Doctoral Dissertation

Identification and analysis of the U-to-C RNA editing enzyme in plants for restoration of genetic disorders

Ruchika

Supervisor: Professor Dr. Toshifumi Tsukahara

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Materials Science
(March 2022)

Abstract

Identification and analysis of the U-to-C RNA editing enzyme in plants for restoration of genetic disorders

Ruchika,
Tsukahara Laboratory, JAIST

RNA editing is a post-transcriptional molecular process through which cells can be describe as site-specific changes in the nucleotide sequences within an RNA molecule that has been transcribed by RNA polymerase enzymes. Among various types of RNA editing, Adenine-to-Inosine, A-to-I(G), Cytidine-to-Uridine, C-to-U and Uridine-to-Cytidine, U-to-C are the most common. Studies have already revealed the mechanisms for A-to-I and C-to-U RNA editing which are generally caused by ADARs and APOBEC-AID deaminase family, respectively. However, the transaminase enzymes responsible for "Reverse" U-to-C editing is not yet discovered, although it is the abundant phenomenon in lower plant species like hornworts and rare in animals. Here I examined the U-to-C RNA editing in *Arabidopsis* tissues at different developmental stages of growth. In this study, the high-throughput RNA sequencing (RNA-seq) of 12-day-old and 20-day-old *Arabidopsis* seedlings was performed. The results showed that DEGs were expressed to higher levels in 12-day-old seedlings than in 20-day-old seedlings. Additionally, pentatricopeptide repeat (PPR) genes were also expressed at higher levels as indicated by the log2FC values. RNA-seq analysis of 12-day and 20-day-old Arabidopsis seedlings revealed candidates of U-to-C RNA editing events. Sanger sequencing of both gDNA and cDNA confirmed the seven U-to-C RNA editing sites.

Further, I investigated the U-to-C RNA editing-related genes in *Arabidopsis* tissues and the effects on mRNA stability, with a special focus on PPR proteins. Here, I have demonstrated the effects of the "reverse" RNA editing on the mRNA stability for all seven edited genes. In addition, the nuclear PPR gene (AT2G19280) in 12-day-old seedlings of *Arabidopsis thaliana* was also identified with U-to-C RNA editing. The U-to-C RNA editing sites were found in the untranslated region (3' UTR) of the mature mRNA and affect its secondary structure. Then, the correlation between U-to-C RNA editing-related genes and their mRNA abundance were examined. Furthermore, I investigated the effects of U-to-C RNA editing in Arabidopsis using the transcription inhibitor actinomycin D (Act D). The addition of Act D to the cell suspension culture of transgenic Arabidopsis generated by Agrobacterium-mediated transformation showed that single nucleotide base conversion adversely affected the mRNA secondary structure and stability.

Pentatricopeptide repeats (PPR) proteins are act as sequence-specific RNA-binding proteins within mitochondria and chloroplasts in almost all land plants. Genome-wide analysis of the hornworts, *Anthoceros agrestis*, revealed the PPR proteins in this species contain unique C-terminal DYW-like domains with specific signatures. In present work, I have explored the study on three different variants of C-terminal PPR proteins of hornworts, GRP-type, DRH- type and DYW-types. I investigated the RNA editing events by cloning the Hornworts PPR genes. A bacterial expression system was developed in which the Hornworts specific PPR protein variants were cloned with PPR56 (truncated DYW), *Physcomitrella patens* (moss) editing factor. We measured the gene expression levels of DYW variants. In addition, we demonstrated the functional homology of DYW domains with APOBEC1in human cells.

The MS2 system were initially developed with ADAR1 and APOBEC1 for A-to-I(G) and C to-U RNA editing, respectively. In this study, an expression system was designed with PPR56, *Physcomitrella patens* (moss) editing factor. We engineered an artificial RNA editing mechanism by combining the deaminase domain of plant derived PPR56 with a guideRNA (gRNA) which is complementary to target mRNA. As a target RNA, we used RNA encoding blue fluorescent protein (BFP) which was derived from the gene encoding GFP by a single T-to-C substitution. Upon transient expression of both components (PPR56 and gRNA), we confirmed the restoration of sequence of GFP revealing an editing efficiency of up to 85-100%, while previous developed system with APOBEC1 only showed about 20 % editing efficiency. Furthermore, we identified that the C-terminal E2-DYW domain of PPR56 is sufficient for C-to-U conversion in the MS2 system.

Lastly, I introduced a non-sequencing approach for the rapid detection of RNA editing using a portable micro-Temperature gradient Gel Electrophoresis (μTGGE). This is based on the principle of electrophoresis, which use temperature to denature the samples as it moves across the polyacrylamide gel. In this method, a fragment of double-

stranded DNA when heated, forming a gradient of double-stranded DNA to partially separated strands to completely separated single-stranded DNA. A sample of RNA editing with different nucleotide bases show the different melting profiles because of their different melting profile. Here, we demonstrated the difference between the melting profiles for edited and non-edited (wild type) RNAs. Reproducibility was evaluated from measuring the pattern similarity scores (PaSS) between the band patterns obtained with the edited and non-edited RNAs. This tool is providing a simple, and cost-effective method for detecting even a single base modification in RNAs. We expect that our rapid analyzing tool will foster further discoveries in this rapidly expanding field of molecular biology.

Keywords: U-to-C RNA editing, *Arabidopsis thaliana*, PPR56, DYW domains, hornworts, RNA-seq

Doctoral Dissertation

Identification and analysis of the U-to-C RNA editing enzyme in plants for restoration of genetic disorders

Supervisors

Supervisor	Dr. Toshifumi Tsukahara
Second Supervisor	Dr. Takumi Yamaguchi
Minor Research Supervisor	Dr. HidekazuTsustui (inside JAIST) Dr. Mizuki Takenaka (outside JAIST)

Members of Review committee

Head of the committee	Prof. Dr. Toshifumi Tsukahara
Internal reviewer	Prof. Dr. Yuzuru Takamura
Internal reviewer	Dr. Hidekazu Tsustui
Internal reviewer	Prof. Takahiro Hohsaka
External reviewer	Dr. Prof. Dr. Ito Masaki

Japan Advanced Institute of Science and Technology
RUCHIKA
(s1920043)

Table of Content

Chapter 1

General Introduction

1.1. Research interest

RNA molecules regulate and participate in a vast array of cellular processes, and the scientific community is now well into a new era in which some aspect of RNA biology- as a tool, a therapeutic, a diagnostic, or a central player in fundamental biological process- is becoming increasingly important. Any mutation in the gene may not be able to successfully produce the protein, which may also cause various diseases. Thus, I challenged this task and working on the development of new genetic disease treatment methods. I attempted to generate a new generation of genetic treatments, which directly works on the genes of the patient and repairs the mutated genes.

In this way, I would like to utilize my knowledge of RNA research to work on next-generation disease treatment that acts directly on RNA to improve the function of RNA.

Among all types of RNA editing, A-to-I, C-to-U and U-to-C are the most common. Researchers are already familiar with the mechanisms for A-to-I and C-to-U RNA editing which are generally caused by ADARs and APOBEC-AID deaminase family, respectively. However, the transaminase enzymes responsible for "Reverse" U-to-C editing is not discovered yet although it is the abundant phenomenon in lower plant species and rare in animals. In this work, for the first time I reported the occurrence of U-to-C RNA editing in plants. further, I was interested in identifying the enzymes responsible for such type of RNA editing. I studied the RNA editing events from four different sources, such as Arabidopsis, hornwort, Bacteria and human cells (Figure1.1). I believe, this study helped me understanding the detailed analysis of the U-to-C RNA editing events that could be use in future for restoration of genetic disorders.

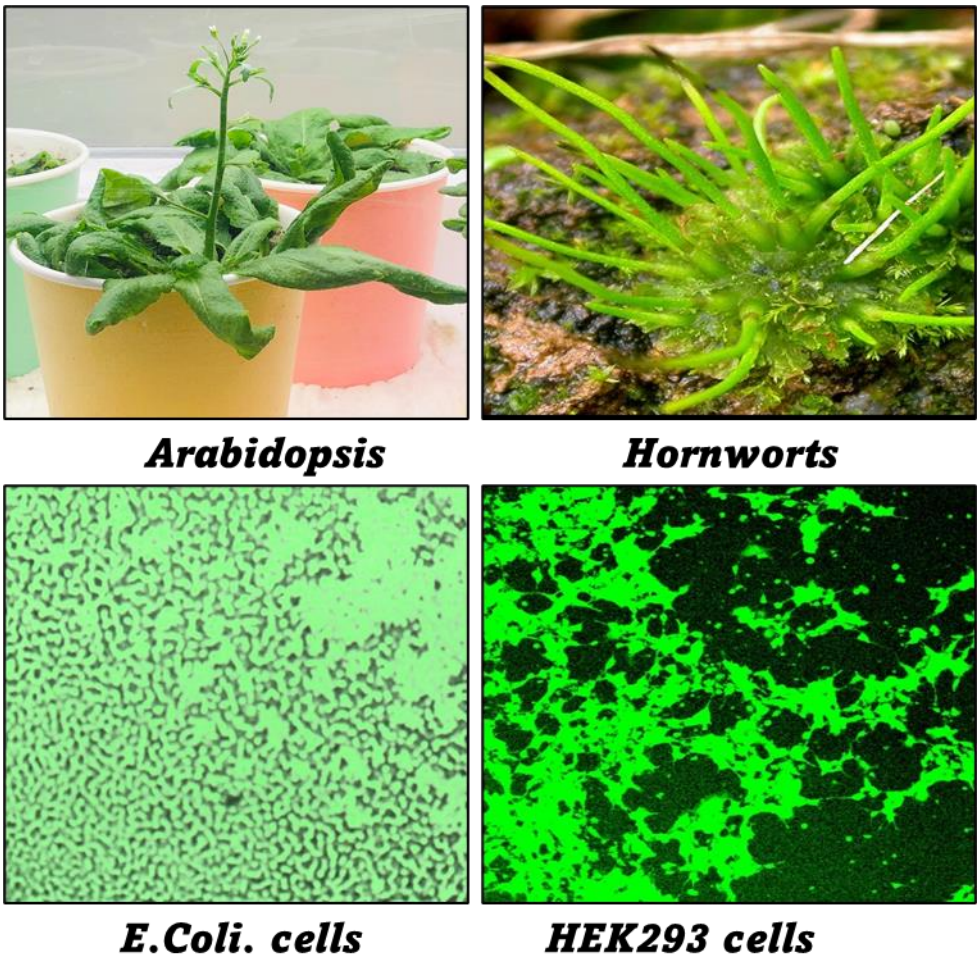


Figure 1.1. Samples used to study the RNA editing events.

1.2. Research Background

1.2.1. Central dogma

In molecular biology, the basic principle of Central dogma is an inheritance mechanism that encodes the genetic information present in DNA and transfers them to RNA and protein[1]. The DNA in genomes does not directly synthesis the protein itself, but instead use RNA molecules as intermediary. When the cell needs to synthesis a particular protein, the nucleotide sequence of immensely long DNA in a chromosome is first copied into RNA (called *transcription*). These

RNA copies of sequence are used directly as a template for proteins (a process called *translation*)[2]. The schematic diagram for central dogma is shown in figure 1.2.

One serious challenge against the principle of central dogma is the discovery for RNA editing. Genetic information's not coded in the genomic template may be transferred into the mRNA template after being transcribed.[3–5]

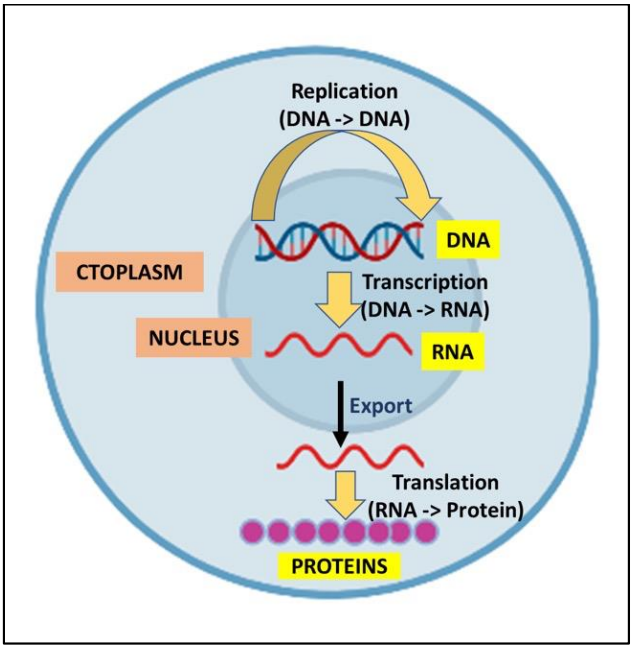


Figure 1.2. **Schematic representation of central dogma**. The flow of genetic information from DNA to RNA (transcription) and from RNA to protein (translation) occurs in all living cells.

