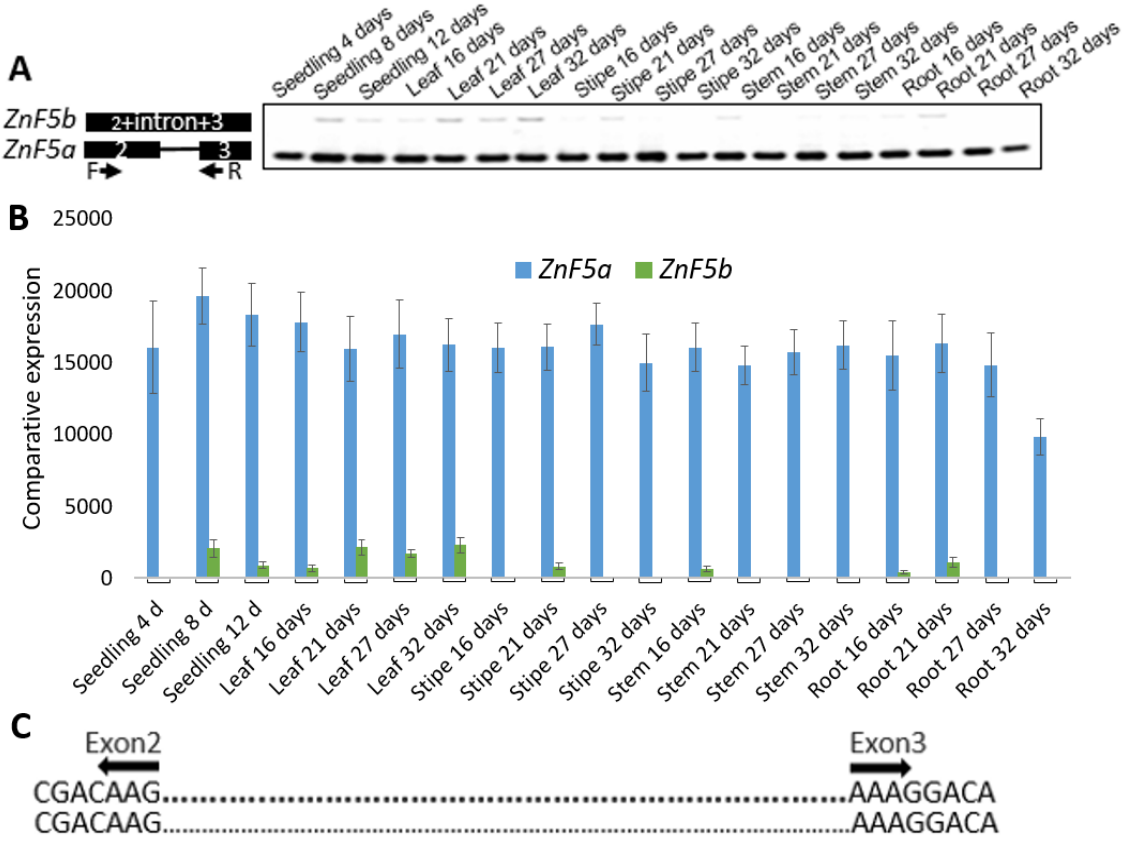


Figure 5. *ZnF5*/AT1G06040, STO. **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in two spliced isoforms. **C.** The genomic sequence of *ZnF5* from exon 2 to exon 3. Arrow indicates exon 2 and exon 3 boundaries, dot indicates intron sequences. Bold dot indicate intron retained.

In *ZnF6*, 89 nt intron retained between exon 5 and 6, creating *ZnF6b* (Figure 6A and C). Intron retention was activated in root 27 days (Figure 6A). In leaf *ZnF6a*, 16, 21, 27 and 32 days expression gradually decreased (Figure 6B).

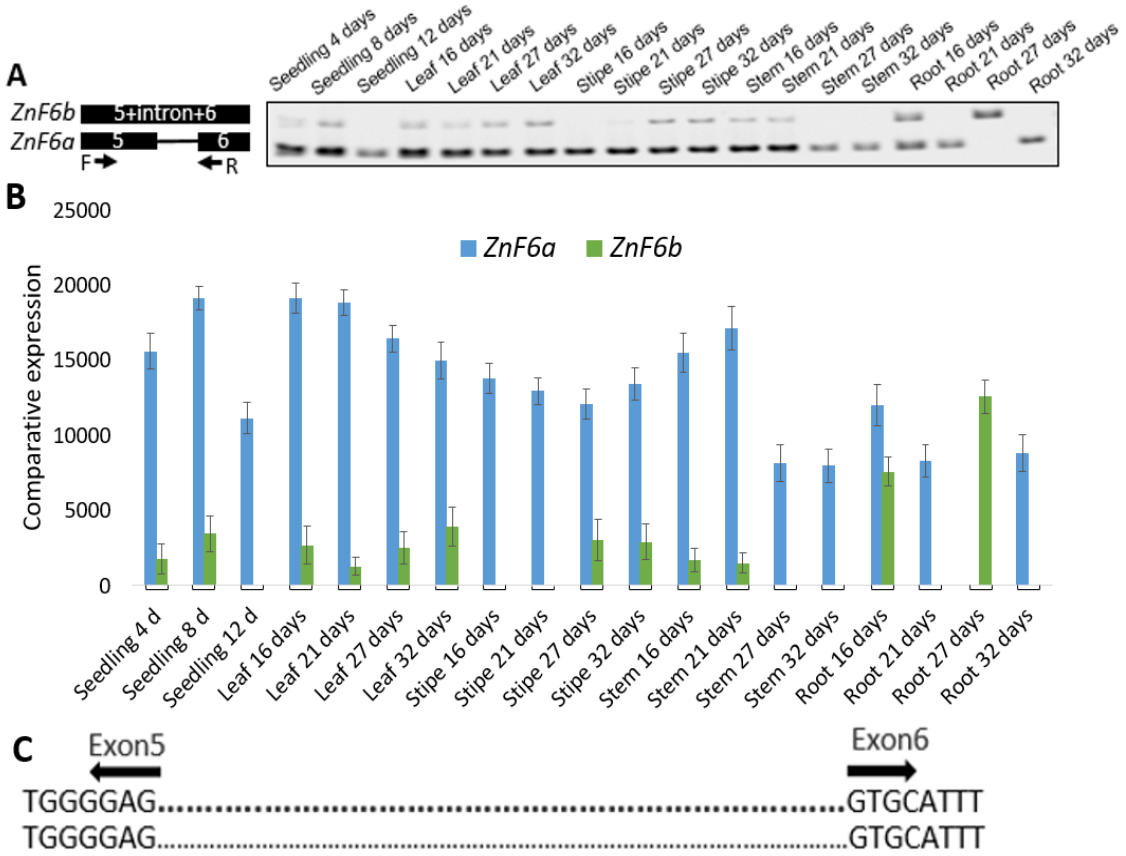


Figure 6. *ZnF6* accession no. AT1G29800. **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in two spliced isoforms. **C.** The genomic sequence of *ZnF6* from exon 5 to exon 6. Arrow indicates exon 5 and exon 6 boundaries, dot indicates intron sequences. Bold dot indicate intron retained.

In *ZnF7*, 110 nt intron retained between exon 2 and 3, creating *ZnF7c* (Figure 7A and C). 5' alternative spliced site, 33 nt added with exon 2 creating *ZnF7b*. This 5' alternative spliced site was selected in all tissues (Figure 7A).

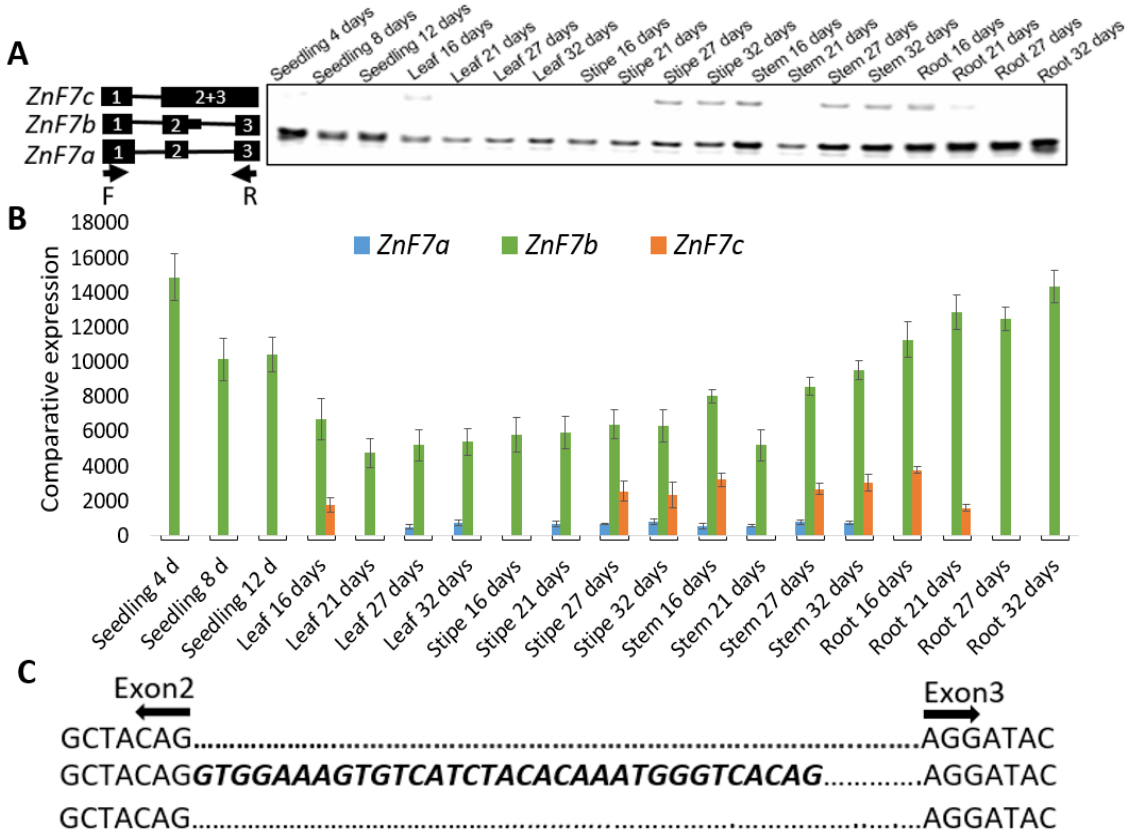


Figure 7. *ZnF7* accession no. AT4G06634. **A.** PCR amplification of seedling, leaf, stipe, stem and root. Estimated fragment size from gel in base pairs (bp). **B.** Comparative expression of different tissues in two spliced isoforms. **C.** The genomic sequence of *ZnF7* from exon 2 to exon 3. Arrow indicates exon 2 and exon 3 boundaries, dot indicates intron sequences. Sequences in bold and italics indicate a 5′ alternative-splicing site. Bold dot indicate intron retained.

3.3.2 Determination of protein affected by alternative splicing

Alternative splicing events located in the coding region of ZnF1, ZnF2, ZnF3, ZnF5, ZnF6 and ZnF7.

Table 2: Summary of alternative splicing affecting protein of studied genes

Tentative gene Name	Alternative splicing events	Effect on	
		Nucleotide	Amino acids
ZnF1a	Reference isoform Shorter	-	-
ZnF1b	Reference isoform 3' alternative spliced site	78 nucleotide addition	26 amino acid addition
ZnF2a	Reference isoform exon skipping	-	-
ZnF2b	Reference isoform longer	171 nucleotide addition	57 amino acid addition
ZnF2c	New unannotated Intron retention	100 nucleotide addition	Frameshift
ZnF2d	New unannotated Intron retention	106 nucleotide addition	Frameshift
ZnF3a	Reference isoform Shorter	-	-
ZnF3b	Reference isoform Intron retention	84 nucleotide addition	28 amino acid addition
ZnF5a	Reference isoform Shorter	-	-
ZnF5b	Reference isoform Intron retention	82 nucleotide addition	Frameshift
ZnF6a	Reference isoform Shorter	-	-
ZnF6b	Reference isoform Intron retention	89 nucleotide addition	Frameshift
ZnF7a	Reference isoform Shorter	-	-
ZnF7b	Reference isoform 5' alternative spliced site	33 nucleotide addition	11 amino acid addition
ZnF7c	Reference isoform Intron retention	110 nucleotide addition	Frameshift

The table represents that alternative splicing affected protein our selected genes of ZnF1, ZnF2, ZnF3, ZnF5, ZnF6 and ZnF7. The alternative splicing sites in case of ZnF4 were selected in UTR regions.

In ZnF1 there was an addition of 26 amino acids with reference isoform ZnF1b (Table 2) which was confirmed by PCR and sequencing. Those 26 amino acid contains Zinc finger C-x8-C-x5-C-x3-H type domain interval 4-26 and E-value 4.44e-05 (Figure 8).

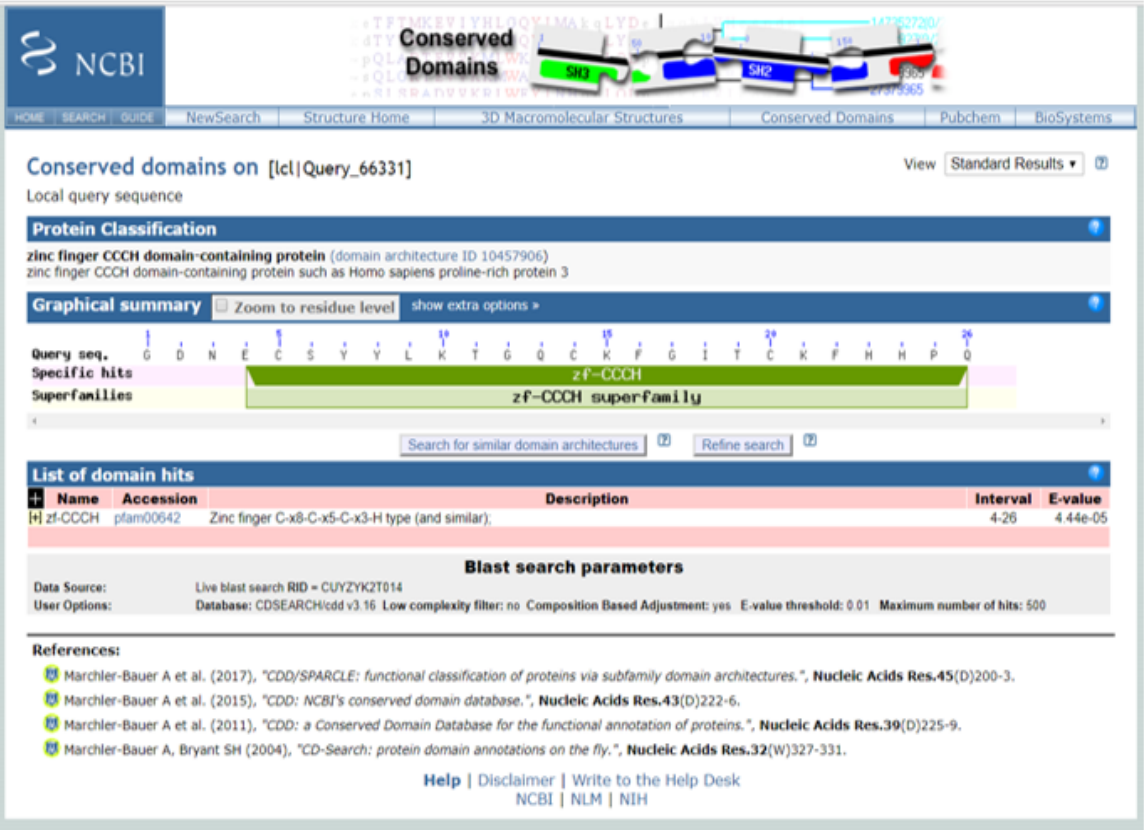


Figure 8. Conserved domain of ZnF1 accession no. AT2G47850. 26 amino acid (interval 4-26) contains Zinc finger C-x8-C-x5-C-x3-H type domain.

In ZnF2, reference isoform is ZnF2a, ZnF2b and another two new unannotated isoform ZnF2c, ZnF2d were identified (Table 2) due to the selection of new intron retention 100 nucleotides, 106 nucleotides respectively (Figure 2A and C). In ZnF2, due to 57 amino acids; QQTTFPNTNAGGVSSSGPPNPFASFTQQSNNQQTAFSNTNAGGLSSSGPPN AFASFN (253-309) addition with ZnF2a. In ZnF2c and ZnF2d additional 100 nucleotides and 106 nucleotides respectively resulted in the frameshift.