1.2.2. RNA editing and types

RNA editing is described as a post-transcriptional modification that triggers any site-specific changes in RNA nucleotide base sequences including insertion, deletions, and nucleotide base conversion[6]. RNA editing was first defined as the modification of RNA ultimately changing its coding capacity [7]. The term "RNA editing" was first coined in 1986 when it was found in *Trypanosoma* mitochondrial cytochrome oxidase (*cox*) subunit II gene, where four nucleotides were added in the RNA which was not integrated in the genomic DNA[8] . The report on mammalian

RNA editing was reported in 1987 [9]. From that time, this term has been used for insertion[8] , deletion or substitution[10] of nucleotides. But recent scientific advancement and application of next-generation sequencing technology have changed the understanding and effect regarding RNA editing. In higher eukaryotes insertion and deletion has not been reported yet but the most widespread phenomenon is base conversion.[7]

RNA editing is one of the most important epigenetic processes by which posttranscriptional modification of gene occurs. This phenomenon results in diversity of transcript by insertion, deletion or conversion of nucleotides ultimately leading to protein diversity. Previously it was assumed that one gene is responsible for one protein but by RNA editing along with another mechanism like splicing, different functional proteins can be produced from the same gene. Therefore, RNA editing is very important for differentiation, growth, and development in life.

RNA editing in plants

- **C-to-U editing**

RNA editing (also **RNA modification**) is a post-transcriptional molecular process through which cells can be describe as site-specific changes in the nucleotide sequences within an RNA molecule that has been transcribed by RNA polymerase enzymes[11]. They may include insertion, deletion or base conversion of nucleotides and has been widely investigated in animals. [12–15] In plants, the post transcriptional modification of RNA editing is specified by the deamination of the C-to-U base substitution which is extensively expressed in cell organelles, the mitochondria and chloroplast and, many nuclear transcripts were also differentially edited[15]. It was first discovered in 1986 where the uridine bases were inserted at different specific sites to restore or change the protein coding sequence in the mitochondrial (kinetoplast) cytochrome *c oxidase II* (*coxII*)

transcripts of *Trypanosoma brucei* to restore the function.[8] In plant, RNA editing was first described in 1989 as C-to-U conversion in mitochondrial transcripts.[16–18] The greater part of the C-to-U substitution is chiefly in the protein-coding regions which safeguard the evolutionary codons. Later, it was found that the C-to-U substitution is the most observed type [19] and has been identified in the mitochondria organelles of diverse plants such as *Arabidopsis thaliana*, *Oryza sativa, Brassica napus Beta vulgaris, Vitis vinifera,* and *Nicotiana tabacum*.[20,21]

- U-to-C editing

In case of some lower plant families, extensive RNA editing of U-to-C in addition to the "reverse" C-to-U has also been identified in plants organelles (plastid and mitochondria).[22,23] Therefore, RNA editing should be treated in as a corrective mechanism at the post-transcriptional level for C-to-T (or T-to-C) changes, which possibly act as support to less favored changes [13]. However, in *Arabidopsis thaliana* computational evidence of A-to-I RNA editing events in nucleus transcriptome suggested to be identified as homolog with ADAR. In plants, the RNA editing is carried out usually in mitochondria and chloroplast as well as nuclear genes. In recent study they reported that some peculiar editing events were also found in plants. In total 12 different types of RNA editing events in *Arabidopsis* and *Salvia;* these events all are intra base substitutions. [24]Also, rare occurrence of U-to-C RNA editing in the nuclear genes was reported in *Arabidopsis*. [25,26]

1.2.3. PPR proteins for RNA editing

In 2017, it was stated that largest plant-specific family of pentatricopeptide repeat (PPR) proteins may have been involved in RNA editing [27]. PPR proteins are described by the presence of

signature tandem repeats of a 35-amino acid in PPR motif [28]. They contain the RNA recognition code, which are used for identification or prediction of different and distinct RNA editing site. In Arabidopsis, large family members of PPR have more than 500 members of which approximately 193 members has an E domain extended, while 87 others have both E and DYW domains in them. Whereas remaining members belongs to P-subfamily 19. In some case of PPR proteins there is an additional C-terminal DYW domain considerably 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 [29]. Furthermore, DYW domain of PPR family proteins are known to have a conserved residues whose sequence is like cytidine deaminases which is an editing enzyme. Therefore, DYW domains is considered to have an active site for C to U editing [30]. Also, it has been shown that the E domain has an important role in RNA editing, especially the PG region. 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 20,21. The detailed motif structure of PPR proteins is shown in Figure 1.3.

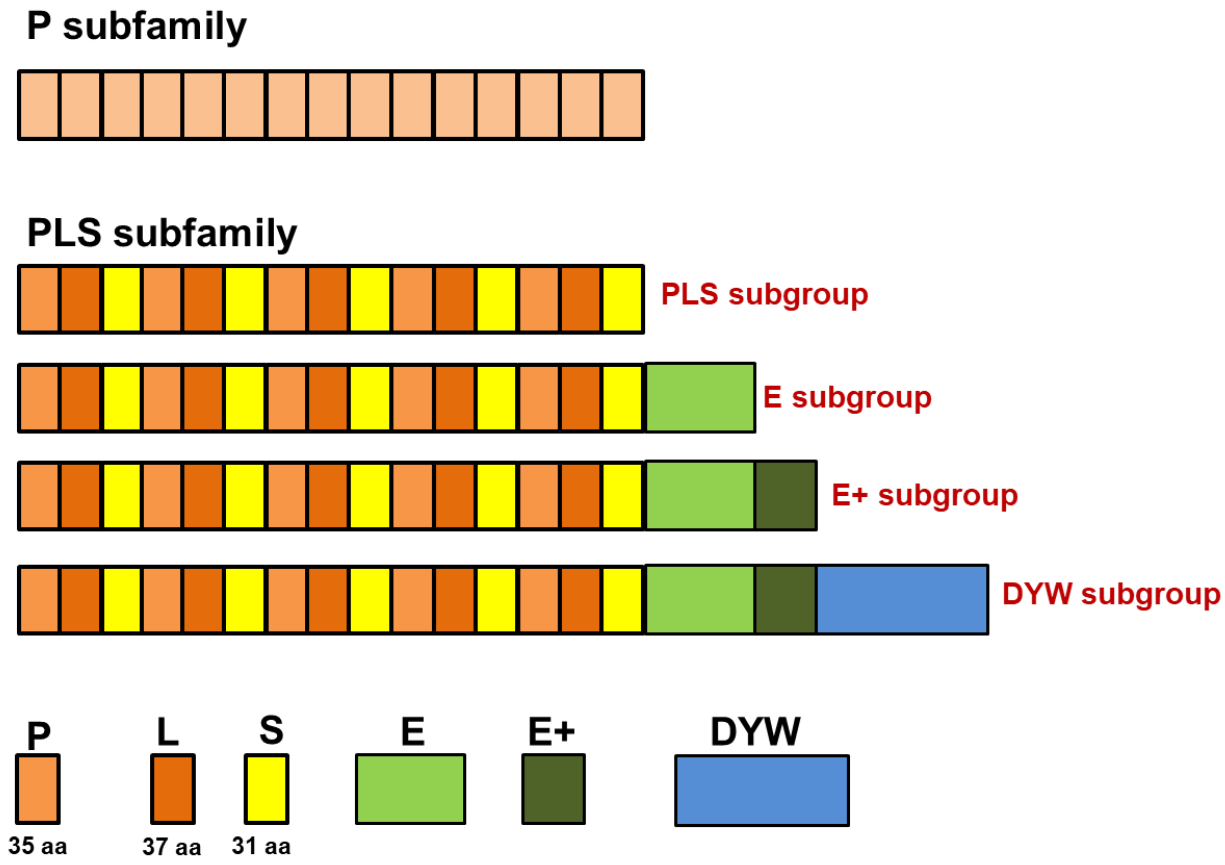


Figure 1.3. Diagrammatic representation of Motif Structure of Arabidopsis PPR Proteins

1.2.4. RNA editing in animals

RNA editing based on MS2 system

MS2 tagging is a technique which is based on the three-hybrid system between the MS2 bacteriophage coat protein and a stem-loop structure from the phage genome [33,34]. This technique is used for biochemical purification of RNA-protein complexes and partnered to GFP expression for the detection of target RNA in the living cells [35]. The MS2 coat protein system and the lambda N—B box system, two tethering systems mainly used in eukaryotic cells. They have similar characteristics. The coat protein will bind to the RNA with stem-loop structure and specific sequence. The molecular weight of λN protein is 12.2 kDa and MS2 coat protein is 13.7 kDa. In terms of binding affinity, λN protein is ~10^{-8}M, and MS2 coat protein is ~10^{-9}M when AUUA used or in case of AUCA even 10^{-10}M [36]. In the previous studies, MS2 had been used for RNA editing by link with ADAR1 or APOBEC1 [37,38]. It shows that the MS2 system is an effective artificial RNA editing system.

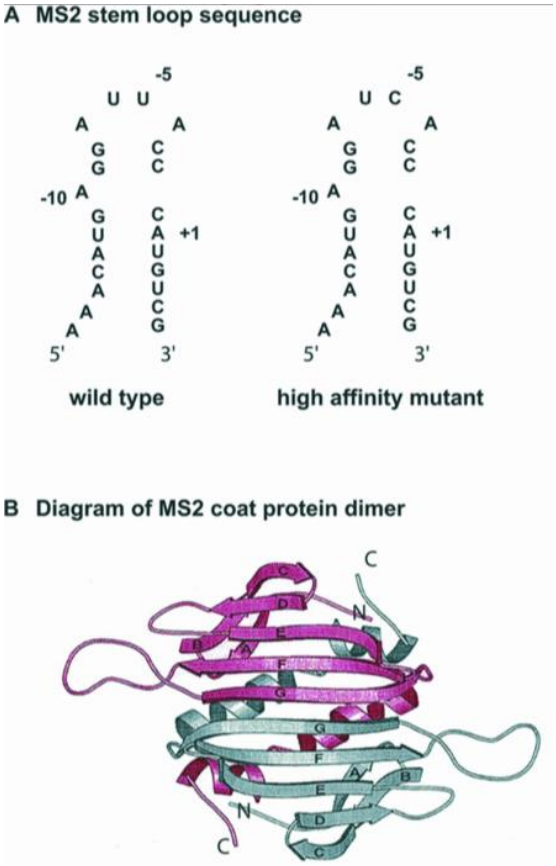


Figure 1.4. MS2 RNA stem-loop and MS2 coat protein. (A) The wild-type (left) and high-affinity C-loop (right) sequence of the MS2 stem-loop. (B) Diagram shows the association between two MS2 coat protein monomers (red and gray). [39]

1.2.6. Site-directed RNA editing

Recently, the methods for site directed RNA editing are developed using engineered deaminases combined with short guide RNAs to recode single A residues at specific sites in any user defined transcript. [37,38,40–42]

Based on the phenomenon of MS2 tagging, which is based on natural interactions between the MS2 bacteriophage coat protein and a stem-loop structure from the phage genome [33], the ADAR1 along with the Guide RNA as coat protein is attached with the ADAR1 construct and Stem loop with the Guide RNA construct, was reported to restore the genetic code mutation (Figure 1.5A). [38]

Also, APOBEC 1 is incorporated to the MS2 coat protein, and the gRNA bound to MS2 stem loop with a view to initiate the binding of the coat protein and stem loop thus allowing the gRNA to guide the deaminase for site-specific RNA editing at the targeted nucleotide sequence (figure 1.5 B and D). [37].