ZnF2a amino-acid sequences:

Met R K E L C R N F Q R G S L G F V F Y I L G F N L K S I D V R R C C R Y G E N C R F L H P Q Q A K P N N P F G F G T Q N Q Q Q Q Q Q Q Q Q Q N S S N P F G F G V Q S G G S S R P N Q F Q N T W S R T A S T P T G G G A A A S T Q Q T G K Q T Q P A D H K C T D P A A C K R V **Met** Q D D F K N E R P **Met** W K L T C Y G H W K Y F P C D V T G D I S Y E E L R A V A Y E E A K R G I P L Q S I V E R E R N L Q N S K I A E F E N F L R N P Y K G S V T A N Q S P F A A T T P S I F P Q S S Q I N S P S P A F S G F N Q Q T A F S N T N A G G L S S S G P P N A F A S F N K Q P N A F S V N T P Q P V P S G P S G F Q T N P S T T F K P A S F G P G P G F A T T P Q N N N I F G Q S T P T P A T N T S Q N N Q T A F N F N V P V A S F T A P A I N T T N T S S G T E L Q I G G D P V D S S I W L K E K W N P G E I P E Q A P P D A F V**Stop**

ZnF2a nucleotide and amino-acid sequences:

```
atgaggaaggaactgtgtagaaactttcagcgtggcagtttaggtttcgtgttctatatt
 M  R  K  E  L  C  R  N  F  Q  R  G  S  L  G  F  V  F  Y  I
ctggggtttaatttgaaatcaatcgatgtgcgaaggtgttgtaggtatggggaaaactgt
 L  G  F  N  L  K  S  I  D  V  R  R  C  C  R  Y  G  E  N  C
agatttcttcatccacaacaagctaagccaaataatccctttggctttggtacacaaaac
 R  F  L  H  P  Q  Q  A  K  P  N  N  P  F  G  F  G  T  Q  N
caacaacaacaacagcagcagcagcaacaaaatagttcgaaccctttggatttggtgta
 Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  N  S  S  N  P  F  G  F  G  V
caaagtggtggttccagcagaccgaatcagttccagaatacttggtcacgaacggcttct
 Q  S  G  G  S  S  R  P  N  Q  F  Q  N  T  W  S  R  T  A  S
acacctactggtggtggtgctgctgcttctacacagcagactggtaaacagacgcagccg
 T  P  T  G  G  G  A  A  A  S  T  Q  Q  T  G  K  Q  T  Q  P
gcagatcataaatgcacagatcctgctgcgtgtaagcgggtaatgcaagacgattttaag
 A  D  H  K  C  T  D  P  A  A  C  K  R  V  M  Q  D  D  F  K
aatgagagacccatgtggaagctcacatgctacggccactggaaatattttccatgtgat
 N  E  R  P  M  W  K  L  T  C  Y  G  H  W  K  Y  F  P  C  D
gttaccggtgatatcagctatgaagaactacgtgcagtggcatatgaagaagctaaacga
 V  T  G  D  I  S  Y  E  E  L  R  A  V  A  Y  E  E  A  K  R
ggaatacctctacagtcgattgttgaaagggagaggaatctgcaaaattccaaaatagcc
 G  I  P  L  Q  S  I  V  E  R  E  R  N  L  Q  N  S  K  I  A
gagtttgaaaactttctacggaatccatacaaaggctctgttaccgccaatcaaagcccg
 E  F  E  N  F  L  R  N  P  Y  K  G  S  V  T  A  N  Q  S  P
tttgctgcaaccactcctagtatcttccctcagtccagtcagatcaattccccttctcca
 F  A  A  T  T  P  S  I  F  P  Q  S  S  Q  I  N  S  P  S  P
gcattctctggctttaatcaacaaactgcattctccaataccaatgcgggtggactcagt
 A  F  S  G  F  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  S
tcatctggacctccgaatgcctttgcaagctttaataaacaaccaaatgctttcagcgtc
 S  S  G  P  P  N  A  F  A  S  F  N  K  Q  P  N  A  F  S  V
aacactcctcaacctgttccttcaggtccctctggtttccaaaccaatccatcaacaaca
 N  T  P  Q  P  V  P  S  G  P  S  G  F  Q  T  N  P  S  T  T
ttcaaaccagcatcatttggacctggacccggatttgccacaactccacaaaacaataac
 F  K  P  A  S  F  G  P  G  P  G  F  A  T  T  P  Q  N  N  N
atctttggtcaatcaactccaacaccggctacaaacacttcccagaacaatcagaccgca
```

```
 I  F  G  Q  S  T  P  T  P  A  T  N  T  S  Q  N  N  Q  T  A
ttcaacttcaatgtccctgttgcatcttttactgctcctgctataaacacgacaaataca
 F  N  F  N  V  P  V  A  S  F  T  A  P  A  I  N  T  T  N  T
tcttccggaaccgagctgcaaataggtggtgatcctgttgatagtagtatctggctaaag
 S  S  G  T  E  L  Q  I  G  G  D  P  V  D  S  S  I  W  L  K
gagaaatggaatccaggggagattccggaacaagcgccgcctgatgcctttgtttaa
 E  K  W  N  P  G  E  I  P  E  Q  A  P  P  D  A  F  V  -
```

ZnF2b amino-acid sequences:

Met R K E L C R N F Q R G S C R Y G E N C R F L H P Q Q A K P N N P F G F G T Q N Q Q Q Q Q Q Q Q Q Q N S S N P F G F G V Q S G G S S R P N Q F Q N T W S R T A S T P T G G G A A A S T Q Q T G K Q T Q P A D H K C T D P A A C K R V **Met** Q D D F K N E R P **Met** W K L T C Y G H W K Y F P C D V T G D I S Y E E L R A V A Y E E A K R G I P L Q S I V E R E R N L Q N S K I A E F E N F L R N P Y K G S V T A N Q S P F A A T T P S I F P Q S S Q I N S P S P A F S G F N Q Q T A F S N T N A G G L S S S G P P N A F A S F N Q Q T T F P N T N A G G V S S S G P P N P F A S F T Q Q S N N Q Q T A F S N T N A G G L S S S G P P N A F A S F N K Q P N A F S V N T P Q P V P S G P S G F Q T N P S T T F K P A S F G P G P G F A T T P Q N N N I F G Q S T P T P A T N T S Q N N Q T A F N F N V P V D I F Y C S C Y K H D K Y I F R N R A A N R W **Stop** S C **Stop Stop Stop** Y L A K G E **Met** E S R G D S G T S A A **Stop** C L C L

ZnF2b nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atgaggaaggaactgtgtagaaactttcagcgtggcagttgtaggtatggggaaaactgt
 M  R  K  E  L  C  R  N  F  Q  R  G  S  C  R  Y  G  E  N  C
agatttcttcatccacaacaagctaagccaaataatccctttggctttggtacacaaaac
 R  F  L  H  P  Q  Q  A  K  P  N  N  P  F  G  F  G  T  Q  N
caacaacaacaacagcagcagcagcaacaaaatagttcgaaccctttggatttggtgta
 Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  N  S  S  N  P  F  G  F  G  V
caaagtggtggttccagcagaccgaatcagttccagaatacttggtcacgaacggcttct
 Q  S  G  G  S  S  R  P  N  Q  F  Q  N  T  W  S  R  T  A  S
acacctactggtggtggtgctgctgcttctacacagcagactggtaaacagacgcagccg
 T  P  T  G  G  G  A  A  A  S  T  Q  Q  T  G  K  Q  T  Q  P
gcagatcataaatgcacagatcctgctgcgtgtaagcgggtaatgcaagacgattttaag
 A  D  H  K  C  T  D  P  A  A  C  K  R  V  M  Q  D  D  F  K
aatgagagacccatgtggaagctcacatgctacggccactggaaatattttccatgtgat
 N  E  R  P  M  W  K  L  T  C  Y  G  H  W  K  Y  F  P  C  D
gttaccggtgatatcagctatgaagaactacgtgcagtggcatatgaagaagctaaacga
 V  T  G  D  I  S  Y  E  E  L  R  A  V  A  Y  E  E  A  K  R
ggaatacctctacagtcgattgttgaaagggagaggaatctgcaaaattccaaaatagcc
 G  I  P  L  Q  S  I  V  E  R  E  R  N  L  Q  N  S  K  I  A
gagtttgaaaactttctacggaatccatacaaaggctctgttaccgccaatcaaagcccg
 E  F  E  N  F  L  R  N  P  Y  K  G  S  V  T  A  N  Q  S  P
tttgctgcaaccactcctagtatcttccctcagtccagtcagatcaattccccttctcca
 F  A  A  T  T  P  S  I  F  P  Q  S  S  Q  I  N  S  P  S  P
gcattctctggctttaatcaacaaactgcattctccaataccaatgcgggtggactcagt
 A  F  S  G  F  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  S
```

```
tcatctggacctccgaatgcctttgcaagcttcaatcaacaaactacattccccaatacc
 S  S  G  P  P  N  A  F  A  S  F  N  Q  Q  T  T  F  P  N  T
aatgctggtggagtcagttcatctggacctccgaacccctttgcaagctttactcaacaa
 N  A  G  G  V  S  S  S  G  P  P  N  P  F  A  S  F  T  Q  Q
tcaaataatcaacaaactgcattctccaataccaacgctggtggtctcagttcatctgga
 S  N  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  S  S  S  G
cctccgaatgcctttgcaagctttaataaacaaccaaatgctttcagcgtcaacactcct
 P  P  N  A  F  A  S  F  N  K  Q  P  N  A  F  S  V  N  T  P
caacctgttccttcaggtccctctggtttccaaaccaatccatcaacaacattcaaacca
 Q  P  V  P  S  G  P  S  G  F  Q  T  N  P  S  T  T  F  K  P
gcatcatttggacctggacccggatttgccacaactccacaaaacaataacatctttggt
 A  S  F  G  P  G  P  G  F  A  T  T  P  Q  N  N  N  I  F  G
caatcaactccaacaccggctacaaacacttcccagaacaatcagaccgcattcaacttc
 Q  S  T  P  T  P  A  T  N  T  S  Q  N  N  Q  T  A  F  N  F
aatgtccctgttgacatcttttactgctcctgctataaacacgacaaatacatcttccgg
 N  V  P  V  D  I  F  Y  C  S  C  Y  K  H  D  K  Y  I  F  R
aaccgagctgcaaataggtggtgatcctgttgatagtagtatctggctaaaggagaaatg
 N  R  A  A  N  R  W  -  S  C  -  -  -  Y  L  A  K  G  E  M
gaatccaggggagattccggaacaagcgccgcctgatgcctttgtttaa
 E  S  R  G  D  S  G  T  S  A  A  -  C  L  C  L
```

ZnF2c amino-acid sequences:

Met R K E L C R N F Q R G S C R Y G E N C R F L H P Q Q A K P N N P F G F G T Q N Q Q Q Q Q Q Q Q Q Q N S S N P F G F G V Q S G G S S R P N Q F Q N T W S R T A S T P T G G G A A A S T Q Q T G K Q T Q P A D H K C T D P A A C K R V **Met** Q D D F K N E R P **Met** W K L T C Y G H W K Y F P C D V T G D I S Y E E L R A V A Y E E A K R G I P L Q S I V E R E R N L Q N S K I A E F E N F L R N P Y K G S V T A N Q S P F A A T T P S I F P Q S S Q I N S P S P A F S G F N Q Q T A F S N T N A G G L S S S G P P N A F A S F N Q Q T T F P N T N A G G V R Y **Stop** L S P Y S I F K Q V C V T F I T S W F P Y F S L A N Y S S A **Stop** F I W T S E P L C K L Y S T I K **Stop** S T N C I L Q Y Q R W W S Q F I W T S E C L C K L **Stop Stop** T T K C F Q R Q H S S T C S F R S L W F P N Q S I N N I Q T S I I W T W T R I C H N S T K Q **Stop**H L W S I N S N T G Y K H F P E Q S D R I Q L Q C P C **Stop** H L L L L L L **Stop** T R Q I H L P E P S C K **Stop** V V I L L I V V S G **Stop** R R N G I Q G R F R N K R R L **Met** P L F

ZnF2c nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atgaggaaggaactgtgtagaaactttcagcgtggcagttgtaggtatggggaaaactgt
 M  R  K  E  L  C  R  N  F  Q  R  G  S  C  R  Y  G  E  N  C
agatttcttcatccacaacaagctaagccaaataatccctttggctttggtacacaaaac
 R  F  L  H  P  Q  Q  A  K  P  N  N  P  F  G  F  G  T  Q  N
caacaacaacaacagcagcagcagcaacaaaatagttcgaaccctttggatttggtgta
 Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  N  S  S  N  P  F  G  F  G  V
caaagtggtggttccagcagaccgaatcagttccagaatacttggtcacgaacggcttct
 Q  S  G  G  S  S  R  P  N  Q  F  Q  N  T  W  S  R  T  A  S
acacctactggtggtggtgctgctgcttctacacagcagactggtaaacagacgcagccg
```

```
 T  P  T  G  G  G  A  A  A  S  T  Q  Q  T  G  K  Q  T  Q  P
gcagatcataaatgcacagatcctgctgcgtgtaagcgggtaatgcaagacgattttaag
 A  D  H  K  C  T  D  P  A  A  C  K  R  V  M  Q  D  D  F  K
aatgagagacccatgtggaagctcacatgctacggccactggaaatattttccatgtgat
 N  E  R  P  M  W  K  L  T  C  Y  G  H  W  K  Y  F  P  C  D
gttaccggtgatatcagctatgaagaactacgtgcagtggcatatgaagaagctaaacga
 V  T  G  D  I  S  Y  E  E  L  R  A  V  A  Y  E  E  A  K  R
ggaatacctctacagtcgattgttgaaagggagaggaatctgcaaaattccaaaatagcc
 G  I  P  L  Q  S  I  V  E  R  E  R  N  L  Q  N  S  K  I  A
gagtttgaaaactttctacggaatccatacaaaggctctgttaccgccaatcaaagcccg
 E  F  E  N  F  L  R  N  P  Y  K  G  S  V  T  A  N  Q  S  P
tttgctgcaaccactcctagtatcttccctcagtccagtcagatcaattccccttctcca
 F  A  A  T  T  P  S  I  F  P  Q  S  S  Q  I  N  S  P  S  P
gcattctctggctttaatcaacaaactgcattctccaataccaatgcgggtggactcagt
 A  F  S  G  F  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  S
tcatctggacctccgaatgcctttgcaagcttcaatcaacaaactacattccccaatacc
 S  S  G  P  P  N  A  F  A  S  F  N  Q  Q  T  T  F  P  N  T
aatgctggtggagtcaggtactaactttctccatattccatattcaaacaagtttgtgtt
 N  A  G  G  V  R  Y  -  L  S  P  Y  S  I  F  K  Q  V  C  V
acctttataacctcttggttcccatatttctctttggccaactattcctctgcatagttc
 T  F  I  T  S  W  F  P  Y  F  S  L  A  N  Y  S  S  A  -  F
atctggacctccgaaccccttgcaagctttactcaacaatcaaataatcaacaaactgc
 I  W  T  S  E  P  L  C  K  L  Y  S  T  I  K  -  S  T  N  C
attctccaataccaacgctggtggtctcagttcatctggacctccgaatgcctttgcaag
 I  L  Q  Y  Q  R  W  W  S  Q  F  I  W  T  S  E  C  L  C  K
ctttaataaacaaccaaatgctttcagcgtcaacactcctcaacctgttccttcaggtcc
 L  -  -  T  T  K  C  F  Q  R  Q  H  S  S  T  C  S  F  R  S
ctctggtttccaaaccaatccatcaacaacattcaaaccagcatcatttggacctggacc
 L  W  F  P  N  Q  S  I  N  N  I  Q  T  S  I  I  W  T  W  T
cggatttgccacaactccacaaaacaataacatctttggtcaatcaactccaacaccggc
 R  I  C  H  N  S  T  K  Q  -  H  L  W  S  I  N  S  N  T  G
tacaaacacttcccagaacaatcagaccgcattcaacttcaatgtccctgttgacatctt
 Y  K  H  F  P  E  Q  S  D  R  I  Q  L  Q  C  P  C  -  H  L
ttactgctcctgctataaacacgacaaatacatcttccggaaccgagctgcaaataggtg
 L  L  L  L  L  -  T  R  Q  I  H  L  P  E  P  S  C  K  -  V
gtgatcctgttgatagtagtatctggctaaaggagaaatggaatccaggggagattccgg
 V  I  L  L  I  V  V  S  G  -  R  R  N  G  I  Q  G  R  F  R
aacaagcgccgcctgatgcctttgttta
 N  K  R  R  L  M  P  L  F
```

ZnF2d amino-acid sequences:

Met R K E L C R N F Q R G S C R Y G E N C R F L H P Q Q A K P N
N P F G F G T Q N Q Q Q Q Q Q Q Q Q Q N S S N P F G F G V Q S G
G S S R P N Q F Q N T W S R T A S T P T G G G A A A S T Q Q T G
K Q T Q P A D H K C T D P A A C K R V **Met** Q D D F K N E R P **Met** W
K L T C Y G H W K Y F P C D V T G D I S Y E E L R A V A Y E E A K
R G I P L Q S I V E R E R N L Q N S K I A E F E N F L R N P Y K G S
V T A N Q S P F A A T T P S I F P Q S S Q I N S P S P A F S G F N Q
Q T A F S N T N A G G L S S S G P P N A F A S F N Q Q T T F P N T
N A G G V S S S G P P N P F A S F T Q Q S N N Q Q T A F S N T N A
G G L R Y Q L L W A F S I F Q I P **Stop** R G A F **Met** K T L F P Y F S
L A N D S S T **Stop** F I W T S E C L C K L **Stop Stop** T T K C F Q R Q

H S S T C S F R S L W F P N Q S I N N I Q T S I I W T W T R I C H N S T K Q **Stop** H L W S I N S N T G Y K H F P E Q S D R I Q L Q C P C **Stop** H L L L L L L **Stop** T R Q I H L P E P S C K **Stop** V V I L L I V V S G **Stop** R R N G I Q G R F R N K R R L **Met** P L F

ZnF2d nucleotide and amino-acid sequences, yellow mark indicates spliced sequence.

```
atgaggaaggaactgtgtagaaactttcagcgtggcagttgtaggtatggggaaaactgt
 M  R  K  E  L  C  R  N  F  Q  R  G  S  C  R  Y  G  E  N  C
agatttcttcatccacaacaagctaagccaaataatccctttggctttggtacacaaaac
 R  F  L  H  P  Q  Q  A  K  P  N  N  P  F  G  F  G  T  Q  N
caacaacaacaacagcagcagcagcaacaaaatagttcgaacccttttggatttggtgta
 Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  N  S  S  N  P  F  G  F  G  V
caaagtggtggttccagcagaccgaatcagttccagaatacttggtcacgaacggcttct
 Q  S  G  G  S  S  R  P  N  Q  F  Q  N  T  W  S  R  T  A  S
acacctactggtggtggtgctgctgcttctacacagcagactggtaaacagacgcagccg
 T  P  T  G  G  G  A  A  A  S  T  Q  Q  T  G  K  Q  T  Q  P
gcagatcataaatgcacagatcctgctgcgtgtaagcgggtaatgcaagacgatttaag
 A  D  H  K  C  T  D  P  A  A  C  K  R  V  M  Q  D  D  F  K
aatgagagacccatgtggaagctcacatgctacggccactggaaatattttccatgtgat
 N  E  R  P  M  W  K  L  T  C  Y  G  H  W  K  Y  F  P  C  D
gttaccggtgatatcagctatgaagaactacgtgcagtggcatatgaagaagctaaacga
 V  T  G  D  I  S  Y  E  E  L  R  A  V  A  Y  E  E  A  K  R
ggaatacctctacagtcgattgttgaaagggagaggaatctgcaaaattccaaaatagcc
 G  I  P  L  Q  S  I  V  E  R  E  R  N  L  Q  N  S  K  I  A
gagtttgaaaactttctacggaatccatacaaaggctctgttaccgccaatcaaagcccg
 E  F  E  N  F  L  R  N  P  Y  K  G  S  V  T  A  N  Q  S  P
tttgctgcaaccactcctagtatcttccctcagtccagtcagatcaattccccttctcca
 F  A  A  T  T  P  S  I  F  P  Q  S  S  Q  I  N  S  P  S  P
gcattctctggctttaatcaacaaactgcattctccaataccaatgcgggtggactcagt
 A  F  S  G  F  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  S
tcatctggacctccgaatgcctttgcaagcttcaatcaacaaactacattccccaatacc
 S  S  G  P  P  N  A  F  A  S  F  N  Q  Q  T  T  F  P  N  T
aatgctggtggagtcagttcatctggacctccgaaccccttgcaagctttactcaacaa
 N  A  G  G  V  S  S  S  G  P  P  N  P  F  A  S  F  T  Q  Q
tcaaataatcaacaaactgcattctccaataccaacgctggtggtctcaggtatcaactt
 S  N  N  Q  Q  T  A  F  S  N  T  N  A  G  G  L  R  Y  Q  L
ttgtgggctttctccattttccaaattccgtaaaggggtgccttcatgaaaactctgttt
 L  W  A  F  S  I  F  Q  I  P  -  R  G  A  F  M  K  T  L  F
ccatatttctctctggccaatgattcctctacatagttcatctggacctccgaatgcctt
 P  Y  F  S  L  A  N  D  S  S  T  -  F  I  W  T  S  E  C  L
tgcaagctttaataaacaaccaaatgctttcagcgtcaacactcctcaacctgttccttc
 C  K  L  -  -  T  T  K  C  F  Q  R  Q  H  S  S  T  C  S  F
aggtccctctggtttccaaaccaatccatcaacaacattcaaaccagcatcatttggacc
 R  S  L  W  F  P  N  Q  S  I  N  N  I  Q  T  S  I  I  W  T
tggacccggatttgccacaactccacaaaacaataacatctttggtcaatcaactccaac
 W  T  R  I  C  H  N  S  T  K  Q  -  H  L  W  S  I  N  S  N
accggctacaaacacttcccagaacaatcagaccgcattcaacttcaatgtccctgttga
 T  G  Y  K  H  F  P  E  Q  S  D  R  I  Q  L  Q  C  P  C  -
catcttttactgctcctgctataaacacgacaaatacatcttccggaaccgagctgcaaa
 H  L  L  L  L  L  L  -  T  R  Q  I  H  L  P  E  P  S  C  K
taggtggtgatcctgttgatagtagtatctggctaaaggagaaatggaatccaggggaga
 -  V  V  I  L  L  I  V  V  S  G  -  R  R  N  G  I  Q  G  R
ttccggaacaagcgccgcctgatgcctttgttta
 F  R  N  K  R  R  L  M  P  L  F
```

In case of ZnF3, in the isoform of ZnF3a, 84 nucleotides were added by intron retention between the exon 2 and 3 (Figure 3A and C) that formed longer isoform ZnF3b (Table 2). I found, the length of 28 amino-acid sequences; GMSHFFHMPKIENNNTSSSIYASSS PVS are highly conserved among ZnF family proteins in between and even among plant species (Figure 9). It certainly indicates that these amino acids have very important conserved functions and these are only being changed due to alternative splicing. This peptide sequence is highly homologous (93%) even with AUXIN RESPONSE FACTOR (ARF) (Figure 9).

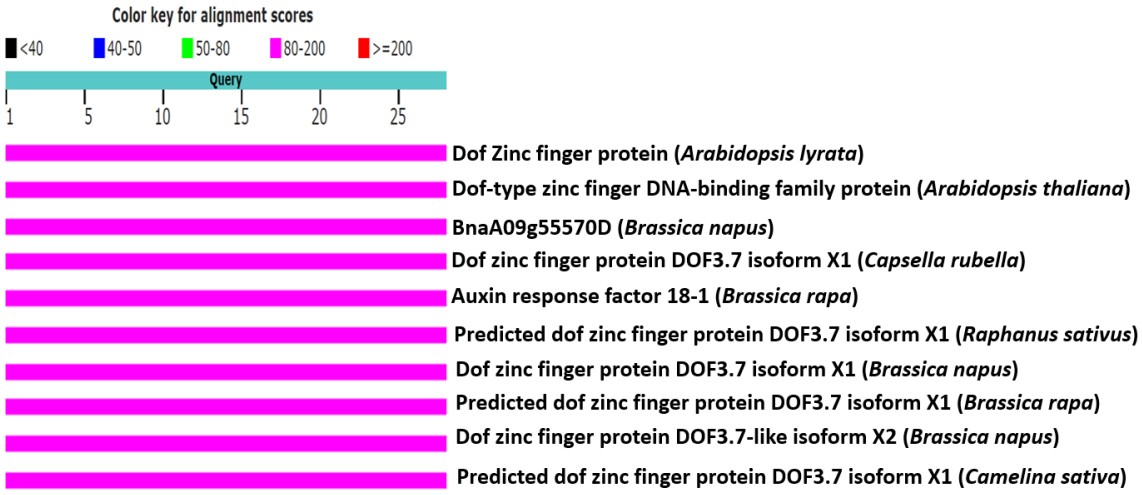


Figure 9. Homology of ZnF3 accession no. AT2G47850 (28 amino acid).

In ZnF5 and ZnF6, 82nt and 89 nt addition resulted in framshift respectively (Table 2). In case of ZnF7, I identified three references isoform ZnF7a, ZnF7b and ZnF7c, 3' alternative spliced site added 33 nucleotides creating ZnF7b (Figure 7A and C) that consist of 11 inframe added amino acids GGKCHLHKWVT (40-50). At the same time, in the reference isoform ZnF7c, 110 nucleotides were added by intron retention between the exon 2 and 3 (Figure 7A and C). In ZnF7c 110 nucleotide addition resulted in

frameshift (Table 2). At the same time, in the reference isoform ZnF7b, 3' alternative spliced site added 33 nucleotides that make an 11 amino acid. This peptide exhibits sequence homology (73%) with threonylcarbamoyl-AMP synthases.

ZnF7a amino-acid sequences:

Met D H Q N Y Q Y Q N P F E R R P I L K S K A P A V K W I K E W V
P Q D I V A T E D T F S R L K E K E K E P D V P E P E P E P T T E I
L F L C S Y D G C G K T F F D V S A L R K H S H I H G E R Q Y V C
D Q E G C G K K F L D S S K L K R H Y L I H T G E R N Y I C T Y E G
C G K L W **Met** Q A F S L D F N L R S H **Met** K T H S Q E N Y H I C P
Y S G C V K R Y A H E Y K L K N H V A A Y H E K N G G G E T P K Y
T P P A E K V L R T V K T P A T V C G P S S D R P Y A C P Y E G C
E K A Y I H E Y K L K L H L K R E H P G H L Q E E N A D T P T L N K
H N G N D R N E I D D G S D Q D V Y R K H A S N G K G Q T H K Q
Q S R A K P N **Met** R T P P A K V G K K G S T S S P A K A R I A K K
P W Q A K E T F E E V E R E E E E D S E E T E E D R D N V E D G W
R F G E N N E D D D D D E E T E Y E D **Stop**

ZnF7a nucleotide and amino-acid sequences:

```
atggatcatcaaaattatcaataccaaaatcctttcgagagacgacccatccttaaatcc
 M  D  H  Q  N  Y  Q  Y  Q  N  P  F  E  R  R  P  I  L  K  S
aaagctcctgctgtgaagtggatcaaagaatgggtaccacaagatattgttgctacagag
 K  A  P  A  V  K  W  I  K  E  W  V  P  Q  D  I  V  A  T  E
gatacttttagcagacttaaagaaaaagagaaagagcctgatgttcctgagcctgaacct
 D  T  F  S  R  L  K  E  K  E  K  E  P  D  V  P  E  P  E  P
gaaccaactacagagattttgtttctctgtagttatgatggttgcgggaagactttcttt
 E  P  T  T  E  I  L  F  L  C  S  Y  D  G  C  G  K  T  F  F
gatgttagtgcattgaggaaacattctcatatccatggagaaagacaatatgtttgtgat
 D  V  S  A  L  R  K  H  S  H  I  H  G  E  R  Q  Y  V  C  D
caggaaggatgtggaaagaaatttctggatagttcaaagttgaagagacattatcttatt
 Q  E  G  C  G  K  K  F  L  D  S  S  K  L  K  R  H  Y  L  I
catactggagagagaaattatatatgtacttatgaaggatgtggaaagttgtggatgcag
 H  T  G  E  R  N  Y  I  C  T  Y  E  G  C  G  K  L  W  M  Q
gcattctcgttggatttaaccttaggtctcacatgaagactcattcacaagagaattat
 A  F  S  L  D  F  N  L  R  S  H  M  K  T  H  S  Q  E  N  Y
catatatgtccttacagtgggtgtgtgaagagatatgctcatgaatacaagctaaagaac
 H  I  C  P  Y  S  G  C  V  K  R  Y  A  H  E  Y  K  L  K  N
cacgttgctgcctaccatgaaaagaatggtggtggagagactcccaaatatacaccacca
 H  V  A  A  Y  H  E  K  N  G  G  G  E  T  P  K  Y  T  P  P
gcagagaaagtattaaggactgtcaaaacacctgcaacagtttgtggcccgtcttcggat
 A  E  K  V  L  R  T  V  K  T  P  A  T  V  C  G  P  S  S  D
cggccatacgcatgcccttacgaagggtgtgagaaagcttacatacatgagtacaagctt
 R  P  Y  A  C  P  Y  E  G  C  E  K  A  Y  I  H  E  Y  K  L
aagctccacttgaagagagaacatccagggcatttacaagaagagaacgcggataccccc
 K  L  H  L  K  R  E  H  P  G  H  L  Q  E  E  N  A  D  T  P
acactgaacaagcacaatggcaatgacaggaatgagatagatgacgggagtgaccaagat
```

```
 T  L  N  K  H  N  G  N  D  R  N  E  I  D  D  G  S  D  Q  D
gtttacaggaaacacgctagtaatgggaaaggccagacacataaacaacagagcagagct
 V  Y  R  K  H  A  S  N  G  K  G  Q  T  H  K  Q  Q  S  R  A
aagccaaacatgaggacaccaccagccaaagttggaaagaaaggttctacctcttcgcct
 K  P  N  M  R  T  P  P  A  K  V  G  K  K  G  S  T  S  S  P
gccaaagcaaggattgcaaaaaaaccatggcaagcaaaagaaacttttgaagaagtagaa
 A  K  A  R  I  A  K  K  P  W  Q  A  K  E  T  F  E  E  V  E
agagaagaagaagaagatagcgaggagacagaggaagatagagataatgtggaggatggc
 R  E  E  E  E  D  S  E  E  T  E  E  D  R  D  N  V  E  D  G
tggaggtttggcgaaaacaacgaggatgatgacgacgatgaagagaccgagtatgaagat
 W  R  F  G  E  N  N  E  D  D  D  D  D  E  E  T  E  Y  E  D
tag
 -
```

ZnF7b amino-acid sequences; yellow mark indicates spliced sequence.

Met D H Q N Y Q Y Q N P F E R R P I L K S K A P A V K W I K E W V
P Q D I V A T G G K C H L H K W V T E D T F S R L K E K E K E P D
V P E P E P E P T T E I L F L C S Y D G C G K T F F D V S A L R K H
S H I H G E R Q Y V C D Q E G C G K K F L D S S K L K R H Y L I H T
G E R N Y I C T Y E G C G K A F S L D F N L R S H **Met** K T H S Q E
N Y H I C P Y S G C V K R Y A H E Y K L K N H V A A Y H E K N G G
G E T P K Y T P P A E K V L R T V K T P A T V C G P S S D R P Y A
C P Y E G C E K A Y I H E Y K L K L H L K R E H P G H L Q E E N A
D T P T L N K H N G N D R N E I D D G S D Q D V Y R K H A S N G K
G Q T H K Q Q S R A K P N **Met** R T P P A K V G K K G S T S S P A K
A R I A K K P W Q A K E T F E E V E R E E E E D S E E T E E D R D
N V E D G W R F G E N N E D D D D D E E T E Y E D **Stop**

ZnF7b nucleotide and amino-acid sequences:

```
atggatcatcaaaattatcaataccaaaatcctttcgagagacgacccatccttaaatcc
 M  D  H  Q  N  Y  Q  Y  Q  N  P  F  E  R  R  P  I  L  K  S
aaagctcctgctgtgaagtggatcaaagaatgggtaccacaagatattgttgctacaggt
 K  A  P  A  V  K  W  I  K  E  W  V  P  Q  D  I  V  A  T  G
ggaaagtgtcatctacacaaatgggtcacagaggatactttagcagacttaaagaaaaa
 G  K  C  H  L  H  K  W  V  T  E  D  T  F  S  R  L  K  E  K
gagaaagagcctgatgttcctgagcctgaacctgaaccaactacagagattttgtttctc
 E  K  E  P  D  V  P  E  P  E  P  E  P  T  T  E  I  L  F  L
tgtagttatgatggttgcgggaagactttctttgatgttagtgcattgaggaaacattct
 C  S  Y  D  G  C  G  K  T  F  F  D  V  S  A  L  R  K  H  S
catatccatggagaaagacaatatgtttgtgatcaggaaggatgtggaaagaaatttctg
 H  I  H  G  E  R  Q  Y  V  C  D  Q  E  G  C  G  K  K  F  L
gatagttcaaagttgaagagacattatcttattcatactggagagagaaattatatatgt
 D  S  S  K  L  K  R  H  Y  L  I  H  T  G  E  R  N  Y  I  C
acttatgaaggatgtggaaaggcattctcgttggattttaaccttaggtctcacatgaag
 T  Y  E  G  C  G  K  A  F  S  L  D  F  N  L  R  S  H  M  K
actcattcacaagagaattatcatatatgtccttacagtgggtgtgtgaagagatatgct
 T  H  S  Q  E  N  Y  H  I  C  P  Y  S  G  C  V  K  R  Y  A
catgaatacaagctaaagaaccacgttgctgcctaccatgaaaagaatggtggtggagag
```

```
 H  E  Y  K  L  K  N  H  V  A  A  Y  H  E  K  N  G  G  G  E
actcccaaatatacaccaccagcagagaaagtattaaggactgtcaaaacacctgcaaca
 T  P  K  Y  T  P  P  A  E  K  V  L  R  T  V  K  T  P  A  T
gtttgtggcccgtcttcggatcggccatacgcatgcccttacgaagggtgtgagaaagct
 V  C  G  P  S  S  D  R  P  Y  A  C  P  Y  E  G  C  E  K  A
tacatacatgagtacaagcttaagctccacttgaagagagaacatccagggcatttacaa
 Y  I  H  E  Y  K  L  K  L  H  L  K  R  E  H  P  G  H  L  Q
gaagagaacgcggatacccccacactgaacaagcacaatggcaatgacaggaatgagata
 E  E  N  A  D  T  P  T  L  N  K  H  N  G  N  D  R  N  E  I
gatgacgggagtgaccaagatgtttacaggaaacacgctagtaatgggaaaggccagaca
 D  D  G  S  D  Q  D  V  Y  R  K  H  A  S  N  G  K  G  Q  T
cataaacaacagagcagagctaagccaaacatgaggacaccaccagccaaagttggaaag
 H  K  Q  Q  S  R  A  K  P  N  M  R  T  P  P  A  K  V  G  K
aaaggttctacctcttcgcctgccaaagcaaggattgcaaaaaaaccatggcaagcaaaa
 K  G  S  T  S  S  P  A  K  A  R  I  A  K  K  P  W  Q  A  K
gaaacttttgaagaagtagaaagagaagaagaagaagatagcgaggagacagaggaagat
 E  T  F  E  E  V  E  R  E  E  E  E  D  S  E  E  T  E  E  D
agagataatgtggaggatggctggaggtttggcgaaaacaacgaggatgatgacgacgat
 R  D  N  V  E  D  G  W  R  F  G  E  N  N  E  D  D  D  D  D
gaagagaccgagtatgaagattag
 E  E  T  E  Y  E  D  -
```

ZnF7c amino-acid sequences:

K V L L K V S L L Y R S F P Q N P N L P Q I L I Q S Y S K L T K E P P L P K **Stop** I N G S S K L S I P K S F R E T T H P **Stop** I Q S S C C E V D Q R **Met** G T T R Y C C Y R W K V S S T Q **Met** G H R **Stop** N C D V L I **Stop** F S C L **Stop** **Met** L N C I I V L L L L P E D T F S R L K E K E K E P D V P E P E P E P T T E I L F L C S Y D G C G K T F F D V S A L R K H S H I H G E R Q Y V C D Q E G C G K K F L D S S K L K R H Y L I H T G E R N Y I C T Y E G C G K A F S L D F N L R S H **Met** K T H S Q E N Y H I C P Y S G C V K R Y A H E Y K L K N H V A A Y H E K N G G G E T P K Y T P P A E K V L R T V K T P A T V C G P S S D R P Y A C P Y E G C E K A Y I H E Y K L K L H L K R E H P G H L Q E E N A D T P T L N K H N G N D R N E I D D G S D Q D V Y R K H A S N G K G Q T H K Q Q S R A K P N **Met** R T P P A K V G K K G S T S S P A K A R I A K K P W Q A K E T F E E V E R E E E E D S E E T E E D R D N V E D G W R F G E N N E D D D D D E E T E Y E D **Stop** F E L R L S H R R P F W L R P F S L C E G G F L V A L F L D S S S F R L P **Stop** H S S S F K A L Y S I P L T F L K T F S F S F G Q L F E R F F F

ZnF7c nucleotide and amino-acid sequences; yellow mark indicates spliced sequence.

```
tcaaagtattgcttaaggtctctcttctctatcgttcttttccccagaaccctaatctcccc
   K  V  L  L  K  V  S  L  L  Y  R  S  F  P  Q  N  P  N  L  P
caaatcctgattcaatcgtactcaaaactcacaaaggaaccacctttacccaaataaatc
 Q  I  L  I  Q  S  Y  S  K  L  T  K  E  P  P  L  P  K  -  I
aatggatcatcaaaattatcaataccaaaatcctttcgagagacgacccatccttaaatc
 N  G  S  S  K  L  S  I  P  K  S  F  R  E  T  T  H  P  -  I
caaagctcctgctgtgaagtggatcaaagaatgggtaccacaagatattgttgctacagg
 Q  S  S  C  C  E  V  D  Q  R  M  G  T  T  R  Y  C  C  Y  R
tggaaagtgtcatctacacaaatgggtcacaggtgaaattgtgatgttttaatctaattt
 W  K  V  S  S  T  Q  M  G  H  R  -  N  C  D  V  L  I  -  F
tcatgtttgtgaatgttaaattgtatcattgtgttgttgttattaccagaggatactttt
 S  C  L  -  M  L  N  C  I  I  V  L  L  L  L  P  E  D  T  F
agcagacttaaagaaaaagagaaagagcctgatgttcctgagcctgaacctgaaccaact
 S  R  L  K  E  K  E  K  E  P  D  V  P  E  P  E  P  E  P  T
acagagattttgtttctctgtagttatgatggttgcgggaagactttctttgatgttagt
 T  E  I  L  F  L  C  S  Y  D  G  C  G  K  T  F  F  D  V  S
gcattgaggaaacattctcatatccatggagaaagacaatatgtttgtgatcaggaagga
 A  L  R  K  H  S  H  I  H  G  E  R  Q  Y  V  C  D  Q  E  G
tgtggaaagaaatttctggatagttcaaagttgaagagacattatcttattcatactgga
 C  G  K  K  F  L  D  S  S  K  L  K  R  H  Y  L  I  H  T  G
gagagaaattatatatgtacttatgaaggatgtggaaaggcattctcgttggattttaac
 E  R  N  Y  I  C  T  Y  E  G  C  G  K  A  F  S  L  D  F  N
cttaggtctcacatgaagactcattcacaagagaattatcatatatgtccttacagtggg
 L  R  S  H  M  K  T  H  S  Q  E  N  Y  H  I  C  P  Y  S  G
tgtgtgaagagatatgctcatgaatacaagctaaagaaccacgttgctgcctaccatgaa
 C  V  K  R  Y  A  H  E  Y  K  L  K  N  H  V  A  A  Y  H  E
aagaatggtggtggagagactcccaaatatacaccaccagcagagaaagtattaaggact
 K  N  G  G  G  E  T  P  K  Y  T  P  P  A  E  K  V  L  R  T
gtcaaaacacctgcaacagtttgtggcccgtcttcggatcggccatacgcatgcccttac
 V  K  T  P  A  T  V  C  G  P  S  S  D  R  P  Y  A  C  P  Y
gaagggtgtgagaaagcttacatacatgagtacaagcttaagctccacttgaagagagaa
 E  G  C  E  K  A  Y  I  H  E  Y  K  L  K  L  H  L  K  R  E
catccagggcatttacaagaagagaacgcggatacccccacactgaacaagcacaatggc
 H  P  G  H  L  Q  E  E  N  A  D  T  P  T  L  N  K  H  N  G
aatgacaggaatgagatagatgacgggagtgaccaagatgtttacaggaaacacgctagt
 N  D  R  N  E  I  D  D  G  S  D  Q  D  V  Y  R  K  H  A  S
aatgggaaaggccagacacataaacaacagagcagagctaagccaaacatgaggacacca
 N  G  K  G  Q  T  H  K  Q  Q  S  R  A  K  P  N  M  R  T  P
ccagccaaagttggaaagaaaggttctacctcttcgcctgccaaagcaaggattgcaaaa
 P  A  K  V  G  K  K  G  S  T  S  S  P  A  K  A  R  I  A  K
aaaccatggcaagcaaaagaaacttttgaagaagtagaaagagaagaagaagaagatagc
 K  P  W  Q  A  K  E  T  F  E  E  V  E  R  E  E  E  E  D  S
gaggagacagaggaagatagagataatgtggaggatggctggaggtttggcgaaaacaac
 E  E  T  E  E  D  R  D  N  V  E  D  G  W  R  F  G  E  N  N
gaggatgatgacgacgatgaagagaccgagtatgaagattagtttgagcttcgtctaagc
 E  D  D  D  D  D  E  E  T  E  Y  E  D  -  F  E  L  R  L  S
cataggagacctttttggttacgtcccttctctttatgcgaaggtggctttttagtagct
 H  R  R  P  F  W  L  R  P  F  S  L  C  E  G  G  F  L  V  A
ttgtttcttgactcttcttcttttcgtttgccttgacactcttcttcatttaaagctttg
 L  F  L  D  S  S  S  F  R  L  P  -  H  S  S  S  F  K  A  L
tacagtattccgttaacatttttgaaaaccttttcttttagttttggtcaacttttttgaa
 Y  S  I  P  L  T  F  L  K  T  F  S  F  S  F  G  Q  L  F  E
agatttttttttt
 R  F  F  F
```

3.4 Discussion

The process of alternative splicing has been the subject of extensive study in mammalian cells, however, much yet remains unknown about the process in plant cells. The process is evidently greatly varied in plants and heavily influences transcript diversity. In the plant *Arabidopsis thaliana*, instances of splicing that involve only single-nucleotide exons have been reported (Guo and Liu 2015). In this study, alternative splicing occurring amongst the Zinc Finger gene family is the main focus. Information obtained from the Uniprot database (http://www.uniprot.org) describes the functions of *ZnF1* and *ZnF5* as being involved in the binding of zinc metal ions, although the exact biological mechanism has not yet been fully elucidated. Conn et al. recently reported that exon 8/9 of *ZnF2* generate circular RNA which may bind with cognate DNA more stably forming DNA-RNA hybrid (Conn et al. 2017). In my study I found 7/8 exon containing isoform *ZnF2b* highly expressed in all tissues whereas exon skipping isoform *ZnF2a* expression level was very low in seedling 8 days; 27 and 32 days old stipe; 16 days old stem and root. However, exon 9 present in both isoforms. The consequence of such DNA-RNA hybrid may be lower transcription efficiency. The functions of *ZnF3*, however, have been identified as binding to chromatin, acting as a transcription factor by binding to DNA, in addition to behaving as a negative regulator of transcription by binding to DNA regulatory regions. *ZnF3* is in the group of Dof (for DNA binding with Zinc finger) family protein. Dof protein is a novel protein with highly conserved DNA binding domain and a C2-C2 zinc finger motif (Yanagisawa & Sheen, 1998). In the Dof protein, there is a nonconserved transcriptional factor in the C-terminal region which may interact with various partners (Wang et al. 2018; Yanagisawa & Sheen, 1998). Regarding *ZnF3*, intron retention was observed mainly in stipe, stem and root (Figure 3A). Shorter isoform expressed in all

tissues (Figure 3A). In maize, Dof1 is a transcriptional activator whereas Dof2 is a transcriptional repressor. In maize, light stimulates Dof1 to bind to (AAAAGG) target site in green leaves and activate transcription (Yanagisawa & Sheen, 1998). However, in non-green leaves, its transcriptional activity was inhibited. In my study, it is interesting that in 4 days non-green seedling *ZnF3b* isoform was not expressed (Figure 3A). Considering the above statements, it can be stated that intron retention in ZnF3b may render the isoform loss of its transcriptional activity. *ZnF4* binds to zinc ions as well as DNA, catalyses hydrolysis reactions, exhibits nuclease activity and binds to RNA. *ZnF6* has been reported to bind to zinc ions as well as phosphatidylinositol in addition to ruffle assembly regulation. CELL DEATH RELATED ENDOSOMAL FYVE/SYLF PROTEIN 1 (CFS1) is responsible for autophagy regulator and repressor of cell death (Sutipatanasomboon et al. 2017). CFS1 and *ZnF6* (AT1G29800) have the similar FYVE and actin binding SH3YL1 Ysc84/Lsb4p Lsb3p plant FYVE (SYLF) domain. CFS1 and At1g29800 display a high degree of amino acid sequence conservation with 75% similarity in the FYVE domain and 71% similarity in the SYLF domain (Sutipatanasomboon et al. 2017). The alternate splicing portion of my interest is located at downstream of SYLF domain. In my study, shorter isoform (*ZnF6a*) expression was higher in almost in all tissues except root 27 days (Figure 6A and B). On the other hand, intron retention (*ZnF6b*) between exon 5 and 6 expressed comparatively lower level in seedling 4, 8 days; in leaves; stipe 27and 32 days; stem 16, 21 days; root 16, 27 days (Figure 6A and B). It can be stated that the ZnF6b play a vital role in the normal physiology of the plants. *ZnF7* accession no. AT4G06634 consist of mainly five C2-H2 zinc-finger motifs that can be classified as the N-terminal four fingers tandem arrays, another one fifth zinc finger is, 114 nucleotides downstream from the fourth zinc finger.

It is reported that *ZnF7* is a negative regulator of abscisic acid (ABA) pathway (Li, Wu, Li, Song, & Liu, 2016). ABA is an important hormone for plant development including seedling growth, seed maturation, and response to various environmental stresses. *ZnF7* has both transcription activation and repression domain. In previous finding expression of longer *ZnF7b* found moderate in seedling, higher in flower and absent in root (Li et al. 2016). This finding is similar with my shorter isoform (*ZnF7c*) that expressed in leaves, stipe, stem, and in 16 days old root. However, I found the longer isoform (*ZnF7b*) expressed in almost all tissues even higher in seedling and root (Figure 7A and B). This discrepancy may be due to my isoform level expression study and larger number sample size with age variation. The important thing that the splicing portion is located in the N-terminal region, therefore according to the finding of (Li et al. 2016) it can be stated that this splicing activity may affect the repression activity of this gene.

The alternative splicing of different *ZnF* genes results in varied expression levels in different tissues. In this study, it was identified that instances of intron retention were better expressed than alternative 5' and 3' splice sites. Previously reported that, in the case of Arabidopsis, intron retention is more common types than other types of alternative splicing (Li et al.2016). Greater expression of *ZnF3* at the tissue-specific level was observed in this study. It has been reported that the editing of RNA is non-existent in *ndhB-149*, *ndhB-1255*, and *ndhD-2* of the roots as well as in lincomycin-treated seedlings (Tseng et al. 2013). It has also been established that *matK-640* is fully edited, while *accD-794*, *atpF-92*, *psbE-214*, *psbF-77*, *psbZ-50*, and *rps14-50* are completely or highly edited in both green and non-green tissues (Tseng et al. 2013). RNA-editing in *ndhD-1* has also been discovered to contain the DYW1 domain which exhibits the zinc-binding active site signature motif HxE(x)nCxxC. This motif is essential for RNA editing (Boussardon et al.

2014). Taking into account the results of this study and the aforementioned findings, it may be said that the alternative splicing of the Zinc Finger family genes may affect RNA editing at the tissue-specific level.

Shifting focus to the effects of alternative splicing on resulting protein, it has been seen that ZnF3b produces a 28 amino acid peptide GMSHFFHMPKIENNNTSSSIYASSSPVS (177-204) which is 93% homology with AUXIN RESPONSE FACTOR (ARF). AUXIN RESPONSE FACTOR whose main function is to regulate gene expression in response to auxins which eventually affects plant growth and development (Korasick et al. 2014; Xu et al. 2016). Another homologous (67%) peptide Ubiquitin carboxyl hydrolase L1 (UCH-L1), whose function is in the maintenance of the ubiquitin proteasome system (UPS). UCH-L1 has been reported, in recent studies, to be related to a number of human diseases (Suong et al. 2014). ZnF7b, which is made up of the 11 amino acids GGKCHLHKWVT (40-50) has 73% homology with threonylcarbamoyl-AMP synthase. Threonylcarbamoyl-AMP synthase catalyzes formation of biosynthetic intermediate with L-threonine and ATP conjointly, that play role in enzymatic pathway (Harris et al. 2015).

This study revealed that, ZnF proteins are greatly diversified as a result of alternative splicing. In another study, it was also discovered that the protein-protein interacting pathway was also altered as a result of tissue-specific alternative splicing (Ellis et al. 2012).

Different transcription factors have several ZnF motifs which are involved in DNA binding and protein-protein interactions. Following the research on plants, several ZnF motifs were identified which play a key role. These proteins actually have a key role in different biological processes such as pathogenic immunity, floral development and light

mediated morphogenesis. This proves that alternative splicing events actually alter binding locations as well as change the binding affinity. Collectively, it can be hypothesized that alternative splicing events in the ZnF transcript largely affecting Arabidopsis plant physiology. More investigation regarding the localization of the alternatively spliced transcript and protein, site-specific and tissue-specific RNA editing need to be investigated for understanding their effect on growth and development. Therefore, understanding such natural editing events in the transcript and how they affecting substrate recognition and plant physiology is interesting. However more experimental results are needed to understand the precise effect of alternative splicing and editing events.

3.5 References

Appelhagen I, Huep G, Lu GH et al (2010) Weird fingers: Functional analysis of WIP domain proteins. FEBS Letters; doi: 10.1016/j.febslet.2010.06.007

Boussardon C, Avon A, Kindgren P, et al (2014) The cytidine deaminase signature HxE(x)nCxxC of DYW1 binds zinc and is necessary for RNA editing of ndhD-1. New Phytol. 2014 Sep; 203(4):1090-5. doi: 10.1111/nph.12928. Epub 2014 Jul 10

Chang T, Huang H, Hsu JB, et al (2013) An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC Bioinformatics 14:1-8

Conn VM, Hugouvieux V, Nayak A, Conos SA, et al (2017) A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. 2017 Apr 18;3: 17053. doi: 10.1038/nplants.2017.53

Ellis JD, Barrios-Rodiles M, Çolak R, et al (2012) Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Mol Cell 46:884–892 . doi: 10.1016/j.molcel.2012.05.037

Graveley BR (2001) Alternative splicing: Increasing diversity in the proteomic world. Trends Genet 17:100–107 . doi: 10.1016/S0168-9525(00)02176-4

Guo L, Liu CM (2015) A single-nucleotide exon found in Arabidopsis. Sci Rep 5:1–5 . doi: 10.1038/srep18087

Harris KA, Bobay BG, Sarachan KL, et al (2015) NMR-based Structural Analysis of Threonylcarbamoyl-AMP Synthase and Its Substrate Interactions. J Biol Chem. 2015 Aug 14;290(33):20032-43. doi: 10.1074/jbc.M114.631242

James AB, Syed NH, Bordage S, et al (2012) Alternative Splicing Mediates Responses of the Arabidopsis Circadian Clock to Temperature Changes. Plant Cell 24:961–981 . doi: 10.1105/tpc.111.093948

Kelemen O, Convertini P, Zhang Z, et al (2013) Function of alternative splicing. Gene 514:1–30 . doi: 10.1016/j.gene.2012.07.083

Kent WJ (2002) BLAT — The BLAST -Like Alignment Tool. Genome Res 12:656–664 . doi: 10.1101/gr.229202

Korasick DA, Westfall CS, Lee SG, et al (2014) Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. https://doi.org/10.1073/pnas.1400074111

Laity JH, Lee BM et al (2001) Zinc finger proteins: new insights into structural and functional diversity. Current Opinion in Structural

Biology; https://doi.org/10.1016/S0959-440X(00)00167-6

Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of alternative splicing. Curr Opin Struct Biol 14:273–282 . doi: 10.1016/j.sbi.2004.05.002

Li, T., Wu, X. Y., Li, H., Song, J. H., & Liu, J. Y. (2016). A Dual-Function Transcription Factor, AtYY1, Is a Novel Negative Regulator of the Arabidopsis ABA Response Network. Molecular Plant, 9(5), 650–661. https://doi.org/10.1016/j.molp.2016.02.010

Li S, Yamada M, Han X, et al (2016) High-resolution expression map of the arabidopsis root reveals alternative splicing and lincRNA regulation. Dev Cell 39:508–522

Meng Y, Chen D, Jin Y, et al (2010) RNA editing of nuclear transcripts in Arabidopsis thaliana. BMC Genomics. doi: 10.1186/1471-2164-11-S4-S12

Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Science (80-) 338:1593–1599 . doi: 10.1126/science.1228186

Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes. Briefings Funct Genomics Proteomics 8:215–230

O'Reilly J-A, Fitzgerald J, Fitzgerald S, et al (2001) Diagnostic Potential of Zinc Finger Protein Specific Autoantibodies and Associated Linear B-Cell Epitopes in Colorectal Cancer. PLOS ONE; DOI:10.1371/journal.pone.0123469

Pan Q, Shai O, Lee LJ, et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415 . doi: 10.1038/ng.259

Stothard P (2000) The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. Bio Tech 28:1102–1104

Sun T, Bentolila S, et al (2016) The Unexpected Diversity of Plant organelle RNA editisomes. Trends Plant Sci: https://doi.org/10.1016/j.tplants.2016.07.005

Suong DN, Thao DT, Masamitsu et al. (2014) Ubiquitin Carboxyl Hydrolase L1 Significance for Human Disease. Protein & Peptide Letters: doi: 10.1073/pnas.1612460114

Sutipatanasomboon, A., Herberth, S., Alwood, E. G., Häweker, H., Müller, B., Shahriari, M., … Schellmann, S. (2017). Disruption of the plant-specific CFS1 gene impairs autophagosome turnover and triggers EDS1-dependent cell death. Scientific Reports. https://doi.org/10.1038/s41598-017-08577-8

Syed NH, Kalyna M, Marquez Y, et al (2012) Alternative splicing in plants - coming of age. Trends Plant Sci 17:616–623 . doi: 10.1016/j.tplants.2012.06.001

Sun T, Shi X, Friso G, et al (2015) A Zinc Finger Motif-Containing Protein Is Essential for Chloroplast RNA Editing. PLOS Genetics: doi: 10.1371/journal.pgen.1005028

Tseng C-C, Lee C-J, Chung Y-T, et al (2013) Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant J 82:375–392

Vasudevan AA, Smits SH, Höppner A, et al. (2013) Structural features of antiviral DNA cytidine deaminase. Biological Chemistry: doi: 10.1515/hsz-2013-0165

Wang, P., Li, J., Gao, X., Zhang, D., Li, A., & Liu, C. (2018). Genome-wide screening and characterization of the Dof gene family in physic nut (Jatropha curcas L.).