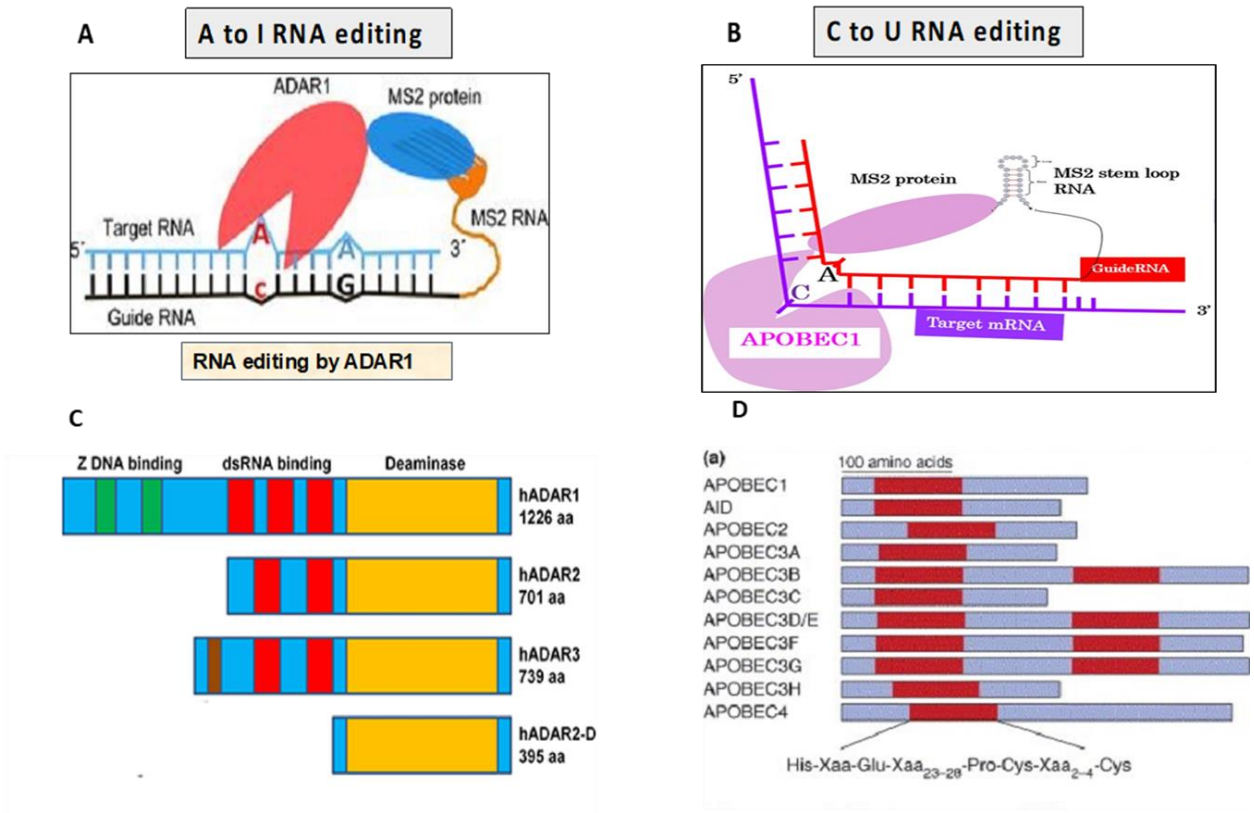


Figure 1.5. Artificial RNA editing based on MS2 system. (A) mechanism for A-to I RNA editing by ADAR1, (B) mechanism for C-to-U RNA editing by APOBEC1, (C) Different types of the ADAR deaminase (D)Members of APOBEC family.

1.3. Objectives of this study

Tsukahara Laboratory in JAIST mainly focusing on development of an artificial RNA editing system to restore the genetic disorder. The team has successfully developed an artificial RNA editing system using deaminase domain of adenosine deaminase acting on RNA (ADAR1) and artificial deaminase (APOBEC1), in combination with MS2 system to target specific A-to-I (G) and C-to-U mutation, respectively at the mRNA level. 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. My major research interest is in developing an amination enzymatic system for "reverse" U-to-C RNA editing.

Considering the above purpose, the objectives of the present study are-

1. Identification and analysis of U to C RNA editing sites in the nuclear genes of *Arabidopsis thaliana*.
2. To study the effects of U to C RNA editing on mRNA stability in *Arabidopsis thaliana.*
3. To identify the candidate DYW domains variants related to U to C RNA editing in hornworts and investigating the RNA editing in *E. coli*.
4. Development of MS2 system with plant derived "DYW" type PPR protein
5. In addition, I developed a non-sequencing approach for the rapid detection of RNA editing.

1.4. Reference

1. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, (1970).
2. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology* **15**, (2017).
3. Blanc, V., Henderson, J. O., Kennedy, S. & Davidson, N. O. Mutagenesis of Apobec-1 Complementation Factor Reveals Distinct Domains That Modulate RNA Binding, Protein-Protein Interaction with Apobec-1, and Complementation of C to U RNA-editing Activity. *Journal of Biological Chemistry* **276**, (2001).
4. Blanc, V., Henderson, J. O., Kennedy, S. & Davidson, N. O. Mutagenesis of Apobec-1 Complementation Factor Reveals Distinct Domains That Modulate RNA Binding, Protein-Protein Interaction with Apobec-1, and Complementation of C to U RNA-editing Activity. *Journal of Biological Chemistry* **276**, (2001).
5. MEHTA, A. & DRISCOLL, D. M. Identification of domains in apobec-1 complementation factor required for RNA binding and apolipoprotein-B mRNA editing. *RNA* **8**, (2002).
6. Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. & Brennicke, A. RNA Editing in Plants and Its Evolution. *Annual Review of Genetics* **47**, (2013).
7. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Reviews Genetics* **2**, (2001).
8. Benne, R. RNA-editing in trypanosome mitochondria. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1007**, (1989).
9. Powell, L. M. *et al.* A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, (1987).

10. Kim, H. & Kim, J.-S. A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics* **15**, (2014).

11. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Reviews Genetics* **2**, (2001).

12. Gott, J. M. & Emeson, R. B. FUNCTIONS AND MECHANISMS OF RNA EDITING. *Annual Review of Genetics* **34**, (2000).

13. Maier, R. M., Neckermann, K., Igloi, G. L. & Kössel, H. Complete Sequence of the Maize Chloroplast Genome: Gene Content, Hotspots of Divergence and Fine Tuning of Genetic Information by Transcript Editing. *Journal of Molecular Biology* **251**, (1995).

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

15. Meng, Y. *et al.* RNA editing of nuclear transcripts in Arabidopsis thaliana. *BMC Genomics* **11**, (2010).

16. Covello, P. S. & Gray, M. W. RNA editing in plant mitochondria. *Nature* **341**, (1989).

17. Gualberto, J. M., Lamattina, L., Bonnard, G., Weil, J.-H. & Grienenberger, J.-M. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature* **341**, (1989).

18. Hiesel, R., Wissinger, B., Schuster, W. & Brennicke, A. RNA editing in plant mitochondria. *Science* **246**, (1989).

19. Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. & Brennicke, A. RNA Editing in Plants and Its Evolution. *Annual Review of Genetics* **47**, (2013).

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

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

22. Knie, N., Grewe, F., Fischer, S. & Knoop, V. 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* **16**, (2016).

23. Yoshinaga, K. Extensive RNA editing of U to C in addition to C to U substitution in the rbcL transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucleic Acids Research* **24**, (1996).

24. Meng, Y. *et al.* RNA editing of nuclear transcripts in Arabidopsis thaliana. *BMC Genomics* **11**, (2010).

25. Ruchika, Okudaira, C., Sakari, M. & Tsukahara, T. Genome-Wide Identification of U-To-C RNA Editing Events for Nuclear Genes in Arabidopsis thaliana. *Cells* **10**, (2021).

26. Ruchika & Tsukahara, T. The U-to-C RNA editing affects the mRNA stability of nuclear genes in Arabidopsis thaliana. *Biochemical and Biophysical Research Communications* **571**, (2021).

27. Andrés-Colás, N. *et al.* Multiple PPR protein interactions are involved in the RNA editing system in Arabidopsis mitochondria and plastids. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 8883–8888 (2017).

28. Lurin, C. *et al.* Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis[W]. *The Plant Cell* **16**, (2004).

29. Salone, V. *et al.* A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Letters* **581**, (2007).

30. Shikanai, T. RNA editing in plants: Machinery and flexibility of site recognition. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1847**, (2015).

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

32. Leu, K. C., Hsieh, M. H., Wang, H. J., Hsieh, H. L. & Jauh, G. Y. Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing. *RNA Biology* (2016) doi:10.1080/15476286.2016.1184384.

33. Johansson, H. E., Liljas, L. & Uhlenbeck, O. C. RNA Recognition by the MS2 Phage Coat Protein. *Seminars in Virology* **8**, (1997).

34. SenGupta, D. J. *et al.* A three-hybrid system to detect RNA-protein interactions in vivo. *Proceedings of the National Academy of Sciences* **93**, (1996).

35. Bertrand, E. *et al.* Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell* **2**, (1998).

36. Montiel-Gonzalez, M. F., Diaz Quiroz, J. F. & Rosenthal, J. J. C. Current strategies for Site-Directed RNA Editing using ADARs. *Methods* **156**, (2019).

37. Bhakta, S., Sakari, M. & Tsukahara, T. RNA editing of BFP, a point mutant of GFP, using artificial APOBEC1 deaminase to restore the genetic code. *Scientific Reports* **10**, (2020).

38. Azad, M. T. A., Bhakta, S. & Tsukahara, T. Site-directed RNA editing by adenosine deaminase acting on RNA for correction of the genetic code in gene therapy. *Gene Therapy* **24**, (2017).

39. Keryer-Bibens, C., Barreau, C. & Osborne, H. B. Tethering of proteins to RNAs by bacteriophage proteins. *Biology of the Cell* **100**, (2008).

40. Azad, Md. T. A., Qulsum, U. & Tsukahara, T. Comparative Activity of Adenosine Deaminase Acting on RNA (ADARs) Isoforms for Correction of Genetic Code in Gene Therapy. *Current Gene Therapy* **19**, (2019).

41. Montiel-Gonzalez, M. F., Vallecillo-Viejo, I., Yudowski, G. A. & Rosenthal, J. J. C. Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed RNA editing. *Proceedings of the National Academy of Sciences* **110**, (2013).

42. Stafforst, T. & Schneider, M. F. An RNA-Deaminase Conjugate Selectively Repairs Point Mutations. *Angewandte Chemie International Edition* **51**, (2012).

Chapter 2

Genome-wide identification of U-to-C RNA editing events for nuclear genes in Arabidopsis thaliana

2.1. Introduction

RNA editing, one of the most promising means of post-transcriptional gene regulation, has been widely investigated in various protozoa [1], mammalian apolipoprotein-B [2], animals [3], fungi [4], bacteria [5,6], and viruses [7,8] as well as in plants [9–11]. A-to-I (Inosine) RNA editing is observed in animal nuclear genes, while C-to-U RNA editing is not limited to animals but is also spreading in plant organelles. The mechanism of cytidine-to-uridine (C-to-U) RNA editing in plant organelles is completely different from that in animal nucleus but also reasonably well understood, mainly owing to the characterization of many RNA editing factors in model systems such as Arabidopsis thaliana and *Physcomitrella patens* [12]. In flowering plants, the RNA editing machinery, collectively described as the editosome, consists of at least four proteins including pentatricopeptide repeat (PPR) protein, Multiple Organellar RNA editing factor (MORF)/RNA editing factor interacting protein (RIP), Organelle RNA Recognition Motif (ORRM) proteins, and organelle zinc-finger protein (OZ1).

PPR proteins constitute a large family of proteins, with more than 400 members [13] and are either directly or indirectly responsible for RNA editing [14–16]. Direct selection of target sites is governed by PLS subclass PPR proteins with additional E1 and E2 domains only or further C-terminal DYW domain, which is most likely to catalyze C to U deamination.

In addition to PPR proteins, MORF/RIP, ORRM, and OZ proteins are also required for successful RNA editing and play an important role in the editosome [17]. In the morf1 loss-of-function mutant, a single amino acid substitution in the conserved MORF domain abrogates the interaction of MORF1 with many PLS-class PPR proteins, im-plying that direct interaction with PPR proteins is required for the RNA editing function of MORF1 [18]. In P. patens, the upstream PPR stretch for RNA recognition linked in cis to the downstream E1, E2, and DYW domains is evident in all

editing factors. Because of the simplicity of this model system, all organelle editing sites in the moss have been assigned to their corresponding DYW-type editing factors (Schallenberg-Rüdinger & Knoop, 2016). Reconstitution of target site-specific C-to-U RNA editing in E. coli cell as well as in vitro with a single DYW-type RNA editing factor from Physcomitrella patens suggests the DYW domain catalyzes the cytidine deamination.

While C-to-U RNA editing occurs in chloroplasts and mitochondria of the majority of terrestrial plants, U-to-C RNA editing is rare in land plants, except in hornworts, lycophytes, and ferns, and is, therefore, referred to as an occasional phenomenon [20]. Because of its rare occurrence only in non-model plants, negligible research has been done to explore the mechanism involved in U-to-C RNA editing. Recent finding of novel types of DYW domain, which are limited to species having U-to-C editing, im-plies that the domains are somehow involved in amination of uridines in plant organelles [21,22]. Therefore, this study was more centered toward the expressed PPR genes. PPR proteins are involved in RNA editing of organellar transcripts. However, their expression and functional role as the editing factors at the nuclear level further need to be uncovered.