International Journal of Molecular Sciences, 19(6). https://doi.org/10.3390/ijms19061598

Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13:59–69. doi: 10.1038/nrg3095

Xu Z, Ji A, Song J, et al (2016) Genome-wide analysis of auxin response factor gene family members in medicinal model plant *Salvia miltiorrhiza*. Biology Open 2016: bio.017178 doi: 10.1242/bio.017178

Yanagisawa, S., & Sheen, J. (1998). Involvement of Maize Dof Zinc Finger Proteins in Tissue-Specific and Light-Regulated Gene Expression. The Plant Cell. https://doi.org/10.2307/3870630

Chapter IV

Analysis of Tissue-Specific RNA Editing Events Involved in RNA Editing in *Arabidopsis thaliana*

4.1 Introduction

RNA editing is defined as the process by which nucleotides in RNA molecules undergo modifications, such as additions or deletions. Therefore, RNA editing alters the flow of genetic information from the genome to mature RNA (Takenaka et al. 2013). Diverse types of RNA editing events have been reported in viruses, primitive eukaryotes, vertebrates, fungi, and plants. Interestingly, C-to-U and A-to-I(G) RNA editing is a common deamination phenomenon in animals and plants, whereas U-to-C RNA editing due to amination is specific to plants (Meng et al. 2010). In plants, a C-to-U deaminase enzyme, DYW domain protein, has been described (Shikanai 2015); however, an A-to-I(G) homologous deaminase has not yet been reported. In *Arabidopsis thaliana*, the RNA editing machinery, collectively referred to as the editosome, consists of at least four protein families. These include PPRs, MORFs/RIPs, ORRM1 and OZ1. The PPR gene family is one of the largest gene families in Arabidopsis, comprising approximately 450 genes (Lurin et al. 2004). These genes are either directly or indirectly responsible for RNA editing (Okuda et al. 2009; Doniwa et al. 2010; Chateigner-Boutin et al. 2013; Schallenberg-Rüdinger et al. 2013; Leu et al. 2016). The MORF/RIP, ORRM and OZ proteins are also directly or indirectly involved in C-to-U RNA editing and play an important role in the editosome (Sun et al. 2016). The plant quintuple editing factor 1 (QED1), PPR protein required for *accD* RNA editing 1 (RARE1) and NUWA proteins are functional homologs of the mammalian APOBEC-1 protein and are responsible for C-to-U RNA editing in plants (Guillaumot et al. 2017; Wagoner et al. 2015). These data indicate that RNA editing is a normal physiological phenomenon in plants.

Different theories have been proposed to describe the purpose of RNA editing. According to one theory, RNA editing is an evolutionary process, which did not occur in primitive organisms. According to another theory, RNA editing is responsible for the adaptation of organisms and repair of DNA damage caused by ultraviolet radiation or other environmental stresses (Barkan & Small, 2014; Takenaka et al. 2013). Plastid and mitochondrial genome editing events have mostly been investigated using bioinformatics and experimental approaches (Chen et al. 2017; He et al. 2016; Ichinose & Sugita, 2017; Rodrigues et al. 2017; Wu et al. 2017). However, RNA editing of nuclear transcripts has been studied only using bioinformatics approaches, for example, to investigate the frequency and distribution of RNA editing events in Arabidopsis (Meng et al. 2010).

The DYW domain, which shows C-to-U conversion activity, has a similar function as the cytidine deaminase protein (Shikanai 2015). The PPR proteins are a specialized type of protein that bind to their cognate RNA (Cheng et al. 2016). Like the PUF protein, the binding specificity of PPR proteins is almost reprogrammable, except for some differences (Cheng et al. 2016; Delannoy, Stanley, Bond, & Small, 2007; Yagi, Nakamura, & Small, 2014). The majority of the PPR family proteins are responsible for RNA editing in plants and play an important role in alternative splicing (Shikanai 2015). RNA editing efficiency varies with the tissue type and developmental stage; for example, RNA editing is more common in green tissues than in non-green tissues (Tseng et al. 2013).

In this study, I investigated RNA editing in various Arabidopsis tissues at different developmental stages of growth, with a special focus on PPR proteins. My data revealed tissue-specific RNA editing events in selected genes. The information gained in this study will advance our understanding of RNA editing in plants.

4.2 Materials and Methods

4.2.1 Growth conditions and sample collection

Arabidopsis ecotype Columbia (Col-0) were sown in paper pots filled with a mixture of horticultural perlite, peat moss, and vermiculite in a 1:2:1 ratio. Pots were covered with plastic, stored in the dark for 3–4 days, and then transferred to a growth room in green farm U.ING (made in Japan) maintained at 22°C and 45% relative humidity, and with a 16 h/8 h light/dark cycle. Plants were watered twice daily, in the morning and evening and fertilizer was applied twice a week. Different tissues were harvested at different developmental stages, including from 4-, 8-, and 12-d-old seedlings and 16-, 21-, 27-, and 32-d-old leaves, stipes, stems and roots.

4.2.2 Extraction of RNA and cDNA synthesis

Total RNA was extracted from the Arabidopsis tissues using QIAGEN Plant Mini Kit (Hilden city, Germany; catalog no. 74904), according to the manufacturer's instructions, and treated with RQ1 RNase-free DNase (Promega, Madison, WI, USA) to remove genomic DNA contamination. The RNA samples were then purified using phenol-chloroform extraction and ethanol precipitation and quantified using a NanoDrop Spectrophotometer (Thermo Scientific). Purified RNA was subjected to cDNA synthesis using reverse transcriptase (Superscript III, Invitrogen) and oligo dT primers. The Arabidopsis *GAPDH* gene was used as a housekeeping gene and was amplified using the forward primer GTTGTCATCTCTGCCCCAAG and reverse primer TGCAACTAGCGTTGGAAACA.

4.2.3 Selection of genes involved in RNA editing

To identify the best candidate genes involved in RNA editing, accession numbers of *PPR*, *ZnF*, *MORF/RIP*, *RRM* and other genes were obtained from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/). These accession numbers were entered in the www.plantgdb.org/AtGDB/ database to determine the genomic map of each gene. The genomic map location of each gene was then used to identify alternatively spliced genes from the Arabidopsis TAIR10 whole genome reference assembly (https://www.arabidopsis.org/). The full-length genomic DNA, mRNA, and cDNA sequences of each gene obtained from TAIR10 were validated against the sequence information (gene, coding sequence [CDS], intron-exon boundaries, transcript, and amino acid) available at http://atgenie.org.

Table 1: The accession number of our studied genes of RNA editing related family proteins

No.	Given gene name/Gene family	Accession.no (Ref isoform)	Annotation	AT°C	Thermal Cycle	Primer sequences		Length (bp)
						Forward	Reverse	
1	*PPR1a*	AT5G24060.1	Pentatricopeptide repeat	55	30	GAACACAGAGGACGGAGGAG	TTGACTTCCCAGTTGGCTTC	429
	PPR1b	AT5G24060.2						
2	*PPR2a*	AT4G19440.1	Pentatricopeptide repeat	58	35	GAGGAGCACGTTAGGATGGA	TGTGACCTGTAAACCGAGGAG	599
	PPR2b	AT4G19440.2						
3	*PPR3a*	AT2G19280.2	Pentatricopeptide repeat	58	30	TGATCCCGAACAGTCTCAAGT	CTCTCATCTTCCTCCGGCTA	304
	PPR3b	AT2G19280.1						
4	*PPR4a*	AT4G38150.1	Pentatricopeptide repeat	59	35	ATTTCCGTCTTTGGCTTGG	CAGTGCCTTTGGAGGATGAT	453
	PPR4b	AT4G38150.2						
5	*PPR5a*	AT1G30610.2, EMB2279	Pentatricopeptide repeat (Embryo defective 2279)	58	30	GGATATGGAAAGCCGTGGTA	GCCAGTGTAAGTCACCACGA	238
	PPR5b	AT1G30610.1, EMB2279						
6	*PPR6a*	AT3G59040.1	Pentatricopeptide repeat	58	30	TTCTTCCTCTCGGCTACTGC	GACCTTGGTCACCGTCAGTT	682
	PPR6b	AT3G59040.2						
7	*PPR7a*	AT4G01990.1	Pentatricopeptide repeat	55	30	GGAAGGCGTACCTGTCAAAA	GCCTTGCCTTTGAGAATTGA	974
8	*PPR8a*	AT3G05240.1	Pentatricopeptide repeat	57	30	TGGGGAAGTTTGATTTCTGG	TTGGGTTTTACCGGCATAGT	932
9	*PPR9a*	AT3G22690.1	Pentatricopeptide repeat	55	35	CGGATGACGATCTCGAAGTT	GTGCCCAGCATTTGAGTCTA	408
	PPR9b	AT3G22690.2						
10	*PPR10a*	AT4G21065.1	Pentatricopeptide repeat	55	35	TCCCATGTCTTACGCTCACA	CCAATCTTTGCACAAGCAGA	507
	PPR10b	AT4G21065.2						
11	*PPR11a*	AT1G18900.1	Pentatricopeptide repeat	58	30	CACCGTTTTTCACGGAGAGT	AGCTGAGGACAAAAGACTGGAG	255
	PPR11b	AT1G18900.3						
	PPR11c	AT1G18900.2						
12	*PPR12a*	AT1G76280.1	Pentatricopeptide repeat	57	30	CTTTGCCGAAAGGTCACAAT	AGGCTGCAACCGTAGCTAGA	731
	PPR12b	AT1G76280.2						
	PPR12c	AT1G76280.3						
13	*PPR13a*	AT2G41720.1	Pentatricopeptide repeat (Embryo defective 2654)	58	30	GAAGTGCAGGGATCGAGTTC	TTACAACAATGTGCCTTCGTG	206
	PPR13b	AT2G41720.2						
14	*PPR14a*	AT1G64430.1	Pentatricopeptide repeat	57	35	GACGAGGCCGAAAGACTACTT	GCGACCGACAAAAGAAACTTA	444
	PPR14b	AT1G64430.2						
15	*PPR15a*	AT1G05670.1	Pentatricopeptide repeat	58	30	AAGAACTGATGCTGGTGCAA	TGGAGAGATGATGTGGCTCTT	429
	PPR15b	AT1G05670.2						
16	*PPR16a*	AT5G27300.1	Pentatricopeptide repeat	58	30	GTTTTGATCCCACCCATCAC	TGAACGTGTTGGGACTGAAG	439
	PPR16b	AT5G27300.2						
17	*ZnF1a*	AT2G47850.1	Zinc finger C-x8-C-x5-C-x3-H type	58	30	CATGAGCCACGTTCCACTAA	GGCGCCTTGCATATAAGAAC	349
	ZnF1b	AT2G47850.2						

			family protein					
18	*ZnF2a* *ZnF2b*	AT1G75340.1 AT1G75340.2	Zinc finger C-x8-C-x5-C-x3-H type family protein	55	35	CCCTTCTC CAGCATTC TCTG	GTTGACG CTGAAA GCATTTG	605
19	*ZnF3a /DAG1a* *ZnF3b /DAG1b* *ZnF3c /DAG1c* *ZnF3d /DAG1d*	AT3G61850.1 AT3G61850.2 AT3G61850.3 AT3G61850.4	Dof-type zinc finger DNA-binding family protein	55	30	CCAGTCG GAGGTAG CTCAAG	GGCATTG TTGGAAA CCCTAA	380
20	*ZnF4a* *ZnF4b*	AT3G51950.1 AT3G51950.2	Zinc finger CCCH-type family protein	55	35	CAACCTGT CGTGTTGG GATA	TGTGTGG TCCATCA TCACAT	296
21	*ZnF5a* *ZnF5b*	AT1G06040.1, STO AT1G06040.2, STO	B-box zinc finger family protein	58	35	CAGCAGC AACAACA ACCTTC	CACCGA AGAATCC CATGTCT	236
22	*ZnF6a* *ZnF6b*	AT1G29800.1 AT1G29800.2	RING/FYVE /PHD-type zinc finger family protein	57	30	TCGCTTGA AGGAAGC ATTTT	CCAACG AGATGA AGCGAA TC	327
23	*ZnF7a* *ZnF7b* *ZnF7c*	AT4G06634.1 AT4G06634.2 AT4G06634.3	Zinc finger (C2H2 type) family protein	55	35	CTTTCGAG AGACGAC CCATC	TTCCCGC AACCATC ATAACT	380
24	*ZnF8a* *ZnF8b*	AT5G63260.1 AT5G63260.2	Zinc finger C-x8-C-x5-C-x3-H type family protein	57	35	CAGCGAT CAAAAGG AGGAAG	CAGCTTT GGGTTCT CCACAT	641
25	*ZnF9a* *ZnF9b*	AT1G78600.1 AT1G78600.2	Light-regulated zinc finger protein 1	55	35	TCGACAA AGCTTCCT CTCGT	TGTCCGG AACAAA CACAGA G	523
26	*MORF1a/RIP1a* *MORF1b/RIP1b*	AT1G53260.1 AT1G53260.2	Hypothetical protein	59	35	TTACAGA CGTCCACC ACCAA	CCTGTAG TTCTGGC CCATGT	340
27	*RRM1a* *RRM1b*	AT5G52040.1 AT5G52040.2	RNA-binding (RRM/RBD/RNP motifs) family protein	58	35	GATCTTGA GCGGCTTT TCAG	TGTCCAC TCAACAC GGAGTC	494
28	*RRM2a* *RRM2b*	AT3G53460.1 AT3G53460.2	RNA-binding (RRM/RBD/RNP motifs) family protein (CP29) Chloroplast RNA binding protein	57	35	GCTCAGCT TGCTCAGC TCTT	CCACTTC TTGGTCC TCTGGA	239
29	*PP01a* *PP01b*	AT4G01690.1 AT4G01690.2	Protoporphyrinogen oxidase family	58	30	ATTGGCG GGTCTAC AAACAC	GGCTACA CCAGCG ACGTAAT	358
30	*GSDA1a* *GSDA1b*	AT5G28050.1 AT5G28050.2	Cytidine/deoxycytidylate deaminase family protein	58	35	TCCACAC ACACCAC ACAAAA	CTTGTTG GTGACCG GAAAAT	216

The enlisted genes were confirmed by PCR and sequencing from PCR products. The table displays annealing temperature, thermal cycle and primer sequences (Forward and Reverse). These primer sequences were used for PCR and sequencing as well.

4.2.4 Primer design

Primers were designed using Primer3 (bioinfo.ut.ee/primer3-0.4.0/primer3/) and verified using the NCBI Primer-BLAST tool. In the case of failure or inappropriate outcome with the first set of primers, new primer sets were designed. Primers were purchased from Eurofins (Japan) in TE buffer at a concentration of 50 pmol/µl in a salt-free condition. Each primer was diluted to a working concentration of 10 pmol/µl using TE buffer.

4.2.5 Sequencing of PCR products

To identify RNA editing sites, PCR products were subjected to three times of direct sequencing on a 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Austin, TX, USA; catalog no. 4336935). Sequences of the reverse strand were reverse complemented using online software (http://www.bioinformatics.org/sms2/reference.html). All sequences were aligned with the Arabidopsis genome sequence using BLAST (Kent, 2002).

4.2.6 Quantification of different RNA editing events

RNA editing events were quantified from the sequencing results using the peak height ratio method. The peak height of the corresponding dual peaks was calculated using ImageJ software. RNA editing (%) was quantified based on the maximum peak height, according to the following equation (Eggington, Greene, & Bass, 2011): RNA editing (%) = [C height / (T height + C height)] × 100,

4.2.7 Prediction of mRNA secondary structure

The secondary structures of mRNAs were predicted using RNAfold version 2.1.6, with default parameters (Lorenz et al. 2011).

4.3 Results

In this study, I mainly focused on RNA editing sites in alternatively spliced isoforms. A total of thirty differentially expressed genes were identified with variable expression patterns in different Arabidopsis tissues. Thirty candidate genes belonging to different families were investigated, including *PPR* (16), *ZnF* (9) *MORF/RIP* (1), *RRM* (2), *PP01* (1), and *GSDA* (1) genes (Table 1). PCR products of all genes were sequenced three times, and transcript positions were determined from the http://atgenie.org database. Overall, I identified RNA editing events in *PPR* (11) genes and *ZnF* (1) gene; no editing events were detected in any of the remaining genes.