In contrast to organellar RNA editing, RNA editing in nuclear genes of plants has not been widely accepted, though it has been suggested by few analyses based on the RNA-seq data. Recently, we also reported U-to-C and A-to-guanosine or inosine (G or I) nucleotide conversions in 12-d-old whole seedlings and leaves of 21-d-old seedlings, respectively [23,24]. However, direct comparison of DNA and cDNA sequencing from the same plant material, which is indispensable to eliminate the possibility of DNA mutations or sequencing errors, was not conducted.

Here, we examined the U-to-C RNA editing in 12-day- and 20-day-old seedlings of Arabidopsis thaliana, which showed distinct RNA editing status at least at a single site in previous analysis. RNA-seq data can be used for sequence differences relative to the reference genome to identify

both genomic SNPs and RNA editing events. The major challenge in identifying U-to-C RNA editing sites using RNA-seq data is the discrimination of RNA editing sites from genome-encoded, single-nucleotide polymorphisms and technical artifacts caused by sequencing or read-mapping errors. We comprehensively analyzed all candidates for U-to-C RNA editing by Sanger sequencing and con-firmed the presence of genuine U-to-C RNA editing events in Arabidopsis nuclear genes.

2.2. Materials and Methods

2.2.1. Plant Growth Conditions and Sample Collection

Seeds of *Arabidopsis thaliana* col-0 were soaked in water and incubated in the dark at 4 °C for 2–3 days. Seeds were sown in paper pots containing a 1:2:1 mixture of horticultural perlite, peat moss, and vermiculite, and covered with a plastic wrap to maintain the moisture content. The pots were placed in a U-ING Green Farm hydroponic grow box (Japan Trend shop, Osaka, Japan) in a growth room at 22 °C temperature, 45% relative humidity, and a 16-h light/8-h dark cycle. After germination, water and fertilizers were supplied twice a week. Seedlings were harvested at different days and intervals (Figure 1).

Figure 2.1: Plant Growth Conditions and Sample Collection

2.2.2. RNA Extraction and cDNA Synthesis

Total RNA was extracted from seedlings using the Qiagen Plant Mini Kit (Hilden, Germany; catalog no. 74904), according to the manufacturer's instructions, and treated with DNase (RQ1 RNase free DNase; Promega, Madison, WI, USA) to remove traces of contaminating genomic DNA. After DNase treatment, the samples were purified by phenol-chloroform extraction, followed by ethanol precipitation. The purified RNA was quantified using a NanoDrop Spectrophotometer (Thermo Fisher Scientific, Wal-tham, MA, USA). Subsequently, cDNA was synthesized using reverse transcriptase (Superscript III; Invitrogen, Carlsbad, CA, USA) and a random hexamer (oligo dT) primer. The sequences of forward and reverse primers are given in Table 2.2.

2.2.3. Library Preparation for Transcriptome Sequencing

The mRNA from 12-d- and 20-day-old samples were enriched using oligo (dT) beads. A total amount of 3 µg RNA per sample was used as input material for the RNA sample preparations. Then, total RNA was extracted and was sent to the company, Novogene Co., Ltd., for Next Generation Sequencing analysis. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, Ipswich, MA, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out us-ing divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer 5X. First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H –). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase. After adenylation of 3′ ends of DNA

fragments, NEBNext Adaptor with hairpin loop structure was ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 150~200 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 µL USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers, and Index (X) Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system. The workflow for library preparation and transcriptome sequencing is shown in supporting Figure 2.2.

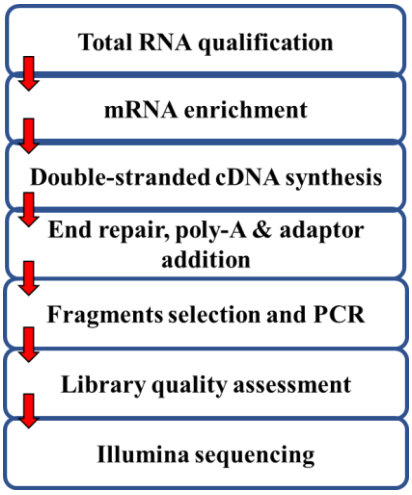


Figure 2.2: The workflow for library preparation and Transcriptome sequencing.

2.2.3.1. Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Gen-eration System using HiSeq PE Cluster Kit cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq platform and 125-bp/150-bp paired-end reads were generated.

2.2.3.2. Data Analysis

Quality control Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing poly- N, and low-quality reads from raw data. At the same time, Q20, Q30, and GC content from the clean data were calculated, as given in supporting Table 2.1. All the downstream analyses were based on the clean data with high quality.

Table 2.1. Datatable for Quality Control

Sample name	Raw reads	Clean reads	raw bases	clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
Day12	24196938	24043946	3.6G	3.6G	0.03	98.03	94.06	45.88
Day20	22829416	22165486	3.4G	3.3G	0.02	98.26	94.59	45.49

Reads mapping to the reference genome. Reference genome (TAIR 10) and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using Bowtie v2.2.3 and paired-end clean reads were aligned to the reference genome using TopHat v2.0.12. We selected TopHat as the mapping tool, as it can generate a database of splice junctions based on the gene model annotation file and, thus, a better mapping result than other non-splice mapping tools.

Quantification of gene expression level. High-throughput sequencing (HTSeq v0.6.1, University of Heidelberg, Heidelberg, Germany) was used to count the reads' numbers mapped to each gene. Then the FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. FPKM, expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced, considers the effect of sequencing depth and gene length for the reads' count at the same time and is currently the most used method for estimating gene expression levels [25]. HTSeq software was used to analyze FPKM, indicating the gene expression levels in this

experiment, using the union mode. The resulting files presented the number of genes with different expression levels and the expression level of single genes.

Differential expression analysis (For DESeq with biological replicates). Differential expression analysis of two conditions/groups (two biological replicates per condition) was performed using the DESeq R package (1.18.0). DESeq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting p-values were adjusted us-ing the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted p-value <0.05 found by DESeq were assigned as differentially expressed. (For DEGSeq without biological replicates.) Prior to differential gene ex-pression analysis, for each sequenced library, the read counts were adjusted by edgeR program package through one scaling normalized factor. Differential expression analysis of two conditions was performed using the DEGSeq R package (1.20.0). The p values were adjusted using the Benjamini and Hochberg method. Corrected p-value of 0.005 and log2(Fold change) of 1 were set as the threshold for significantly differential expression.

SNP analysis. Picard-tools v1.96 and samtools v0.1.18 were used to sort, mark duplicated reads, and reorder the bam alignment results of each sample. Genome Analysis Toolkit, GATK2 (v3.2) software was used to perform SNP calling. The map-ping status of reads was provided in BAM files, which were visualized using the Inte-grative Genomics Viewer (IGV) software.

2.2.4. Sanger Sequencing

After doing PCR with equal amounts of cDNA (100 ng) in each reaction of 20 μL volume, the PCR products were purified by 1% agarose gels and the bands were cut out and frozen. DNA was purified using the QIAquick Gel Extraction kit, and concentra-tion was measured by Nano-Drop. Sequencing of the purified DNA was performed using the Big Dye Terminator v3.1 Cycle

Sequencing Kit (Thermo Fisher Technologies, Waltham, MA, USA) using the forward and reverse primers (Table 2.2). The raw se-quencing data were analyzed using the Sequence Scanner software version 2 (Applied Biosystems) and DNADynamo software.

Table 2.2. List of candidate genes for U-to-C RNA editing sites with the forward and reverse primer sequences.

Gene ID	Read counts		Position	Primer sequence		Length
	12days	20days		Forward	Reverse	
AT4G14940	225	186	8544440 @UTR	TTCCGGTTATGCCAACTCTT	TGAAAGGATTTTGCCGATGG	300bp
AT4G16380	268	99	9255546 @ CDS	ACAAACTCTGTTCCAAAGGC	GGCGGCATTCCATATCCATT	393bp
AT4G29950	1650	713	14657330 @ UTR	GGATTCGGAGAGAGAATAAAAACG	CGGAAGAACACCGAGATTCA	315bp
AT4G32190	1241	505	15546833 @ CDS	GAAGTCCGAACAGGCTTCTC	CTTTCCACCACCATTTCAGC	209bp
AT2G07709	107	16	3392826 @ Intron	TATTTCTTCGGAGCCGTTTC	TTCCCTTTACTAGTAGGGTG	428bp
AT2G07715	129	10	3412532 @ CDS	CCAAGCCAATAGGCGAAAGG	CGATCACTACATAAGCCGCT	310bp
AT2G16586	407	432	7191444 @ UTR	CGGTAGATTAGTTGGAACGA	GAGTAACATGGCGTTCATGT	403bp
AT3G41768	647	240	14198871 @ UTR	CCATAAACGATGCCGACCAG	GCCTCCACGTAGCTAGTTAG	301bp
AT3G47965	91	79	17708862 @ UTR	CACATGATAGAAGCTCCTGGTG	GATACCAAAGGCGATTCAGC	153bp
AT3G54000	80	355	19998466 @ UTR	GTGGAGAGGACGAGGTTTTG	CCCGTGACAGACTGACATTT	122bp
AT5G02670	61	49	603074@UTR	TTATTCACCACGAAGAAAAA	CAGGAAGCTTCCATTGTTGA	302bp
AT5G08740	146	86	2848835 @ CDS	TACTGGCTGGAATATATGGG	GACCTTTGTGTCTCAGAGAG	380bp
AT5G23380	67	52	7869982 @ CDS	CCTCCATTTGGAGCAATGAC	GAGAGAGAGAGAGAGATGGGAATG	237bp
AT5G43970	345	276	17692876 @ UTR	GAGGAGGCGCTCTTTATTCC	CGTCTTCCTTGGGACACGAT	206bp
AT4G32430	07	19	15653919@UTR	CGGTAGATTAGTTGGAACGA	GAGTAACATGGCGTTCATGT	301bp
AT5G52530	675	293	21319578 @ CDS	TGCCAATCTGGTTATGACTCC	GAAGGCACCTCCTTGTATGG	291bp
AT5G52530			21320395 @ CDS	GGGATCGTAAATTCAGGGGC	CCTCACTCTTTCCGTCATCCTC	200bp
AT5G62220	298	127	24989428 @ CDS	GCCGCTGCGTAATTTGACCA	CGCACTTTCTAACGACGGGT	197bp
AT1G22190	1169	2256	7836325 @ CDS	CCCTCCTCTCAGAAATCTACACA	CGGAAGTACCGGTTTGTTTC	313bp
AT5G42320	420	159	16918673@UTR	ACTCCAGTGATGATTTAATG	TGTCTAATGTGTTTTTCAGG	302
AT1G29340	1934	2638	10266697 @ UTR	CGGATGTCTCAGTTCCGATA	AACGAACGATCACAATGCAA	258bp
AT1G56290	190	69	21077241 @ CDS	GAAACGAAGACTCGAAGCTG	CTGGGCTTCTTCATGTTTCC	244bp
AT3G56550	20	8	20953075 @UTR	CCGGAGTCTGCAGTTATTTA	CCCGATTGAATGCTTTTGAT	231bp
AT2G16586	407	432	7191297@ UTR	CCTAGGCTGTCCCGAAGGTA	TGTAAACCAAACTCAACAAT	300bp

2.3. Results

2.3.1. Identification and Analysis of Differentially Expressed Genes (DEGs)in 12-d-old seedlings by RNA-Seq

The level of gene expression was measured by determining transcript abundance; the greater the transcript abundance, the higher the gene expression level [26]. In RNA-seq analysis, gene

expression level is estimated by counting the number of reads mapped onto genes or exons. Read count was proportional not only to the actual gene expression level but also to gene length and sequencing depth. Transcriptome data in-dictated that a total of 33,641 genes were expressed, of which 2140 were specifically expressed in 12-day-old seedlings' genes and 1485 in 20-day-old seedlings' (Figure 2.3A). Volcano plots were used to infer the overall distribution of differentially expressed genes (DEGs). In experiments without biological replicates, the threshold is normally set as $|\log2(\text{Fold Change})| > 1$ and q-value < 0.005. By contrast, in experiments with biological replicates, DESeq eliminates biological variation; therefore, we set our threshold as adjusted p-value (padj) < 0.05. (Figure 2.3B). The correlation coefficient is an important indicator of the reliability of the experiment: the closer the value of the correlation coefficient to 1, the greater the similarity between samples. The square of the Pearson's correlation coefficient (R2) should be greater than 0.92 under ideal experimental conditions. In this study, R2 was greater than 0.8, indicating a slight difference in gene expression between 12- and 20-d-old seedlings (Figure 2.3C).

The FPKM is the most well-known method of gene expression estimation in RNA-seq, as it considers the effects of both sequencing depth and gene length on read counts (Figure 2.3D). Figure 2.4A shows that read counts and FPKM values were higher in 12-d-old seedlings than in the control sample (20-d-old seedlings), indicating higher expression of genes in 12-d-old seedlings. A total of 2712 genes were differentially expressed, of which 1642 were upregulated and 1070 were downregulated (Figure 2.4B), further indicating higher expression in 12-d-old seedlings.

To compare gene expression levels under different conditions, an FPKM distribution diagram was used. The final FPKM value represents the mean of biological replicates. In general, an FPKM

value of 0.1 or 1 was used as a threshold to determine whether a gene is expressed or not. The number of genes with different expression levels is shown in Figure 2.4C.

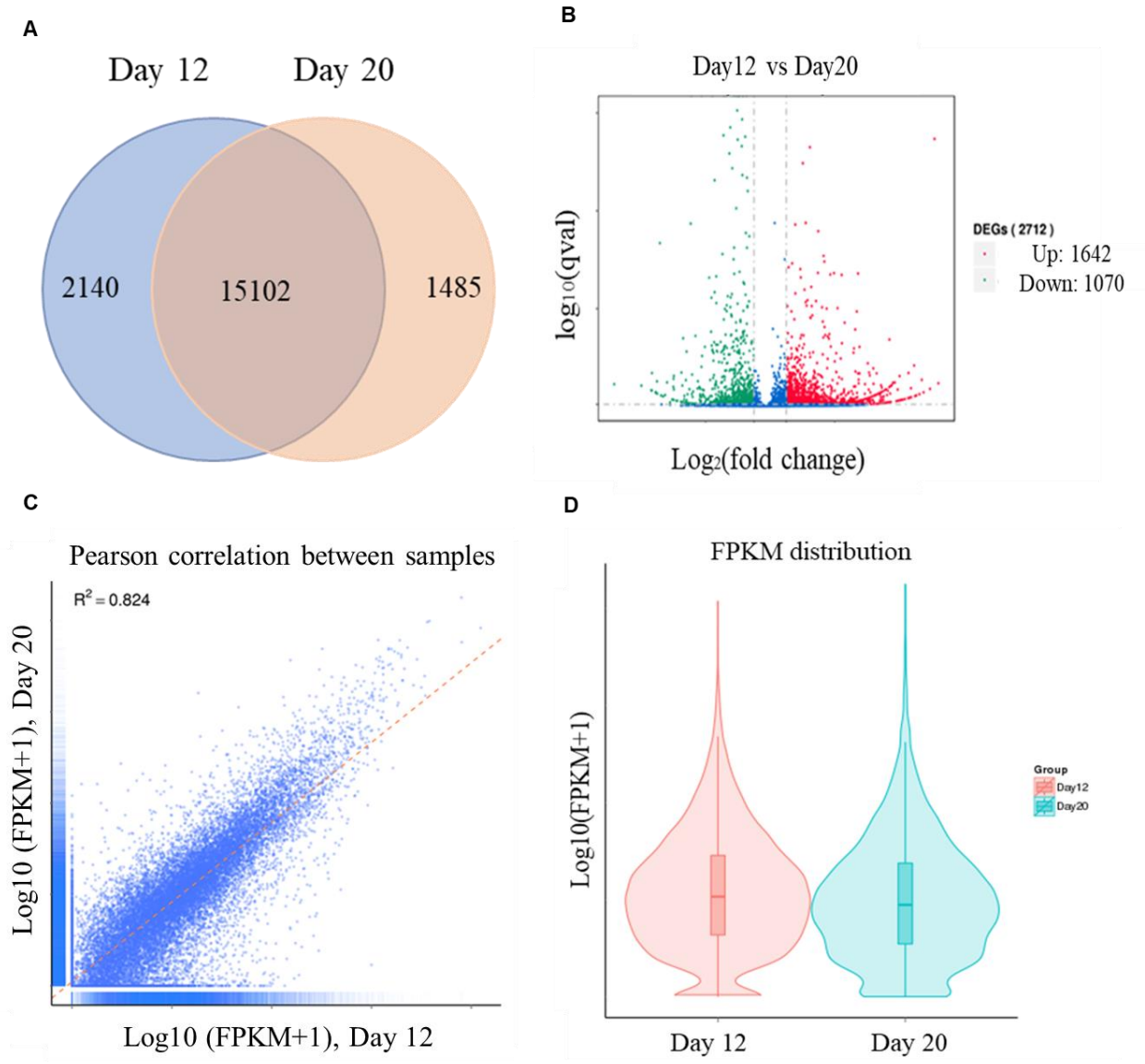


Figure 2.3. Analysis of genes differentially expressed between 12- and 20-d-old Arabidopsis seedlings. (A) Venn diagram of differentially expressed genes (DEGs). The sum of the numbers in each circle represents the total number of genes expressed within a sample, and the overlap represents genes expressed in both samples. (B) Volcano plot of DEGs. The x-axis shows the fold change in gene expression between different samples and the y-axis shows the statistical

significance of the differences in gene expression. Significantly up- and downregulated genes are highlighted in red and green, respectively. Genes showing no differential expression between 12- and 20-d-old seedlings are shown in blue. (C) Correlation analysis of gene expression between samples. R2 indicates the square of the Pearson's correlation coefficient. (D) Comparison of FPKM values of DEGs between 12- and 20-d-old seedlings.

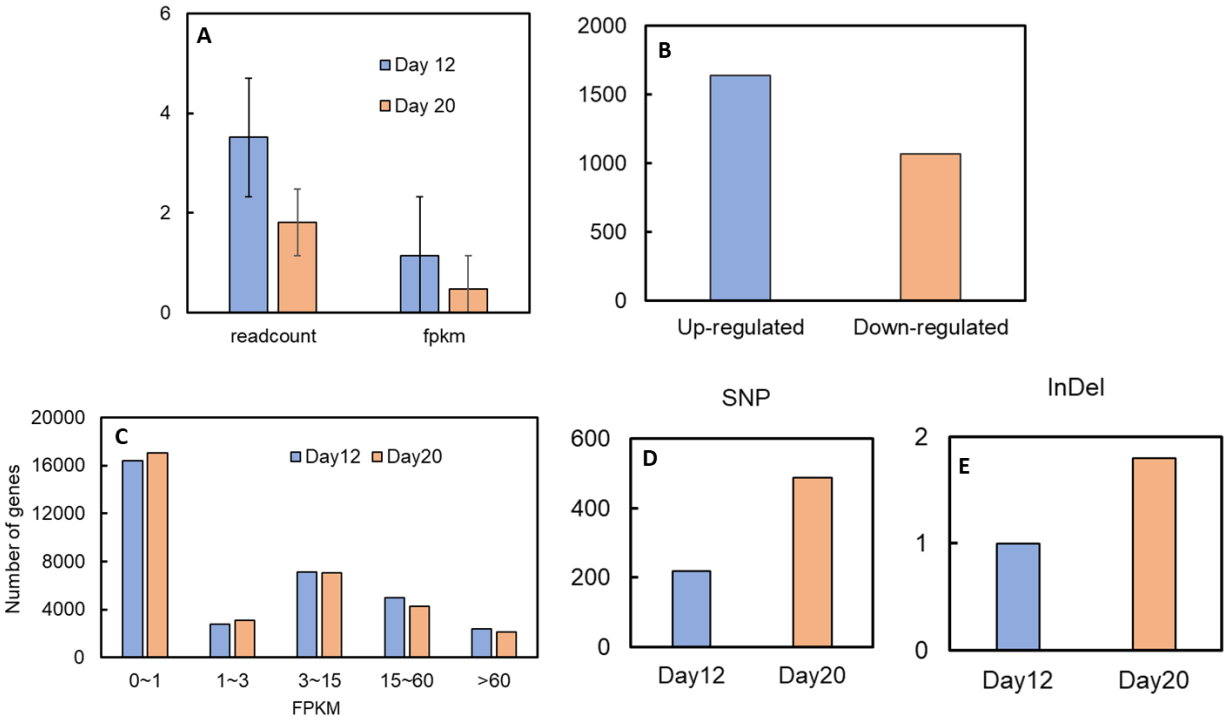


Figure 2.4. Summary of DEGs in RNA-seq analysis

2.3.2. Comparison of Nucleotide Differences between Genomic DNA in Database and RNA-Seq of 12- or 20-D-Old Seedlings

Comparison of RNA-seq data of 12-day- or 20-day-old Arabidopsis Col-0 plants to genomic DNA sequence in the database identified 12 types of possible nucleotide conversion patterns in transcripts: G-to-A, C-to-U, U-to-C, U-to-A, A-to-G, C-to-A, A-to-T, G-to-T, C-to-G, A-to-C, G-to-C, and U-to-G. Among these patterns, U-to-C was the third most predominant after G-to-A and

C-to-U . RNA-seq analysis revealed 590 different sites, of which 79 sites (13%) represented possible U-to-C conversion. Out of 253 genes showing nucleotide differences, 50 contained possible U-to-C conversion (Figure 2A, B). A list of candidate U-to-C RNA editing sites detected in Arabidopsis seedlings is given in Table 2.3.

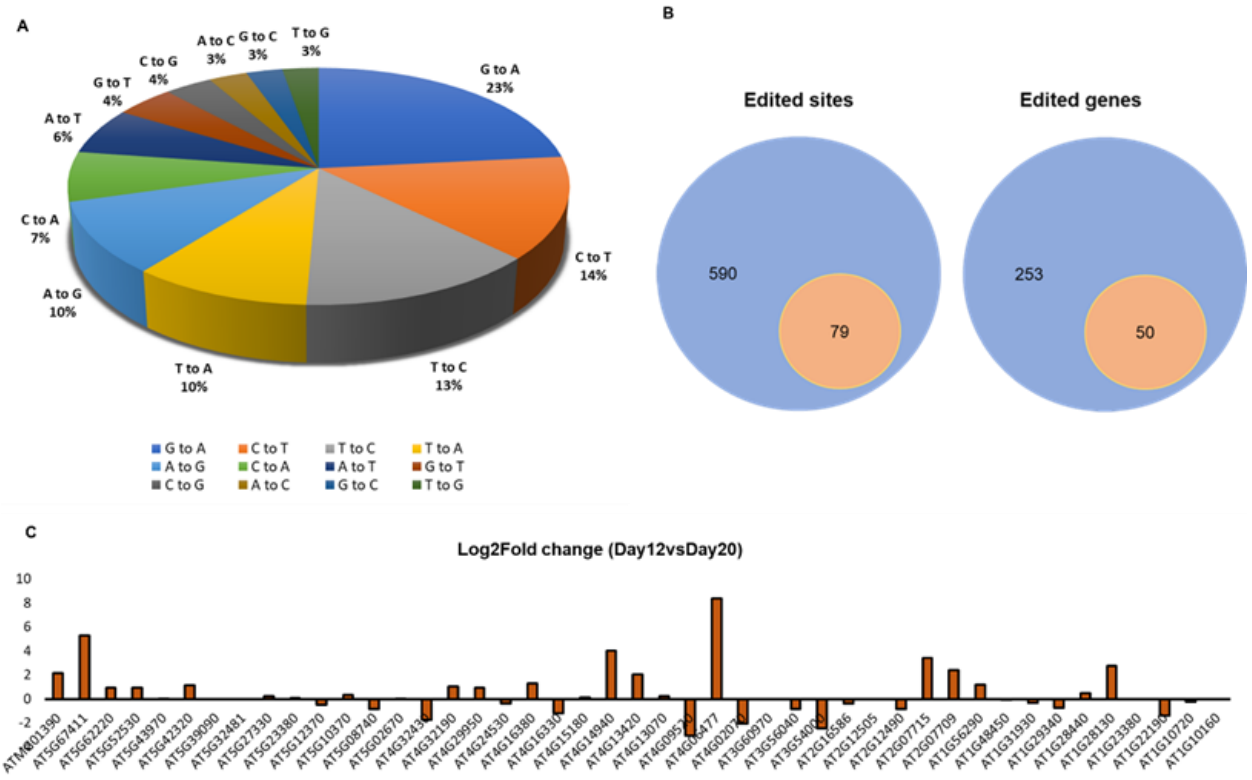


Figure 2.5. Analysis of single-nucleotide base conversions identified in 12-d-old Arabidopsis seed-lings by RNA-seq. (A) Pie chart showing the percentage for genes identified with single-nucleotide base conversions. (B) Number of total edited sites and edited genes (blue), and number of sites and genes with U-to-C mutations (orange). (C) Log2FC values for the genes identified with U-to-C nucleotide conversion. Genes were expressed to higher levels in 12-d-old seedlings than in 20-d-old seedlings (C).

Table 2.3 List of Candidate U-to-C RNA editing sites detected in Arabidopsis seedlings at different developmental stages.