Table 2: Details of RNA editing events and the corresponding amino acid changes detected in Arabidopsis tissues at different developmental stages

Gene name-editing site in transcript	Editing site in genome and edited sequence	RNA editing in tissues (editing %) and edited sequence			Editing location (codon/ nucleotide change)	Amino acid change
PPR2a-2969	3267 TGAGA	16-d-old leaf (-)	21-d-old leaf (14%) TGGGA	27-d-old leaf (-)	3' UTR, A>I(G)	NA
PPR2b-2720	2983 AGACA	21-d-old stem (-)	27-d-old stem (37%) AGGCA	32-d-old stem (-)	3' UTR, A>I(G)	NA
PPR2b-2694	2957 AGATT	16-d-old stem (-)	21-d-old stem (-)	27-d-old stem (28%) AGGTT	3' UTR, A>I(G)	NA
PPR3b-2833	3089 CTAGT	16-d-old leaf (-)	21-d-old leaf (54%) CTGGT	32-d-old leaf (-)	3' UTR, A>I(G)	NA
PPR4b-1008	1333 CCAAA	8-d-old seedling (-)	16-d-old stem (23%) CCGAA	NA	3' UTR, A>I(G)	NA
PPR5a-2293	2999 GCAGG	4-d-old seedling (-)	8-d-old seedling (-)	12-d-old seedling (28%) GCTGG	GCA>GCU, A>U	A710A
PPR6a-134	134 TTTTC	4-d-old seedling (-)	8-d-old seedling (-)	12-d-old seedling (40%) TTCTC	UUC>CUC, U>C	F7L
PPR7a-1035	1144 CAAGA	16-d-old leaf (-)	21-d-old leaf (34%) CAGGA	27-d-old leaf (-)	AAG>AGG, A>I(G)	K336R
PPR8a-1260	1260 ATCAC	4-d-old seedling (-)	8-d-old seedling (-)	12-d-old seedling (30%) ATTAC	AUC>AUU, C>U	I420I
PPR9a-2749	2648 GAGCC	4-d-old seedling (-)	8-d-old seedling (-)	12-d-old seedling (26%) GACCC	AGC>ACC, G>C	S883T
PPR10a-355	355 ATGAG	16-d-old leaf (-)	21-d-old leaf (-)	27-d-old leaf (10%) ATAAG	AUG>AUA, G>A	M111I
PPR10a-512	512 TTCAT	16-d-old leaf (-)	21-d-old leaf (-)	27-d-old leaf (9%) TTTAT	CAU>UAU, C>U	H164Y
PPR11b-2809	3404 TTTAT	4-d-old seedling (-)	8-d-old seedling (-)	12-d-old seedling (41%)	3' UTR, U>G	NA

				TTGAT		
PPR11b-2852	3224 GGAGA	4-d-old seedling (-)	8-d-old seedling (23%) GGCGA	12-d-old seedling (-)	GAG>GCG, A>C	E831A
DAG1a-615	1296 CCTCT	16-d-old leaf (-)	21-d-old leaf (28%) CCACT	27-d-old leaf (-)	UCU>ACU, U>A	S130T
DAG1a-794	1559 AGAGG	16-d-old leaf (-)	21-d-old leaf (-)	27-d-old leaf (16%) AGTGG	AGA>AGU, A>U	R189S

Arabic numbers represent editing site in transcript and genome. Underline base indicates the edited base, "-" indicates the absence of RNA editing. NA, not applicable; 3' UTR, 3' untranslated region.

4.3.1 Tissue-specific RNA editing events

In this study, I identified deamination/amination and others tissue-specific RNA editing events in gene transcripts (Table 2). In the course of study, I identified 21 RNA editing sites in total 12 genes which are expressed in different tissues. Among these I observed deamination/amination type (11) and others type (10). In case of *PPR* gene family I found 19 editing sites among these deamination/amination type (11) and others type (9). But in case of *ZnF* gene I only detected others type (2) editing site.

Another thing I observed is RNA editing sites expressed in different tissues. In my analysis, I identified 10 editing sites which were expressed in seedling tissues including deamination/amination type (3) and others (7), 8 editing sites were expressed in leaf tissues including deamination/amination type (5) and others (3) and only deamination/amination type (3) was expressed in stem tissues.

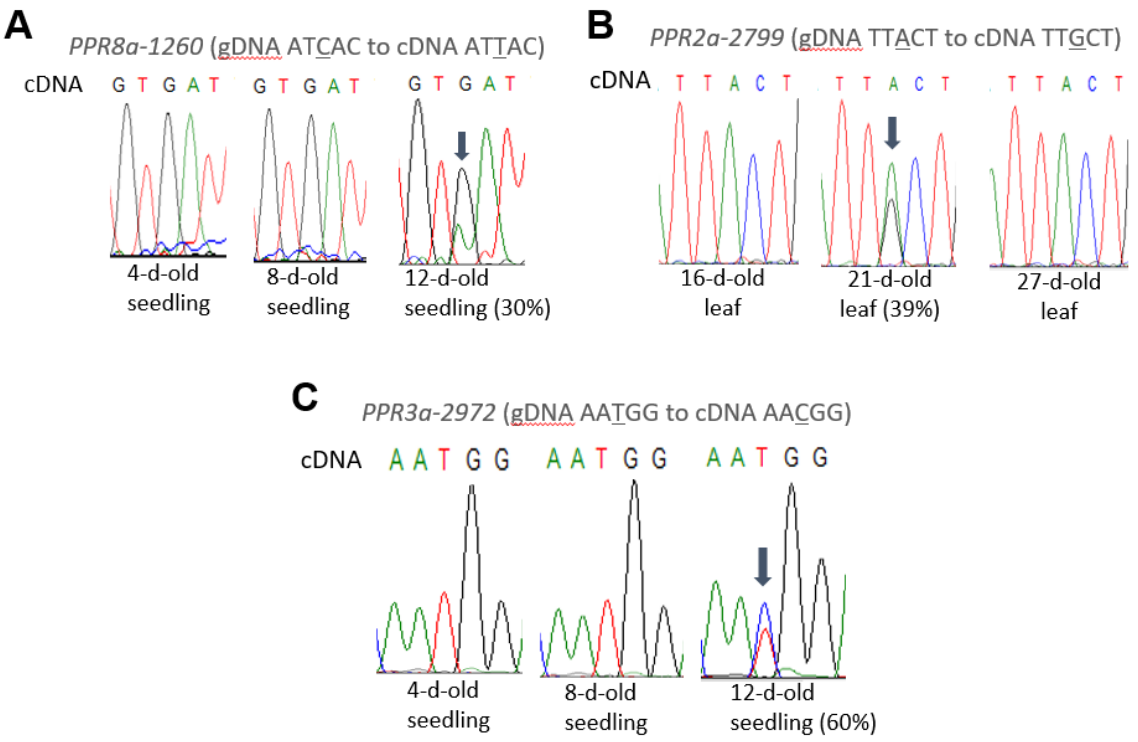


Figure 1. Validation of deaminase/aminase types of RNA editing events in Arabidopsis tissues at different developmental stages via cDNA sequencing. Sequencing of C-to-U editing events in 12-d-old seedlings using reverse primers (**A**), A-to-I(G) editing in 21-d-old leaves using forward primers (**B**), and U-to-C editing in 12-d-old seedlings using reverse primers (**C**). Arrow indicates the edited base. gDNA, genomic DNA.

In *PPR8a* gene, 30% C-to-U RNA editing was detected at 1,260 nt in 12-d-old seedlings but not in 4- or 8-d-old seedlings (Figure 1A). In *PPR2a* gene, 39% A-to-I(G) another deaminase-type editing was detected at 2,799 nt in 21-d-old leaves but not in 16- or 27-d-old leaves (Figure 1B). In *PPR3a* gene, 60% U-to-C an amination-type editing was detected at 2972 nt in 12-d-old seedlings but not in 4- or 8-d-old seedlings (Figure 1C).

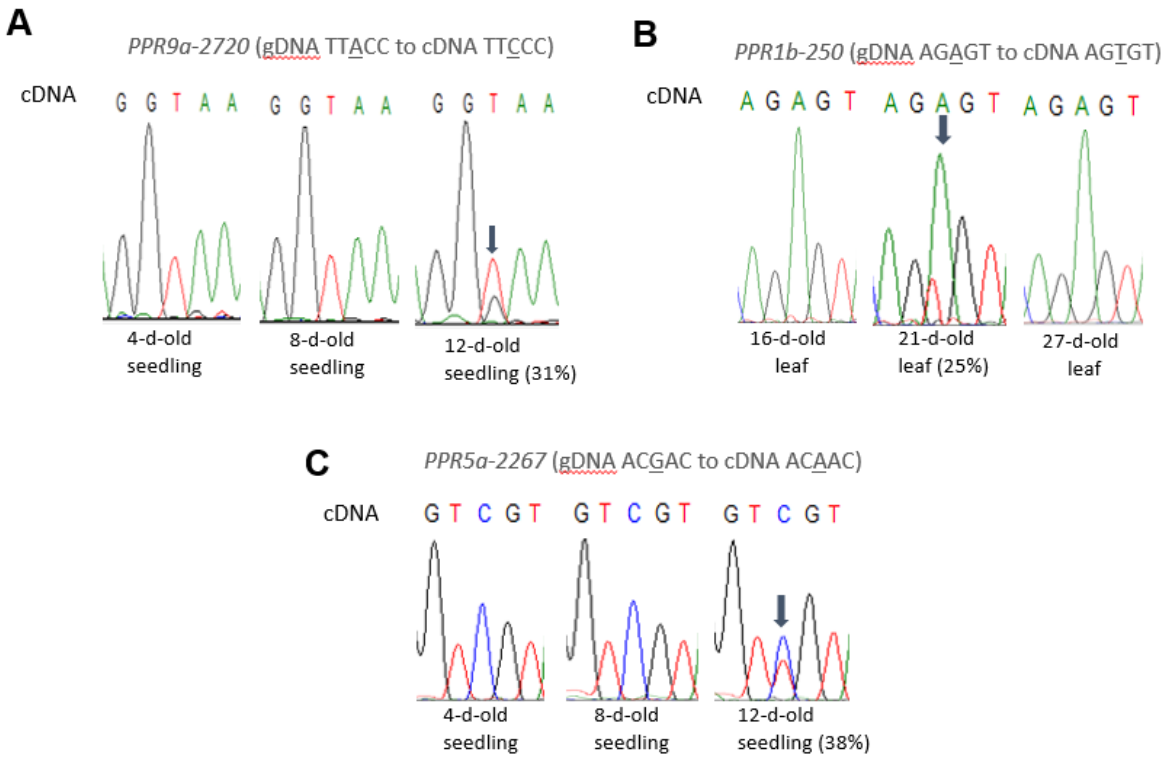


Figure 2. Validation of others types of RNA editing events in Arabidopsis tissues at different developmental stages via cDNA sequencing. Sequencing of A-to-C editing events in 12-d-old seedlings using reverse primers, amino acid change (L to F) (**A**), A-to-U editing in 21-d-old leaves using forward primers, amino acid change (E to V) (**B**), and G-to-A editing in 12-d-old seedlings using reverse primers, amino acid change (D to N) (**C**). Arrow indicates the edited base. gDNA, genomic DNA.

In *PPR9a* gene, 31% A-to-C another type of RNA editing was observed at 2,720 nt in 12-d-old seedlings but not in 4- or 8-d-old seedlings (Figure 2A). In *PPR1b* gene, 25% A-to-U RNA editing was identified at 250 nt in 21-d-old leaves, although it was not detected in 16- and 27-d-old leaves (Figure 2B). In *PPR5a* gene, 38% G-to-A RNA editing was identified at 2,267 in 12-d-old seedlings but not in 4- or 8-d-old seedlings (Figure 2C). RNA editing sites were more frequent in seedling and leaf tissues and less common in stem tissues but not in root and stipe tissues (Table 2). Additionally, RNA editing was

more predominant in the protein-coding regions of genes than in untranslated regions (Table 2).

4.3.2 Analysis of amino acid substitutions induced by RNA editing

RNA editing resulted in several amino acid substitutions in target genes (Table 2). In *PPR1b-250*, *PPR5a-2267*, *PPR6a-134*, *PPR7a-1035*, *PPR9a-2720*, *PPR9a-2749*, *PPR10a-355*, *PPR10a-512*, *PPR11b-2852*, *DAG1a-615*, and *DAG1a-794*, RNA editing changed glutamic acid to valine (E-to-V), aspartic acid to asparagine (D-to-N), phenylalanine to leucine (F-to-L), lysine to arginine (K-to-R), leucine to phenylalanine (L-to-F), serine to threonine (S-to-T), methionine to isoleucine (M-to-I), histidine to tyrosine (H-to-Y), glutamic acid to alanine (E-to-A), serine to threonine (S-to-T) and arginine to serine (R-to-S), respectively, as determined using the ExPASy translation tool. Additionally, two silent mutations were detected in *PPR5a-2293* and *PPR8a-1260* (Table 2).

Characterization of RNA editing events

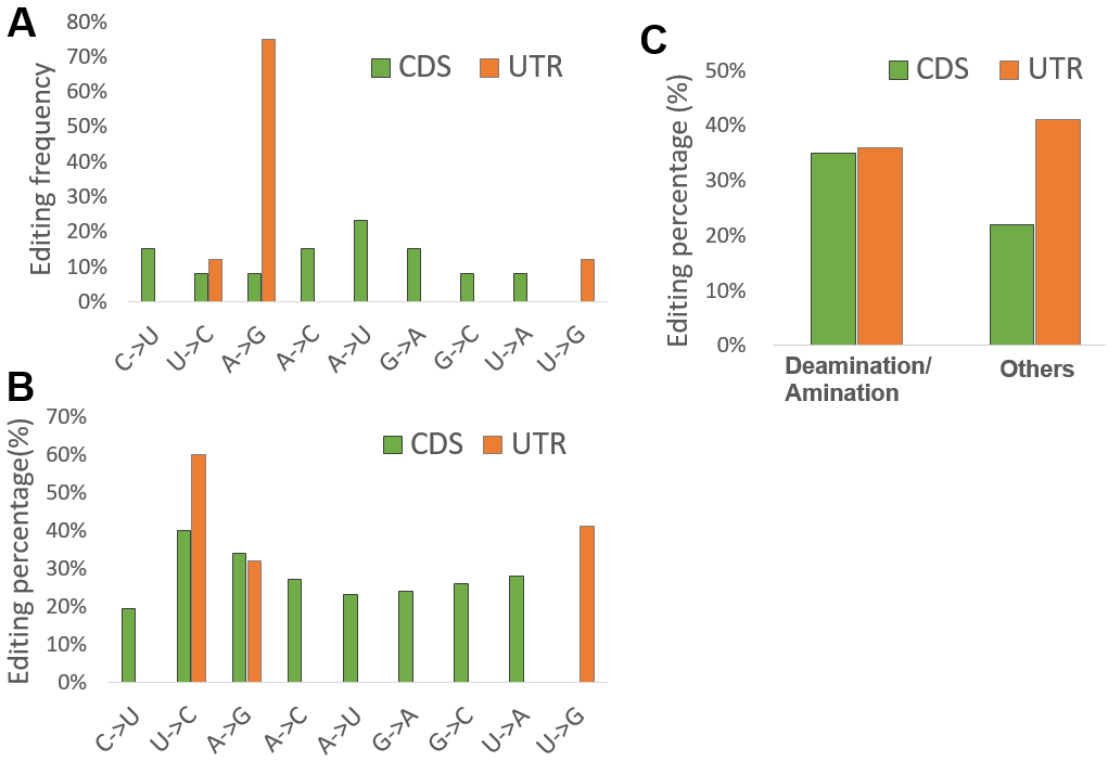


Figure 3. Schematic representation of RNA editing events in Arabidopsis tissues. **A.** Percentage of RNA editing events in the coding sequence (CDS) and untranslated region (UTR) of genes. **B.** Percentage of RNA editing in the CDS and UTR. **C.** Percentage of deamination/amination and others types of RNA editing events in the CDS and UTR. Green bar represents protein coding region, and orange bar indicates the untranslated region.

In this study, I identified nine types of RNA editing events. These events incorporated all possible intra-base substitutions: C-to-U, U-to-C, A-to-I(G), A-to-C, A-to-U, G-to-A, G-to-C, U-to-A, and U-to-G. The A-to-I(G) substitution was the most frequent type in the UTR regions (75%), whereas A-to-U substitution was the most frequent type in CDSs (23%) (Figure 3A). A group of substitutions, including C-to-U, A-to-C, and G-to-A, were the second most frequent editing type; each of these nucleotide changes occurred at a 15%

in the CDS of genes (Figure 3A). Depending on the editing percentage, U-to-C conversion was the most common (60%) and U-to-G was the second most common type (41%), among all of the UTR-specific RNA editing events (Figure 3B). In the case of A-to-I(G) editing, the editing frequency was lower than the editing percentage in CDSs of genes, whereas the reverse was true in the UTR regions (Figure 3A and B). Additionally, the frequency of deamination/amination types of RNA editing showed the highest percentage in the CDS regions and lowest in UTR regions (Figure 3C).

4.3.3 Distribution and effects of RNA editing

The distribution of RNA editing events showed that the frequency and type of RNA editing events were notably variable among the different Arabidopsis gene families. I identified 21 RNA editing sites in a total of 12 genes (Table 3). Of these 21 editing sites, 13 were located in the CDS of genes and 8 in the UTR regions (Table 3). RNA editing altered the secondary structure of three gene transcripts, each of which was edited in the UTR region (Table 3).