S.No.	Position	Reads		Gene ID	Description
		12 Days	20 Days		
1	3412532	56	0	AT2G07715	Ribosomal Proteins L2, RNA binding domain
2	8544440	34	0	AT4G14940	Amine oxidase
3	26898977	2	0	AT5G67411	GRAS family transcription factor
4	8297931	4	0	AT1G23380	KNOTTED1-like homeobox gene 6
5	14657330	14	0	AT4G29950	Ypt/Rab-GAP domain of gyp1p superfamily protein
6	3392826	107	16	AT2G07709	-
7	362386	175	44	ATMG01390	-
8	7191444	105	197	AT2G16586	Unknown
9	3061212	498	2	AT4G06477	-
10	9226791	28	69	AT4G16330	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
11	5816271	12	6	AT3G17050	-
12	9255546	268	99	AT4G16380	Heavy metal transport/detoxification superfamily protein
13	14198871	647	240	AT3G41768	-
14	16918673	55	46	AT5G42320	Zn-dependent exopeptidases superfamily protein
15	17708862	0	21	AT3G47965	Unknown
16	24989428	27	0	AT5G62220	glycosyltransferase 18
17	21320395	0	12	AT5G52530	dentin sialophosphoprotein-related
18	2848835	146	86	AT5G08740	NAD(P)H dehydrogenase C1
19	15546833	13	0	AT4G32190	Myosin heavy chain-related protein
20	3392918	144	14	AT2G07709	-
21	21319578	0	5	AT5G52530	dentin sialophosphoprotein-related
22	21077241	0	2	AT1G56290	CwfJ-like family protein
23	7622202	0	2	AT4G13070	RNA-binding CRS1/YhbY (CRM) domain protein
24	17692876	29	0	AT5G43970	translocase of outer membrane 22-V
25	10266697	46	6	AT1G29340	plant U-box 17
26	7869982	17	0	AT5G23380	Protein of unknown function (DUF789)
27	7836325	19	0	AT1G22190	Integrase-type DNA-binding superfamily protein
28	19998466	36	0	AT3G54000	Unknown
29	603074	13	5	AT5G02670	Unknown
30	22561577	0	2	AT3G60970	multidrug resistance-associated protein 15
31	6025041	0	27	AT4G09520	Cofactor-independent phosphoglycerate mutase
32	909133	0	2	AT4G02070	MUTS homolog 6
33	7797368	0	4	AT4G13420	high affinity K+ transporter 5
34	8662474	0	3	AT4G15180	SET domain protein 2
35	12669828	0	2	AT4G24530	O-fucosyltransferase family protein
36	15653919	0	2	AT4G32430	Pentatricopeptide repeat (PPR) superfamily protein
37	5075516	0	2	AT2G12490	-
38	17587422	0	2	AT2G42200	squamosa promoter binding protein-like 9
39	17958701	0	2	AT2G43200	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
40	526197	0	5	AT3G02515	-
41	20795012	69	64	AT3G56040	UDP-glucose pyrophosphorylase 3
42	3264804	0	2	AT5G10370	helicase domain-containing protein/IBR domain-containing protein/zinc finger protein-related
43	9633752	0	2	AT5G27330	Prefoldin chaperone subunit family protein
44	12108844	0	9	AT5G32481	-
45	15644809	0	4	AT5G39090	HXXXD-type acyl-transferase family protein
46	3332097	0	2	AT1G10160	-
47	3564739	0	2	AT1G10720	BSD domain-containing protein
48	9825469	0	6	AT1G28130	Auxin-responsive GH3 family protein
49	9997031	0	2	AT1G28440	HAESA-like 1
50	4006628	0	13	AT5G12370	exocyst complex component sec10
51	5097198	0	5	AT2G12505	-
52	11465954	0	11	AT1G31930	extra-large GTP-binding protein 3
53	7014676	0	2	AT3G20087	N/A
54	15766171	0	2	AT2G37585	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
55	7191297	249	171	AT2G16586	Unknown
56	17908527	0	2	AT1G48450	Protein of unknown function (DUF760)

2.3.2. The next-generation sequencing (NGS) data of Arabidopsis for expressed PPR genes

We analyzed the next-generation sequencing (NGS) data of Arabidopsis for expressed PPR genes using the Bioinformatics & Evolutionary Genomics website. Out of 465 expressed PPR genes, 10 genes including AT3G62470, AT1G50270, AT1G16830, AT1G63080, AT1G06580, AT3G56550, AT1G09820, AT3G53360, AT2G22410, and AT4G32430 showed nucleotide conversion (Figure 2.6A). Out of 54 U-to-C variant genes, one gene, AT4G32430, was found as PPR gene (Figure 2.6B). The list of expressed genes, PPR genes that differed in base nucleotide conversions, the genes that differed in U-to-C base conversion, and the PPR gene that differed in U-to-C base conversion are shown in Table 2.4.

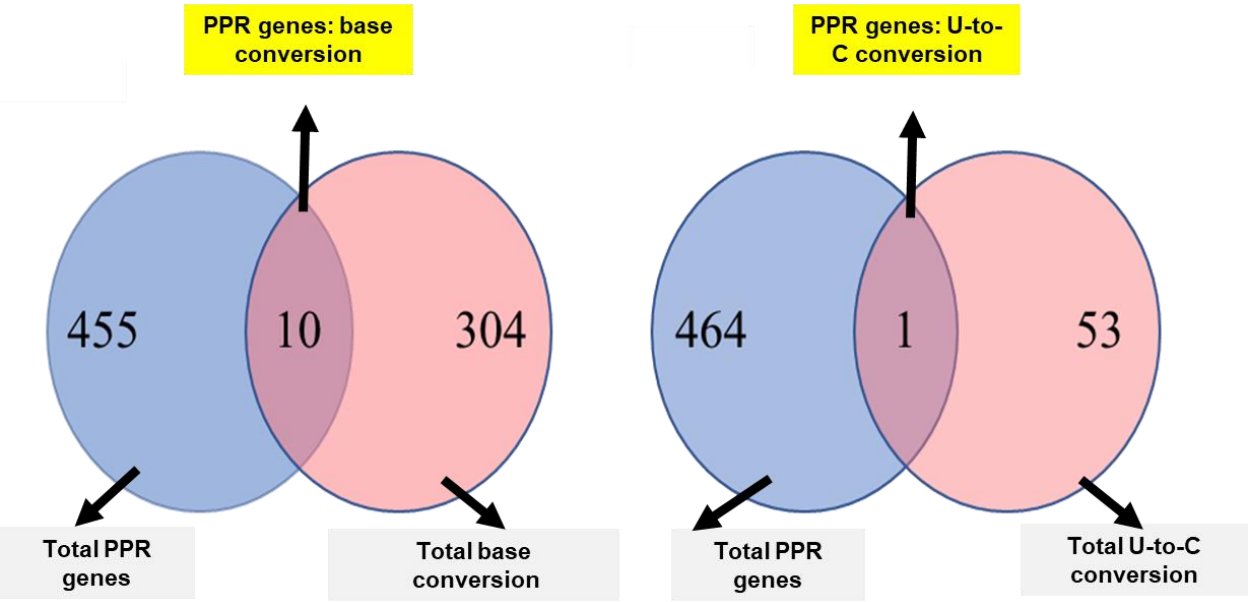


Figure 2.6. The next-generation sequencing (NGS) data of Arabidopsis for expressed PPR genes. Out of 465 expressed PPR genes, 10 genes including AT3G62470, AT1G50270, AT1G16830, AT1G63080, AT1G06580, AT3G56550, AT1G09820, AT3G53360, AT2G22410, and AT4G32430 showed nucleo-tide conversion (A). Out of 54 U-to-C variant genes, one gene, AT4G32430, was found as PPR gene (B).

Table 2.4: The list of expressed PPR genes that differed in base nucleotide conversions.

Names	total	elements
PPR genes that differ in base nucleotide conversion	10	AT3G62470 AT1G50270 AT1G16830 AT1G63080 AT1G06580 AT3G56550 AT1G09820 AT3G53360 AT2G22410 AT4G32430
Expressed PPR genes	455	AT1G19290 AT2G37320 AT5G59200 AT1G15510 AT3G18970 AT3G05240 AT1G69290 AT3G29230 AT1G80270 AT4G17616 AT3G60960 AT2G41080 AT1G79540 AT1G77340 AT1G03540 AT2G33760 AT1G62350 AT1G11290 AT3G49240 AT3G49710 AT1G74750 AT1G62910 AT4G26680 AT1G62260 AT5G50990 AT5G39680 AT5G44230 AT2G16880 AT4G33170 AT2G20720 AT2G36240 AT1G76280 AT5G01110 AT2G13600 AT1G02060 AT5G40405 AT3G46790 AT5G14080 AT1G53330 AT5G48910 AT5G02830 AT5G61400 AT3G15200 AT4G20770 AT5G15340 AT3G06430 AT1G11900 AT4G08210 AT5G66500 AT5G39710 AT5G08305 AT3G47840 AT1G73710 AT3G46870 AT1G80880 AT5G48730 AT3G59040 AT4G02820 AT1G51965 AT4G37170 AT4G21170 AT3G02650 AT3G04130 AT5G67570 AT5G57250 AT3G58590 AT1G77170 AT4G14050 AT3G29290 AT5G25630 AT3G26630 AT4G35130 AT4G14190 AT1G60770 AT1G28020 AT5G28370 AT2G17525 AT5G24830 AT1G11710 AT1G43010 AT2G01740 AT1G13410 AT3G49730 AT4G18840 AT1G09900 AT3G56030 AT2G17033 AT4G01570 AT2G32230 AT2G21090 AT2G37230 AT1G22830 AT1G32415 AT1G17630 AT2G34370 AT2G44880 AT2G27800 AT1G55890 AT2G29760 AT4G14820 AT1G77150 AT5G55740 AT1G80550 AT1G64430 AT5G11310 AT2G03380 AT4G26800 AT2G37310 AT5G27110 AT3G46610 AT3G61170 AT1G63330 AT2G18940 AT1G73400 AT1G52620 AT4G30700 AT2G19280 AT1G26460 AT5G62370 AT5G08490 AT3G13880 AT2G03880 AT1G71420 AT5G14820 AT1G08610 AT5G61990 AT5G65570 AT5G02860 AT3G22470 AT3G17370 AT4G31070 AT3G13770 AT2G06000 AT5G40400 AT3G02010 AT1G28690 AT1G74580 AT2G20540 AT3G09060 AT2G17140 AT5G41170 AT5G19020 AT2G39230 AT2G34400 AT2G48000 AT1G05600 AT5G06540 AT4G18520 AT3G22150 AT4G02750 AT5G50280 AT3G01580 AT3G49740 AT5G14770 AT1G31790 AT5G36300 AT4G30825 AT1G12250 AT3G51320 AT1G20300 AT4G32450 AT1G16480 AT5G60960 AT2G26790 AT1G30610 AT3G25060 AT4G39530 AT2G32630 AT2G25580 AT3G50420 AT2G04860 AT1G62680 AT2G17670 AT2G36980 AT5G28380 AT5G18475 AT1G68930 AT1G34160 AT2G02150 AT2G20710 AT1G12775 AT1G22960 AT1G09220 AT2G36730 AT5G61370 AT3G62890 AT4G14850 AT5G52850 AT1G31920 AT1G18485 AT1G62720 AT3G04750 AT2G41720 AT4G21190 AT5G52630 AT1G62930 AT5G18950 AT2G35130 AT4G01400 AT1G31430 AT5G14350 AT1G61870 AT1G02370 AT2G15630 AT3G03580 AT5G13230 AT1G05670 AT4G16470 AT3G49142 AT1G55630 AT3G15130 AT5G37570 AT2G39620 AT1G80150 AT2G02750 AT4G21065 AT4G21705 AT3G14580 AT1G15480 AT1G19520 AT5G43820 AT3G47530 AT3G07290 AT5G66631 AT1G06710 AT3G15590 AT4G16835 AT1G31840 AT3G25970 AT5G15010 AT5G28340 AT3G22670 AT3G11380 AT2G44920 AT1G09190 AT4G04790 AT4G20740 AT3G23020 AT5G12100 AT3G02330 AT5G47360 AT1G13630 AT1G33350 AT3G61520 AT5G10690 AT3G16710 AT4G19220 AT1G18900 AT4G20090 AT5G39350 AT5G43790 AT5G28460 AT5G59900 AT3G16010 AT1G63150 AT1G74400 AT1G64310 AT5G56310 AT4G22760 AT3G04760 AT5G42310 AT4G18975 AT1G71210 AT5G64320 AT3G53170 AT4G01030 AT3G24000 AT4G31850 AT1G63130 AT5G53490 AT1G06140 AT5G46460 AT2G33680 AT5G38730 AT3G06920 AT2G38420 AT5G15280 AT1G64583 AT3G25210 AT2G01860 AT5G59600 AT1G62670 AT3G09040 AT1G26500 AT1G71060 AT4G34830 AT1G66345 AT1G13800 AT1G12700 AT2G42920 AT5G55840 AT1G74600 AT4G13650 AT1G10330 AT5G46680 AT4G04370 AT4G33990 AT1G62914 AT3G48810 AT1G02150 AT1G79490 AT1G08070 AT2G30100 AT3G08820 AT2G46050 AT5G04780 AT3G42630 AT5G18390 AT3G13150 AT1G26900 AT1G71460 AT1G52640 AT1G43980 AT1G63400 AT1G63320 AT1G20230 AT5G06400 AT2G40720 AT3G18020 AT1G03510 AT2G01360 AT3G61360 AT4G01990 AT3G60980 AT5G13770 AT3G28640 AT3G16890 AT5G42450 AT1G10270 AT4G39620 AT4G18750 AT4G38150 AT1G74900 AT3G49170 AT2G18520 AT2G17210 AT1G13040 AT4G19440 AT1G11630 AT1G19720 AT5G09450 AT3G28660 AT5G15300 AT3G15930 AT3G14330 AT3G23330 AT3G13160 AT1G12300 AT1G77360 AT1G07740 AT2G35030 AT5G39980 AT3G21470 AT2G01510 AT4G25270 AT1G69350 AT4G38010 AT4G16390 AT5G27460 AT1G03100 AT1G30290 AT3G57430 AT2G45350 AT1G63070 AT2G31400 AT3G26782 AT1G74850 AT3G26540 AT4G21300 AT1G03560 AT1G09410 AT2G28050 AT1G10910 AT4G36680 AT3G20730 AT2G15980 AT5G66520 AT4G11690 AT3G60050 AT1G64580 AT5G46580 AT5G03800 AT1G68980 AT5G65560 AT1G59720 AT2G30780 AT5G16860 AT1G56570 AT1G01970 AT5G40410 AT5G09950 AT5G27270 AT3G02490 AT1G53600 AT2G15690 AT1G09680 AT1G14470 AT5G65820 AT5G46100 AT1G62590 AT3G18840 AT1G74630 AT4G15720 AT3G18110 AT5G15980 AT5G47460

2.3.4. Identification of Genes Harboring U-to-C RNA Editing Site

We selected the genes of both samples that had a minimum number of reads to be able to infer an editing event. This minimum number should be reasonably high to minimize the impact of sequencing artifacts. For example, the T-to-C change at position 14,198,871 in AT3G41768 was supported by 647(29%) and 240 (19%) in 12-d-old and 20-d-old seedlings, respectively. In addition, there were some variants that were sup-ported by 100% of the reads in both samples (12- and 20-d-old). Therefore, these are several editing events that seem to be polymorphisms. For the same gene, we found many reads in the same U-to-C conversions. Genes with higher read coverage were further examined for the confirmation of U-to-C RNA editing sites. Genes, such as AT2G16586, AT5G02670, AT5G42320, AT4G16380, and AT5G08740, showed 249, 13, 55, 268, and 146 reads at the converted sites, respectively. Genes showing extremely low reads (0, 2) were also analyzed by RT-PCR. However, very few sites were confirmed as editing events. Because many reads mapped to each U-to-C conversion site, we considered that these nucleotide conversions were caused by RNA editing [27]. The flowchart for methodology for identification of U-to-C RNA editing sites is shown in Figure 2.7.

2.3.5. Validated the RNA-seq-based candidates experimentally by Sanger sequencing of both genomic, gDNA, and cDNA for all candidate genes

Furthermore, we validated the RNA-seq-based candidates experimentally by Sanger sequencing of both genomic, gDNA, and cDNA for all candidate genes. We extracted DNA and mRNA from the same aliquots of seedling samples. By sequencing the paired DNA and cDNA samples and analyzing each chromatogram by two individual independently we confirmed the U-to-C RNA edited sites. The cDNA showed a double peak, representing T and edited C nucleotides, while no double peak was observed in gDNA sequencing. The sequencing was performed using sense

primer targeting at the editing sites. Validation using PCR and Sanger sequencing verified seven genes, including AT2G16586, AT5G42320, AT5G02670, AT3G41768, AT4G32430, AT3G47965, and AT5G52530, containing U-to-C RNA editing sites. The Sanger sequence chromatograms for all seven edited genes are showed in Figure 2.8. The raw sequencing data were analyzed using the software, DNADyanamo and Sequence Scanner version 2 (Applied Biosystems). When the edited and unedited products were presented together, a dual peak (T (unedited) and C (edited)) was observed at the target site. The editing efficiency was calculated from peak area and a list of genes showing U-to-C RNA editing in 12-d- and 20-day old Arabidopsis seedlings is given in Table 2.5. Furthermore, we also investigated the editing efficiency at different developmental stages of Arabidopsis, such as four days and eight days. It was found that no editing occurred at early stages of development, like in four days, while a few editing could be identified in 8-day-old seedlings (Table 2.6). The U-to-C RNA editing sites were majorly located within the UTRs of mature mRNAs.

Table 2.5. List of genes identified with U-to-C RNA editing sites in 12-day and 20-day old Arabidopsis seedlings

S.No.	Position	Edited Site	Gene ID	RNA Editing Efficiency (in %)		Encoded Protein
				12 Days	20 Days	
1.	14198871	5′ UTR	AT2G16586	77.30	65.74	Transmembrane protein
2.	16918673	CDS	AT5G42320	24.20	0	Zn-dependent exopeptidase superfamily protein
3.	603074	5′ UTR	AT5G02670	0	22.80	Hypothetical protein
4.	7191297	3′ UTR	AT3G41768	45.54	49.65	Ribosomal RNA
5.	15653919	3′ UTR	AT4G32430	0	20.43	PPR-like superfamily protein
6.	17708862	3′ UTR	AT3G47965	24.54	22.48	Hypothetical protein
7.	21320395	CDS	AT5G52530	20.65	0	Dentin sialophosphoprotein-like protein

Table 2.6: List of genes identified with U-to-C RNA editing in Arabidopsis seedlings at different developmental stages.

Gene ID	RNA editing efficiency (in %)			
	4 days	8 days	12days	20 days
AT2G16586	0	67.54	77.30	65.74
AT5G42320	0	20.47	24.20	0
AT5G02670	0	0	0	22.80
AT3G41768	0	40.76	45.54	49.65
AT4G32430	0	0	0	20.43
AT3G47965	0	0	24.54	22.48
AT5G52530	0	0	20.65	0

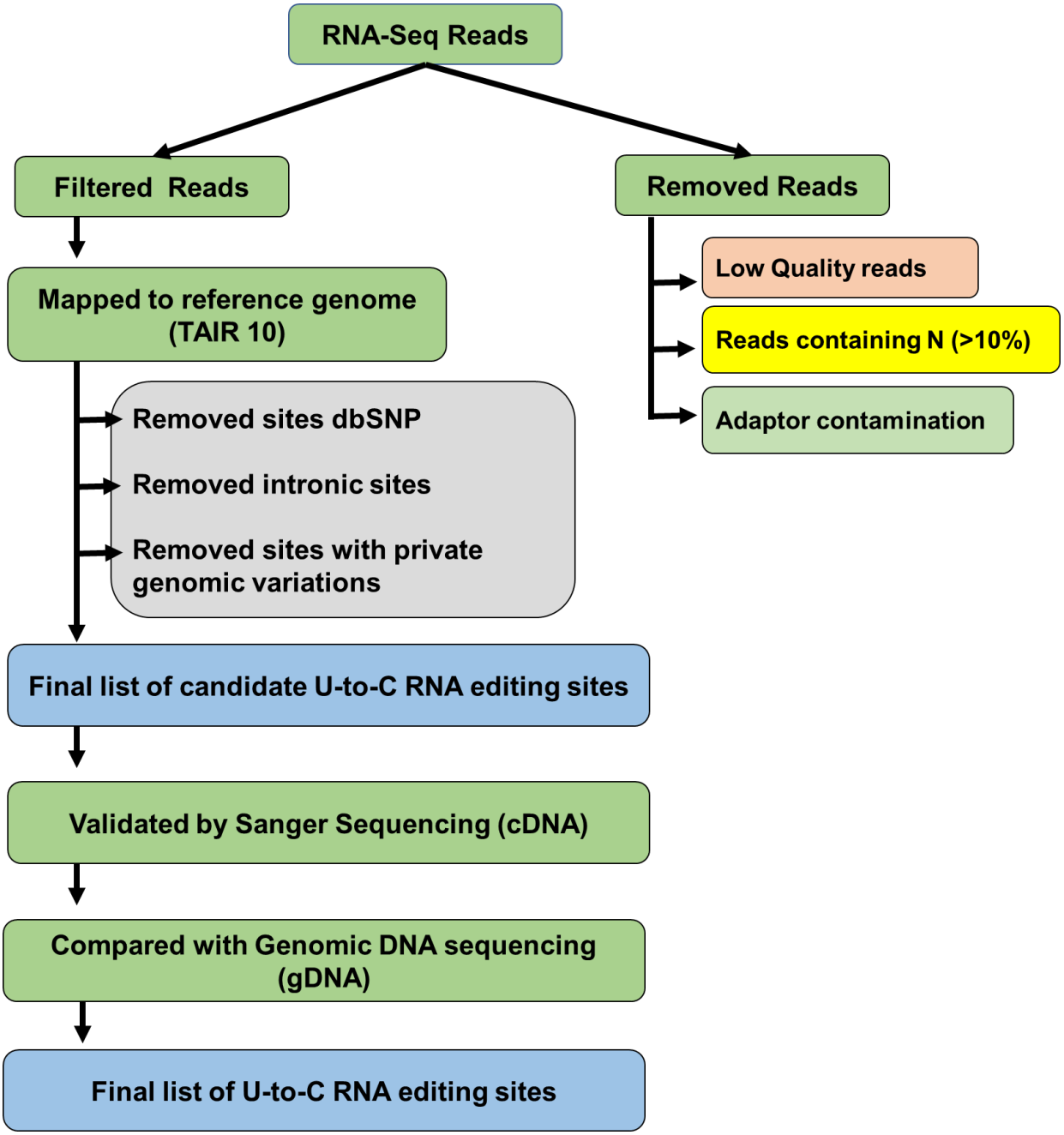


Figure 2.7. The flowchart for methodology for identification of U-to-C RNA editing site. (A) Raw reads are filtered to remove reads containing adapters or reads of low quality, so that downstream analyses are based on clean reads. The filtering process is as follows. (1) Discard reads with adaptor contamination. (2) Discard reads when uncertain nucleotides constitute more than 10% of either read (N > 10%). (3) Discard reads when low-quality nucleotides (base quality less than 20)

constitute more than 50% of the read. For mapping sequences, TopHat2 was chosen for plant genomes. The mismatch parameter was set to 2 and other parameters were set to default. Appropriate parameters were also set, such as the longest intron length. Only filtered reads were used to analyze the mapping status of RNA-seq data to the reference genome. Edited sites were further validated and confirmed by RT-PCR.