Table 3: Distribution of RNA editing sites in genic regions and their effect on mRNA structure

Gene family	Given gene name	Number of RNA editing sites		Effect on secondary structure of mRNA
		CDS	UTR	
PPR	*PPR1b*	1	0	-
	PPR2a/ PPR2b	0	4	+
	PPR3a/ PPR3b	0	2	+
	PPR4b	0	1	-
	PPR5a	2	0	-
	PPR6a	1	0	-
	PPR7a	1	0	-
	PPR8a	1	0	-
	PPR9a	2	0	-
	PPR10a	2	0	-
	PPR11b	1	1	+
ZnF	*DAG1a*	2	0	-

Arabic number indicates numbers of RNA editing site in CDS/UTR; "-" indicates no structural effect, and "+" indicates that RNA editing altered the mRNA structure. NA, not applicable; CDS, coding sequence; UTR, untranslated region.

4.3.4 Structural effects of RNA editing

To determine the potential functional effects of RNA editing, I predicted the secondary structure of mRNA using RNAfold version 2.1.6. The mRNA secondary structure was affected by A-to-I(G), U-to-C, and U-to-G RNA editing events. Alterations observed in the mRNA secondary structure was significant.

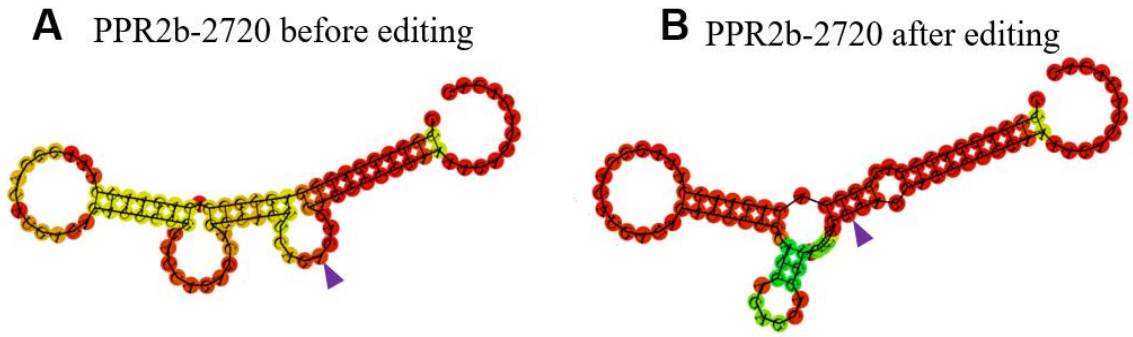


Figure 4. Effect of RNA editing on the secondary structure of mRNA. **A.** mRNA structure before editing. **B.** mRNA structure after A-to-I(G) editing.

In *PPR2b-2720* mRNA, 'A' was located in the bulge loop before editing, whereas the edited nucleotide 'G' was located in the stem after editing (Figure 4A and B).

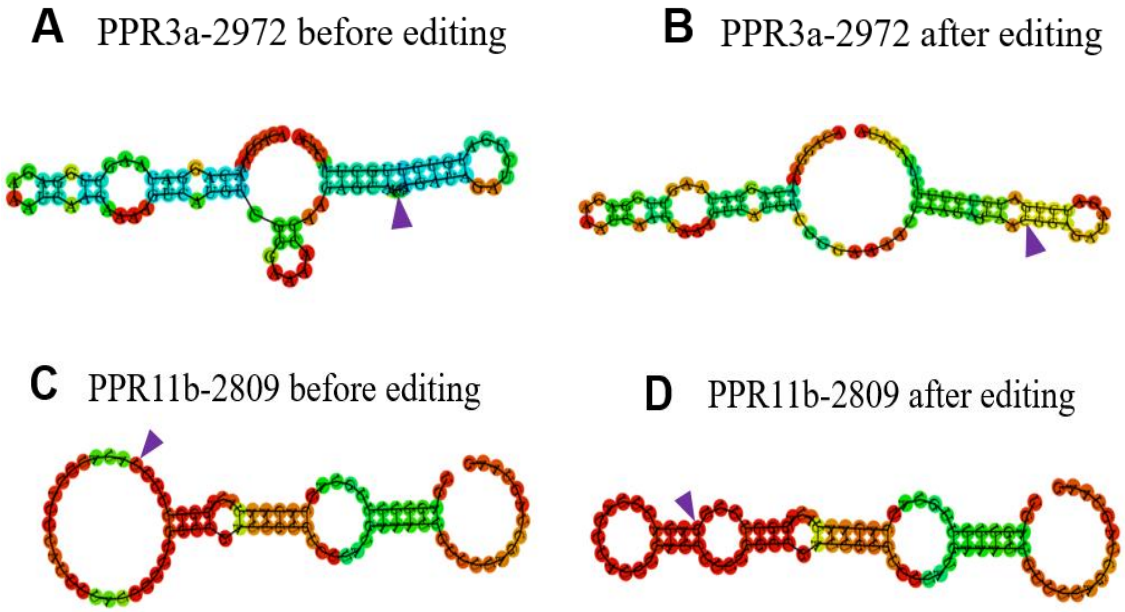


Figure 5. Effect of RNA editing on the secondary structure of mRNA. **A–B.** Secondary structure of mRNA before editing (**A**) and after U-to-C editing (**B**). **C–D.** Secondary structure of mRNA before editing (**C**) and after U-to-G editing (**D**). Arrow indicates the edited base.

In the case of *PPR3a-2972* transcript, 'U' was located in the bulge loop before editing, whereas the altered nucleotide 'C' was located in the hairpin stem after editing (Figure 5A and B). Similarly, in *PPR11b-2809*, 'U' was located in the hairpin loop before editing, whereas the edited nucleotide 'G' was located in the hairpin stem after editing; additionally, a single large loop was divided into two small loops after editing (Figure 5C and D).

4.3.5 Nucleotide preferences in the editing sites in gene CDSs

In this study, I identified 13 RNA editing sites located in the CDS of *PPR* (11) and *ZnF* (2) gene families. The frequency of RNA editing at the first, second, and third nucleotide positions of the codon was 38%, 31%, and 31%, respectively (Figure 6A). Among the RNA editing sites in which the first nucleotide position was altered, 50% of the sites contained 'A' (purine) at the second nucleotide position of the codon (Table 2 and Figure 2C). Sites in which the second nucleotide position of the codon was edited were of the configuration G_G (2), A_G, or A_C (Table 2 and Figure 2B). The codon sequence of sites in which the third nucleotide position was edited was of the configuration GC_, AU_ (2), UU_, or AG_ (Table 2 and Figure 2A). Most of the RNA editing sites were located in the gene CDS and resulted in amino acid changes. Of the 13 RNA editing sites observed in this study, 2 were silent; both of these occurred in (*PPR5a-2293* and *PPR8a-1260*) and resulted in codon change but no amino acid change (Table 2). Among a total of 21 RNA editing events, 11 resulted in amino acid alterations. RNA editing also affected the hydrophobicity/hydrophilicity of the encoded proteins; the proportion of hydrophobic amino acids increased from 38.46% before editing to 53.85% after editing (Figure 6B).

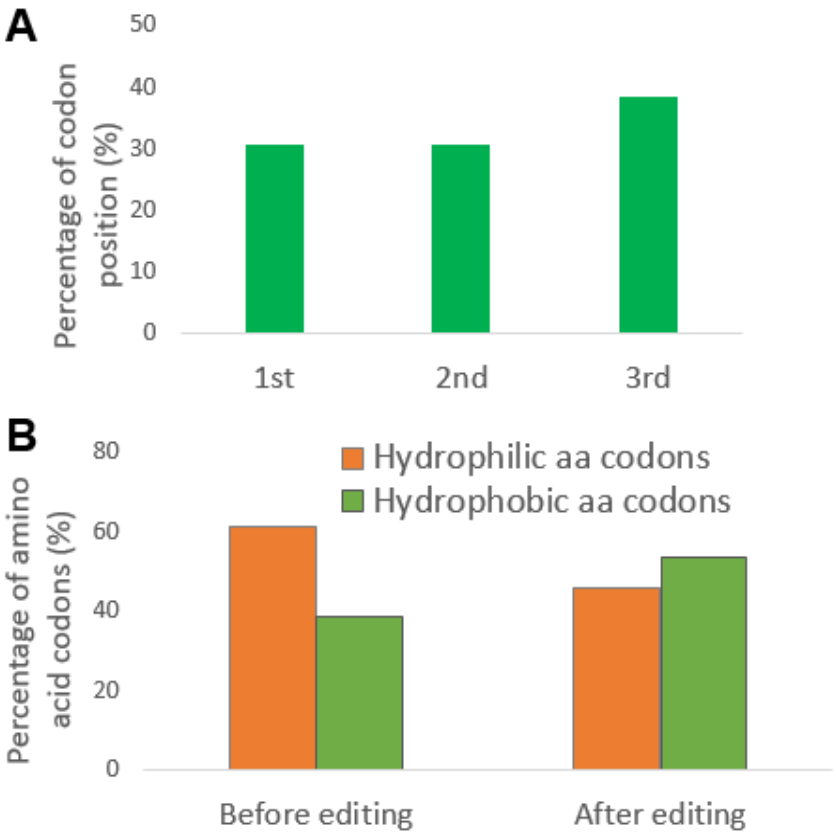


Figure 6. Schematic representation of RNA editing events in Arabidopsis. **A.** Editing percentage at different nucleotide positions of the codon. **B.** Amino acid changes caused by RNA editing.

4.4 Discussion

In this study, it is interesting that the RNA editing components could be edited. I identified nine types of RNA editing events, including all possible nucleotide substitutions, except C-to-A, C-to-G, and G-to-U, in Arabidopsis. Previously, 12 types of RNA editing events have been reported in *Salvia miltiorrhiza* and *Arabidopsis thaliana*; in Arabidopsis, these RNA editing events have been reported in nuclear transcripts using bioinformatics approaches (Meng et al. 2010; Wu et al. 2017). To my knowledge, this is the first report of tissue- and development-specific RNA editing in Arabidopsis. Although the nine types

of RNA editing events identified in this study were based on a small-scale screening of target genes (approximately 21%), these events were specific to seedling, leaf, and stem tissues of different developmental ages (Figure 1, 2 and Table 2). The tissue-specific characteristic of these RNA editing events implies that these were post-transcriptional modifications, not genomic mutations. I observed A-to-U editing events surrounding positions with a higher number of 'A' nucleotide. Previous research indicates that a *cis*-regulatory element neighboring the editing site is required for its recognition by the PPR-associated editing enzyme in plants (Shikanai 2006). Further studies are needed to better understand the processes involved in RNA editing, including the identification of *cis*-regulatory elements, isolation of editing enzymes, and validation of editing sites. I also identified RNA editing events in the UTR regions of genes. RNA editing in introns and UTR regions affects mRNA stability or splicing due to the modification of its secondary structure (Drescher et al. 2002; Jean-Claude Farré et al. 2012; Zeng, Liao, & Chang, 2007; Wu et al. 2017). RNA-Seq analysis in Arabidopsis has shown that more than 61% of multi-exon genes generate alternative mRNA isoforms, which are expressed in a tissue-specific manner (Marquez et al. 2012). However, RNA editing at some sites has an adverse effect on plant growth, development, and fertility (Hammani & Giegé, 2014). In the Arabidopsis nuclear transcripts *AT1G29930.1* and *AT1G52400.1*, C-to-U and U-to-C RNA editing has been reported at the translation borders (Meng et al. 2010). In my study U-to-C RNA editing found in 12 days old seedling (Figure 1B and Table 2). These deamination and amination reactions occur in adjacent sites; therefore, the deamination reaction is considered as the donor of amino group for the amination reaction, although the frequency of the amination reaction is higher than that of the deamination reaction (Meng et al. 2010). It is possible that another factor acts as the amino group donor for the

amination reaction. Additionally, A-to-I(G) editing events were prevalent in the present study. Such editing events have also been reported in previous studies using bioinformatics approaches. However, plant homologs of cytidine deaminase or adenosine deaminase responsible for these editing events have not yet been discovered. It is possible that the adenosine deaminase family of enzymes is derived from the cytidine deaminase family, as the only difference between the two enzymes is the deaminase groove, which facilitates the deamination reaction.

Codon bias is an important consideration for RNA editing, although targeting a nucleotide mainly depends on its neighboring nucleotides (Kuttan & Bass, 2012). In this study, I found that the nucleotide in the third position in the codon showed higher editing frequency than those in the first and second positions. I observed that, for codons in which the third nucleotide was edited, the neighboring nucleotides were GC_, AU_ (2), UU_, and AG_. In contrast to my data, He and colleagues reported a 3-fold higher frequency of RNA editing at the second nucleotide position of the codon than at the first and third nucleotide positions (He et al. 2016). This discrepancy may be due to differences in the preferences of editing enzymes for the adjacent nucleotides; in the case of adenosine deaminase, editing is greatly affected by neighboring nucleotides.

My data showed that different kinds of RNA editing events occur not only in chloroplast and mitochondrial transcripts but also in nuclear transcripts. These events were very frequent in the PPR than ZnF gene families but not in the *MORF/RIP*, *RRM*, *PP01*, and *GSDA* genes, indicating that RNA editing is gene-specific. The *PPR* genes were highly affected by RNA editing. Additionally, RNA editing events are more frequent in seedlings than in any other plant tissues. Seedlings play a very important role in plant

physiology. Early chloroplast development 1 (ECD1), which belongs to the PLS subfamily of the PPR protein family, is responsible for early chloroplast development in seedlings (Jiang et al. 2018). In this study, I observed that single nucleotide conversion adversely affected mRNA secondary structure. Therefore, it can be stated that the RNA editing in the nuclear transcript of such important family genes might have big impact on the plant physiology. Further investigation is needed to determine the algorithm via which PPR proteins recognize and bind to their cognate mRNAs and to understand the precise effect of RNA editing events on the highly programmable PPR protein family and other protein families.

4.5 References

Barkan, A., & Small, I. (2014). Pentatricopeptide Repeat Proteins in Plants. Annual Review of Plant Biology. https://doi.org/10.1146/annurev-arplant-050213-040159

Chateigner-Boutin AL, Colas Des Francs-Small C, Fujii S, et al (2013) The E domains of pentatricopeptide repeat proteins from different organelles are not functionally equivalent for RNA editing. Plant J. 74:935-945

Chen, T. C., Liu, Y. C., Wang, X., Wu, C. H., Huang, C. H., & Chang, C. C. (2017). Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in Phalaenopsis aphrodite subsp. formosana. Botanical Studies. https://doi.org/10.1186/s40529-017-0193-7

Cheng, S., Gutmann, B., Zhong, X., Ye, Y., Fisher, M. F., Bai, F., … Small, I. (2016). Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. Plant Journal, 85(4), 532–547.

https://doi.org/10.1111/tpj.13121

Delannoy, E., Stanley, W. A., Bond, C. S., & Small, I. D. (2007). Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. Biochemical Society Transactions, 35(6), 1643–1647. https://doi.org/10.1042/BST0351643

Doniwa Y, Ueda M, Ueta M, et al (2010) The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. Gene 454:39–46

Drescher, A., Hupfer, H., Nickel, C., Albertazzi, F., Hohmann, U., Herrmann, R., & Maier, R. (2002). C-to-U conversion in the intercistronic ndhI/ndhG RNA of plastids from monocot plants: Conventional editing in an unconventional small reading frame? Molecular Genetics and Genomics, 267(2), 262–269. https://doi.org/10.1007/s00438-002-0662-9

Eggington, J. M., Greene, T., & Bass, B. L. (2011). Predicting sites of ADAR editing in double-stranded RNA. Nature Communications, 2(May), 319. https://doi.org/10.1038/ncomms1324

Guillaumot, D., Lopez-Obando, M., Baudry, K., Avon, A., Rigaill, G., Falcon de Longevialle, A., … Lurin, C. (2017). Two interacting PPR proteins are major Arabidopsis editing factors in plastid and mitochondria. Proceedings of the National Academy of Sciences, 114(33), 201705780. https://doi.org/10.1073/pnas.1705780114

Hammani, K., & Giegé, P. (2014). RNA metabolism in plant mitochondria. Trends in

Plant Science, 19(6), 380–389. https://doi.org/10.1016/j.tplants.2013.12.008

He, P., Huang, S., Xiao, G., Zhang, Y., & Yu, J. (2016). Abundant RNA editing sites of chloroplast protein-coding genes in Ginkgo biloba and an evolutionary pattern analysis. BMC Plant Biology. https://doi.org/10.1186/s12870-016-0944-8

Ichinose, M., & Sugita, M. (2017). RNA editing and its molecular mechanism in plant organelles. Genes. https://doi.org/10.3390/genes8010005

Jean-Claude Farré, Cindy Aknin, Alejandro Araya, B. C. (2012). RNA Editing in Mitochondrial for Splicing Trans -Introns Is Required Jean-Claude. PLoS ONE, 7(12), 1–10. https://doi.org/10.1371/Citation

Jiang T, Zhang J, Rong L, et al. (2018) ECD1 functions as an RNA-editing trans-factor of rps14-149 in plastids and is required for early chloroplast development in seedlings. J Exp Bot 69:3037–3051

Kent, W. J. (2002). BLAT — The BLAST -Like Alignment Tool. Genome Research, 12, 656–664. https://doi.org/10.1101/gr.229202

Kuttan, A., & Bass, B. L. (2012). Mechanistic insights into editing-site specificity of ADARs. Proceedings of the National Academy of Sciences, 109, E3295–E3304. https://doi.org/10.1073/pnas.1212548109

Leu KC, Hsieh MH, Wang HJ, et al (2016) Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing. RNA Biol. 13:593–604

Lorenz R, Bernhat SH, Höner Zu, et al. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov 24;6:26

Lurin C, Andrés C, Aubourg S, et al (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089–2103

Marquez Y., Brown JW., Simpson C., Barta A & Kalyna M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome res: 2012 Jun; 22(6): 1184–1195. http://www.genome.org/cgi/doi/10.1101/gr.134106.111

Meng, Y., Chen, D., Jin, Y. F., Mao, C., Wu, P., & Chen, M. (2010). RNA editing of nuclear transcripts in Arabidopsis thaliana. BMC Genomics. https://doi.org/10.1186/1471-2164-11-S4-S12

Okuda, K., Chateigner-Boutin, A.-L., Nakamura, T., Delannoy, E., Sugita, M., Myouga, F., … Shikanai, T. (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, 146–156. https://doi.org/10.1105/tpc.108.064667

Rodrigues, N. F., Christoff, A. P., da Fonseca, G. C., Kulcheski, F. R., & Margis, R. (2017). Unveiling Chloroplast RNA Editing Events Using Next Generation Small RNA Sequencing Data. Frontiers in Plant Science. https://doi.org/10.3389/fpls.2017.01686

Schallenberg-Rüdinger M, Lenz H, Polsakiewicz M, et al (2013) A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. RNA Biol 10:1549–1556

Shikanai, T. (2006). RNA editing in plant organelles: machinery, physiological function

and evolution. Cell Mol Life Sci. 2006; 63: 698–708. doi: 10.1007/s00018-005-5449-9

Shikanai, T. (2015). RNA editing in plants: Machinery and flexibility of site recognition. Biochimica et Biophysica Acta, 1847(9), 779–785. https://doi.org/10.1016/j.bbabio.2014.12.010

Sun, T., Bentolila, S., & Hanson, M. R. (2016). The Unexpected Diversity of Plant Organelle RNA Editosomes. Trends in Plant Science, 21(11), 962–973. https://doi.org/10.1016/j.tplants.2016.07.005

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., & Brennicke, A. (2013). RNA Editing in Plants and Its Evolution. Annual Review of Genetics. https://doi.org/10.1146/annurev-genet-111212-133519

Tseng C-C, Lee C-J, Chung Y-T, et al (2013) Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant J 82:375–392

Wagoner, J. A., Sun, T., Lin, L., & Hanson, M. R. (2015). Cytidine deaminase motifs within the DYW domain of two pentatricopeptide repeat-containing proteins are required for site-specific chloroplast RNA editing. Journal of Biological Chemistry, 290(5). https://doi.org/10.1074/jbc.M114.622084

Wu, B., Chen, H., Shao, J., Zhang, H., Wu, K., & Liu, C. (2017). Identification of Symmetrical RNA Editing Events in the Mitochondria of Salvia miltiorrhiza by Strand-specific RNA Sequencing. Scientific Reports. https://doi.org/10.1038/srep42250

Yagi, Y., Nakamura, T., & Small, I. (2014). The potential for manipulating RNA with pentatricopeptide repeat proteins. Plant Journal. https://doi.org/10.1111/tpj.12377

Zeng, W. H., Liao, S. C., & Chang, C. C. (2007). Identification of RNA editing sites in chloroplast transcripts of Phalaenopsis aphrodite and comparative analysis with those of other seed plants. Plant and Cell Physiology, 48(2), 362–368. https://doi.org/10.1093/pcp/pcl058

Chapter V

Final Discussion and Future Prospective

5.1 Final discussion and future prospective

RNA editing is a post-transcriptional modification of genes whereas alternative splicing is post and co-transcriptional regulation of gene expression. Major genetic modification which regulates the protein structure and function by creating diversity in protein family is alternative pre-mRNA splicing. Splicing events in RNA creates a lot of diverse protein from a single gene. This post-transcriptional modification creates an evolutionary advantage having multiple proteins with variable functions sometimes producing opposite effect. Processes such as development of an organism and cellular differentiation involves RNA splicing. RNA editing is another mechanism where the nucleotides in RNA are edited by enzyme or editosome complexes so the translation process creates a different functional protein compared to its parent protein (Takenaka et al. 2013). In this study, I tried to find out the tissue-specific expression and alternative splicing of RNA editing related family genes (PPRs and ZnFs) and the RNA editing events occurring in those genes.

In chapter II, In plants, alternative splicing highly diversified. It greatly affects transcript diversity, in Arabidopsis even alternatively spliced single nucleotide exon also reported (Guo and Liu 2015). In this chapter, I focused on alternative splicing of *PPR* family genes. PPR family is one of the largest family gene in Arabidopsis consisting of approximately 450 genes. These genes are responsible for organeller RNA editing directly or indirectly (Okuda et al. 2009; Doniwa et al. 2010; Chateigner-Boutin et al. 2013; Leu et al. 2016). I identified five new unannotated alternatively spliced isoforms of PPR1 (1=new isoform), PPR2 (1=new isoform), PPR4 (2=new isoform) and PPR6 (1=new isoform), which were confirmed by PCR and sequencing. Among these, three contain 3' alternative-splicing

sites, one contains a 5' alternative-splicing site, and the remaining gene contains a 3'-5' alternative-splicing site. The new isoforms of two genes *PPR1* and *PPR2* affect protein, and three other alternative-splicing sites are located in 3' UTRs in PPR4 (2=new isoform) and PPR6 (1=new isoform).

I investigated the expression pattern of the alternatively spliced transcript in different tissues at different ages such as 4 days, 8 days, 12 days, 16 days, 21 days, 27 days and 32 days during the development of the Arabidopsis plant. Tissue-specific expression differs by alternative splicing of *PPR* genes. In my analysis, I found intron retention expressed well in all tissues than 3′ and 5′ alternative splice site. These findings suggest that tissue-specific expression of different alternatively spliced transcripts occurs in Arabidopsis, even at different developmental stages. For better understanding and biological impact of the PPR protein, how the editosome complex is organized and work that needed to understand. I also found TTT and ATGTTT intronic *cis*-regulatory element in *PPR1*, *PPR2* and *PPR4*, TCAAT transcriptional *cis*-regulatory elements in *PPR1* and *PPR4;* ATTGA reverse transcriptional *cis*-regulatory elements in *PPR2.*

Finally, it can be hypothesized that PPR transcripts expression is highly regulated by alternative splicing mechanism which in turn affects the physiology of Arabidopsis.

Redefining the algorithm that confirm the common binding strategy of PPR with its cogent RNA is still needed to discover. My results with newly found alternately spliced isoforms will certainly be helpful to build up the hidden algorithm of PPR protein-RNA interaction. This important finding will have a big impact on RNA research by manipulation of targeted RNA in all living organisms and also of the viruses.

In chapter III, Recently, it has been identified that other family members like MORF/RIP, ORRM and OZ are also involved in RNA editing along with *PPR* genes (Sun et al. 2016). The aim of this study is to understand the alternative splicing events and tissue specific expression of ZnF family proteins. My study reports three new unannotated isoforms of ZnF2 (2=new isoform) and ZnF4 (1=new isoform) genes of ZnF family proteins which are conformed using PCR and sequencing the cDNA. As it is reported that MORF/RIP, ORRM and OZ are working as associated RNA editing factors therefore it is very important to understand the interaction of PPR with these protein family. Better understanding of the interaction will be certainly helpful for understanding the mechanism of RNA editing in plants.

In chapter IV, In the course of study of alternative splicing, tissue-specific RNA editing events were found in PPRs and ZnFs family proteins. I collected samples of different tissues of different developmental stages from Arabidopsis. Such as seedling (whole plant) 4, 8, 12 days; 16, 21, 27 and 32 days old leaf, stem and root. I studied thirty (30) candidate genes of PPRs, ZnFs, MORFs/RIPs, ORRMs, PP01 and GSDA family proteins and mainly focused on RNA editing sites located in alternatively spliced portion. In *Arabidopsis thaliana,* 12 types of RNA editing events in nuclear transcript previously reported merely base on bioinformatics approaches (Meng et al. 2010). During my analysis, I Identified 21 different RNA editing sites in total 12 genes (including 11 PPR and 1 ZnF family proteins) which are expressed in different tissues. In my analysis, I confirmed 9 types of RNA editing events these are C-to-U, U-to-C, A-to-I(G), A-to-C, A-to-U, G-to-A, G-to-C, U-to-A and U-to-G in targeted genes. Even the U-to-C type RNA editing only confirmed in 12 days old seedling. The important thing is that, there

have no tissue and development specific RNA editing events reported yet in plants. This is the first report that RNA editing could be regulated in tissue and development specific manner. It is reported that such events in the intron and UTR regions may affect to RNA stability or splicing due to modification of secondary structure (Drescher et al. 2002; Jean-Claude Farré, Cindy Aknin, Alejandro Araya, 2012; Zeng, Liao, & Chang, 2007; Castandet B, Choury D et al. 2010; Wu et al. 2017). In my observation, I identified A-to-I(G) in *PPR2b-2720*, U-to-C in *PPR3a-2972* and U-to-G in *PPR11b-2809* RNA editing affecting secondary structure of RNA in UTR regions which may greatly affect mRNA stability. RNA editing events are very frequent in *PPR* genes than the *ZnF* family genes. It indicates that this phenomenon is gene specific. It also indicates that, PPR proteins may target various RNA substrate by altering its protein domain structures. Another thing, at the same time RNA editing events are very frequent in seedling than any other tissues. So, I explored that those kind of RNA editing events gene-specific and tissue-specific.

Deamination and amination both type of editing have been found in my study. In one hypotheses, it is stated that deamination reaction may be act as amino group donor for the amination reaction. In an another hypothesis, it is stated that the mutation of an enzyme performing the deamination or transamination type of editing. However, confirmation of the above statements is necessary. The RNA editing events are found in plant grown in laboratory environments. In the natural growing conditions how much these events are affected, these need to be investigated. Side by side in adverse climatic conditions how these editing sites behaves; answer to these query may provide important answers regarding the effect of these RNA editing events.

In summary, I successfully performed the tissue-specific expression pattern of the alternatively spliced transcript in different tissues at different developmental stages of PPR and ZnF family proteins. At the same time I also investigated the tissue-specific RNA editing events in the nuclear transcript. This is the first reports in plants. Thus I believed that my study will assist to understand the tissue-specific expression pattern of alternative splicing and the precise effect of editing events in this highly programmable PPR protein family and other family proteins in the future.

5.2 References

Castandet B, Choury D, et al (2010) Intron RNA editing is essential for splicing in plant mitochondria. Nucleic Acids Res. doi: 10.1093/nar/gkq591

Chateigner-Boutin AL, Colas Des Francs-Small C, Fujii S, et al (2013) The e domains of pentatricopeptide repeat proteins from different organelles are not functionally equivalent for RNA editing. Plant J. doi: 10.1111/tpj.12180

Doniwa Y, Ueda M, Ueta M, et al (2010) The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. Gene 454:39–46 . doi: 10.1016/j.gene.2010.01.008

Drescher, A., Hupfer, H., Nickel, C., Albertazzi, F., Hohmann, U., Herrmann, R., & Maier, R. (2002). C-to-U conversion in the intercistronic ndhI/ndhG RNA of plastids from monocot plants: Conventional editing in an unconventional small reading frame? Molecular Genetics and Genomics, 267(2), 262–269. https://doi.org/10.1007/s00438-002-0662-9

Guo L, Liu CM (2015) A single-nucleotide exon found in Arabidopsis. Sci Rep 5:1–5 . doi: 10.1038/srep18087

Jean-Claude Farré, Cindy Aknin, Alejandro Araya, B. C. (2012). RNA Editing in Mitochondrial for Splicing Trans -Introns Is Required Jean-Claude. PLoS ONE, 7(12), 1–10. https://doi.org/10.1371/Citation

Leu KC, Hsieh MH, Wang HJ, et al (2016) Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing. RNA Biol. doi: 10.1080/15476286.2016.1184384

Meng, Y., Chen, D., Jin, Y. F., Mao, C., Wu, P., & Chen, M. (2010a). RNA editing of nuclear transcripts in Arabidopsis thaliana. BMC Genomics. https://doi.org/10.1186/1471-2164-11-S4-S12

Okuda K, Chateigner-Boutin A-L, Nakamura T, et al (2009) Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. Plant Cell 21:146–156 . doi: 10.1105/tpc.108.064667

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., & Brennicke, A. (2013). RNA Editing in Plants and Its Evolution. Annual Review of Genetics. https://doi.org/10.1146/annurev-genet-111212-133519

Wu, B., Chen, H., Shao, J., Zhang, H., Wu, K., & Liu, C. (2017). Identification of Symmetrical RNA Editing Events in the Mitochondria of Salvia miltiorrhiza by Strand-specific RNA Sequencing. Scientific Reports. https://doi.org/10.1038/srep42250

Zeng, W. H., Liao, S. C., & Chang, C. C. (2007). Identification of RNA editing sites in chloroplast transcripts of Phalaenopsis aphrodite and comparative analysis with those of other seed plants. Plant and Cell Physiology, 48(2), 362–368. https://doi.org/10.1093/pcp/pcl058

5.3 List of Publications

1. Tissue specific alternative splicing of Pentatricopeptide Repeat (PPR) family genes in *Arabidopsis thaliana.* **Umme Qulsum** and Toshifumi Tsukahara (Accepted).
2. Analysis of tissue-specific RNA editing events in *Arabidopsis thaliana.* **Umme Qulsum** and Toshifumi Tsukahara (Submitted).
3. Identification of tissue-specific alternative splicing of Zinc Finger (*ZnF*) family genes in *Arabidopsis thaliana*. **Umme Qulsum** and Toshifumi Tsukahara (To be submitted).
4. Md.Thoufic Anam Azad, **Umme Qulsum** and Toshifumi Tsukahara. Comparative activity of adenosine deaminase acting on RNA (ADARs) isoforms for correction of genetic code in gene therapy (Accepted).

5.4 Conference Presentation

1. Alternative Splicing of RNA Editing Pentatricopeptide Repeat (PPR) Family Gene in *Arabidopsis thaliana.* **Umme Qulsum**, John Munene, Toshifumi Tsukahara. The RNA Society of Japan, Annual meeting 2017, Toyama, japan.
2. Tissue-specific Alternative Splicing of RNA Editing Related Family Genes in *Arabidopsis thaliana.* **Umme Qulsum**, Toshifumi Tsukahara. Molecular Biology Society of Japan, Annual meeting 2017, Kobe, Japan.
3. Analysis of tissue-specific alternative splicing of P-subfamily PPR genes in *Arabidopsis thaliana.* **Umme Qulsum**, Toshifumi Tsukahara. JAIST world conference, JWC 2018, Ishikawa, Japan.
4. Tissue-specific site-directed RNA editing events in *Arabidopsis thaliana.* **Umme Qulsum**, Toshifumi Tsukahara. JAIST-INDIA Symposium on Materials Science 2018, Ishikawa, Japan.
5. Investigation of tissue-specific alternative splicing of RNA editing related family genes in *Arabidopsis thaliana.* **Umme Qulsum**, Toshifumi Tsukahara. 8th World Congress on Plant Genomics and Plant Science, Osaka Japan.

ACKNOWLEDGEMENTS

First of all, I would like to express my great gratitude and thanks to my honourable supervisor Professor Dr. Toshifumi Tasukahara for his continuous support with great patience and wisdom throughout my doctoral study. He supervise my study with great care through the whole process with creative ideas during my PhD. He also financially helps me providing me TA, RA and Researcher position at my doctoral study.

I greatly acknowledge the support and scholastic advices of my second supervisor Associate Professor Takumi Yamaguchi and advisor for minor research Professor Takahiro Hohsaka for their great support and constructive advices.

I have to acknowledge the department of Botany, University of Rajshahi, Bangladesh for granting me study leave and understanding during my stay and study in Japan.

I would like to thanks all past members of Professor Tsukahara Lab. Specially Assistant professor Dr. Md. Thoufic Anam Azad, Professor Dr. AHM Khurshid Alam, Dr. Shafiul Alam, Dr. Vu Thi Luyen, Mr. Shirakawa Yoshiaki, Mr. Iwata and present member Dr. Sonali Bhakta, Mr. Saifullah, Mr. John Munene, Dr. Sakari Matomo, Ms. Saitaka Mamikio, Mr. Jiarui Li, Mr. Tanaka Ryuta for their support, advices and ultimate tolerance. I would like to thanks to Professor Hohsaka Lab and Associate Professor Yamaguchi Lab. members as well.

I would like to thank to all my colleagues and friends mainly Mr. Ehsanul Haque, Mr. Earul Islam, Mr. Md Mosaddik Hasan, Dr. S. M. Nizamuddin, Mr. Eamin Ali Akanda, Dr. Abdus Sattar, Dr. Muhammad Samir Ullah, Mr. Mohammed Khaled Alom, Dr.Tofiq Ahmed, Dr. Redwan Newaz, Dr. Md Anwarul Islam, Mr. Md Abul Kalam Siddiki, Dr. Rashed Mozumder, Dr. Abu Rashed, Dr. H. M. Belal, Mr. Farhan Ferdows, Mr. Mohammad Mesbahul Karim, Ms. Farha Diba Trisha, Dr. Shahiduzzaman Sohel, Ms. Sanchita Afroz, Dr. Reza Kaher, Dr. Asif Ali, Mr. Mohammad Nur Hasan, Mr. Amr Ashmawy, Mr. Md. Arif Hossain, Mr. Shen Qian, Ms. Savvy, Ms. Gargi Joshi, Dr. Trinh Xuan Dai, Mr. Azadul Kabir, Mr. Faiz Al Faisal, Dr. SM Ahasanul Hamid for their motivation and company during my stay in JAIST.

Last but not least, I am really grateful to my beloved husband, daughter, son and family. I am thankful to my husband for his inspiration and continuous support. I would not be able to compensate his suffering and sacrifice during my research work.

Qulsum Umme
JAIST, Ishikawa, Japan
04/01/2018