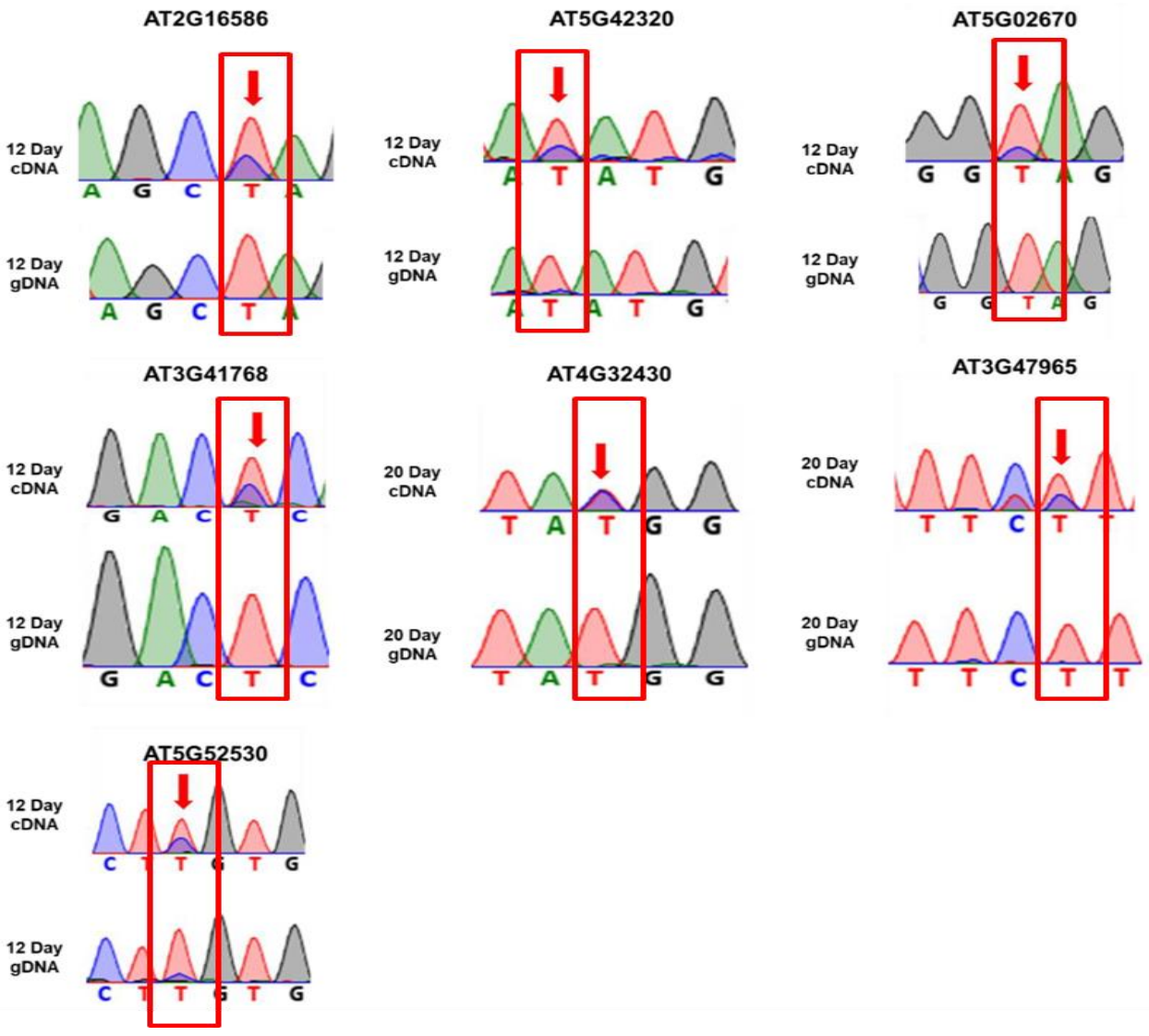


Figure 2.8. The Sanger sequence chromatogram depicting the U-to-C types of RNA editing events in 12-d- and 20-day-old seedlings from the same tissues of Arabidopsis via cDNA and genomic, gDNA using forward primers. Arrows indicate the position of RNA editing.

2.4. Discussion

In our knowledge, this is the first report of U-to-C RNA editing for nuclear genes confirmed by both RNA-seq and Sanger sequencing in flowering plants.

Total RNA isolated from 12-d-old seedlings was examined by NGS, and DEGs were identified based on FPKM values and read counts. The results showed that DEGs were expressed to higher levels in 12-d-old seedlings than in 20-d-old seedlings. This was confirmed by higher FPKM values and read counts and more upregulated genes in 12-d-old seedlings than in 20-d-old seedlings. The ANOVA test was performed for comparing the gene expression levels. The summary for regression analysis of differentially expressed genes among the replicates of 12-d- and 20-d-old seedlings is given in Table 2.7.

Table 2.7: Summary for regression analysis of differentially expressed genes among the replicates of 12 day and 20day old seedlings.

Regression Statistics	
Multiple R	0.839863
R Square	0.705371
Adjusted R Square	0.705262
Standard Error	139.8466
Observations	2711

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	50.41472	2.761485	18.25638	2.56E-70	44.99989	55.82955	44.99989	55.82955	
5.347892	0.314265	0.003902	80.53325	0	0.306613	0.321916	0.306613	0.321916	

ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	1.27E+08	1.27E+08	6485.604	0	
Residual	2709	52980142	19557.08			
Total	2710	1.8E+08				

Additionally, PPR genes were also expressed to higher levels in 12-d-old than in 20-d-old seedlings, as indicated by the log2FC values. These data suggest that DEGs are more likely to be expressed in young Arabidopsis seedlings than in older seedlings. Therefore, more mutations could occur at this stage of development because RNA editing events are more frequent in seedlings than in any other plant tissue.

While investigating for RNA editing events to create a global map of high-quality candidates, an appropriate balance between sensitivity (identifying a highly inclusive set of possible edits) and specificity (being more confident that a call is, in fact, a true RNA edit) is required. We considered it better to have a fewer number of candidate RNA editing events that are more likely to be true than to have a larger number with an increased percentage of false positives. We undoubtedly did not score a substantial number of true, low-level, U-to-C RNA editing events in the process. Up to 90% of nucleotide variants that are not SNPs (either in dbSNP or private genomic SNVs) are U-to-C calls; this suggests they are likely to be U-to-C editing candidates. Furthermore, more than 85% of these candidates are located in UTRs. Our candidate U-to-C RNA editing sites had a different variant frequency from known SNPs. They tended to cluster predominantly in the untranslated regions.

We investigated single-nucleotide base changes and the percentage of read coverage was calculated (Table 2.2). We predicted 12 types of nucleotide differences, including possible U-to-C conversions. RT-PCR products of the genes including the candidate U-to-C conversions were subjected to Sanger sequencing. A total of seven genes, AT2G16586, AT5G42320, AT5G02670, AT3G41768, AT4G32430, AT3G47965, and AT5G52530, were identified as targets for U-to-C RNA editing (Table 2.5). The UTRs of genes encoding proteins involved in RNA metabolism and RNA binding, including PPR proteins, Zn-finger (ZnF)-related proteins, ribosomal protein L2,

transmembrane proteins, and two hypothetical proteins, were identified as target of U-to-C editing. Interestingly, the ribosomal RNA, AT3G41768, was identified for 45.65% of U-to-C RNA editing efficiency. Since about 50% of genes are affected with editing, it might had had significant effect on their functions. Similarly, the transmembrane protein, AT2G16586, was identified with 77.3% of U-to-C RNA editing efficiency, which may affect its general physiology. In addition, the PPR gene, AT4G32430, was also identified with 20.43% U-to-C RNA editing.

While RNA editing in introns or UTR regions can affect mRNA stability, translation, or splicing activity because of the modification of its secondary structure, those in coding region can also affect the translated polypeptide sequence [28–30]. In this study, we demonstrated that most U-to-C RNA editing events are in UTRs, which may affect the secondary structure and, consequently, the stability of mRNA.

In Arabidopsis, C-to-U and U-to-C RNA editing have been reported at the trans-lation borders of nuclear transcripts, AT1G29930.1 and AT1G52400.1 [31].These deamination(C-to-U) and amination (U-to-C) events accumulated at adjacent sites; therefore, the possibility that the deamination reaction serves as the amino group donor for the amination reaction was proposed, although the frequency of amination was higher than that of deamination [31]. Although this hypothesis is attractive, we could not detect the same RNA editing events in our RNA-seq data. Thus, the amino group donor of the U-to-C amination in plants is unclear. However, in cDNA AT3G47965 there is also a small T superposing with the C just upstream the edited T, showing the possible immediate donor of amino group. Previously, an extensive research on editing sites in nuclear transcripts for mRNA by Parallel Analysis of RNA Ends (PARE) and Massively Parallel Signature Sequencing (MPSS) data was reported. It showed that all 12 RNA editing patterns may exist in the nuclear genes and that perhaps the numbers of editing sites in a specific pattern may

vary. The study suggested that RNA editing is an essential RNA-based regulatory layer not only for mitochondrial and chloroplast genes but also for nuclear genes. However, a global vision of RNA editing in plant nu-clear protein-coding transcripts has not been realized. Therefore, this work intended to uncover the occurrence of RNA editing events in the nuclear genes of Arabidopsis. The day-specific characteristic of the U-to-C RNA editing events implied that these were post-transcriptional modifications, not genomic mutations. These editing were identified as a growth-dependent RNA editing efficiency alteration. Day 4 seedlings did not have RNA editing, at least (Table 2.6). It indicates that the enzyme important for this editing events might have been expressed at defined stages of seedling development.

Next, to validate whether the identified RNA editing sites were true positive, we searched for evidence of the identified RNA editing sites in Arabidopsis RNA-seq data generated by public laboratories, using online software http://signal.salk.edu/atg1001/3.0/gebrowser.php. All seven identified U-to-C RNA editing sites AT2G16586, AT5G42320, AT5G02670, AT3G41768, AT4G32430, AT3G47965, and AT5G52530 were aligned against the publicly available RNA-seq databases and confirmed our findings. The target T sites were identified as edited C sites in various databases. The comparative analysis of Arabidopsis RNA-seq is shown in Figure 2.9. The edited sites are indicated within red boxes. Further studies are needed to better understand the processes involved in U-to-C RNA editing, including the identification of cis or trans regulatory elements, isolation of editing enzymes, and validation of editing sites.

Figure 2.9: Validation of target U-to-C RNA editing sites on *Arabidopsis* RNA-seq database. **A)** AT2G16586, **B)** AT5G42320, **C)** AT5G02670, **D)** AT3G41768, **E)** AT4G32430, **F)** AT3G47965, **G)** AT5G5253. The identified U-to-C RNA editing sites were compared against the

publicly available RNA-seq databases. The target “T” sites are identified as edited “C” sites in various databases. The edited sites are indicated within red boxes. The data is generated from an online software http://signal.salk.edu/atg1001/3.0/gebrowser.php.

2.5. Conclusions

Our findings confirm the uridine-to-cytidine RNA editing sites in some nuclear genes in Arabidopsis thaliana. A comprehensive analysis of RNA-seq data to detect nucleotide base conversions was performed. In this study, we examined U-to-C RNA editing in Arabidopsis seedlings at different developmental stages. Sanger sequencing identified the sites and efficiency of seven U-to-C editing events. Most U-to-C RNA editing here identified occurred in the UTR of mature mRNAs. Thus, we confirmed the presence of U-to-C RNA editing in nuclear genes of plants. We provided the experimental basis to explore the mechanism involved in the amination of U-to-C editing and functions and effects of U-to C RNA editing on mRNA stability, other RNA modifications, and translation.

2.6. Reference

1. Benne, R. RNA-editing in trypanosome mitochondria. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1007**, (1989).
2. Chen, S. *et al.* Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238**, (1987).
3. Walkley, C. R. & Li, J. B. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. *Genome Biology* **18**, (2017).
4. Liu, H. *et al.* Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes. *Genome Research* **26**, (2016).
5. Wolf, J. tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *The EMBO Journal* **21**, (2002).
6. Bar-Yaacov, D., Pilpel, Y. & Dahan, O. RNA editing in bacteria: occurrence, regulation and significance. *RNA Biology* **15**, (2018).
7. Thomas, S. M., Lamb, R. A. & Paterson, R. G. Two mRNAs that differ by two nontemplated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5. *Cell* **54**, (1988).
8. Cattaneo, R., Kaelin, K., Baczko, K. & Billeter, M. A. Measles virus editing provides an additional cysteine-rich protein. *Cell* **56**, (1989).
9. Covello, P. S. & Gray, M. W. RNA editing in plant mitochondria. *Nature* **341**, (1989).
10. Gualberto, J. M., Lamattina, L., Bonnard, G., Weil, J.-H. & Grienenberger, J.-M. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature* **341**, (1989).

11. Hiesel, R., Wissinger, B., Schuster, W. & Brennicke, A. RNA editing in plant mitochondria. *Science* **246**, (1989).

12. Barkan, A. & Small, I. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology* **65**, (2014).

13. Lurin, C. *et al.* Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis[W]. *The Plant Cell* **16**, (2004).

14. Chateigner-Boutin, A.-L. Developmental co-variation of RNA editing extent of plastid editing sites exhibiting similar cis-elements. *Nucleic Acids Research* **31**, (2003).

15. Schallenberg-Rüdinger, M., Kindgren, P., Zehrmann, A., Small, I. & Knoop, V. A DYW-protein knockout in *Physcomitrella* affects two closely spaced mitochondrial editing sites and causes a severe developmental phenotype. *The Plant Journal* **76**, (2013).

16. Leu, K.-C., Hsieh, M.-H., Wang, H.-J., Hsieh, H.-L. & Jauh, G.-Y. Distinct role of *Arabidopsis* mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in *nad1* RNA editing. *RNA Biology* **13**, (2016).

17. Sun, T., Bentolila, S. & Hanson, M. R. The Unexpected Diversity of Plant Organelle RNA Editosomes. *Trends in Plant Science* **21**, (2016).

18. Bayer-Császár, E. *et al.* The conserved domain in MORF proteins has distinct affinities to the PPR and E elements in PPR RNA editing factors. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, (2017).

19. Schallenberg-Rüdinger, M. & Knoop, V. Coevolution of Organelle RNA Editing and Nuclear Specificity Factors in Early Land Plants. in (2016). doi:10.1016/bs.abr.2016.01.002.

20. Gerke, P. *et al.* Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*. *New Phytologist* **225**, (2020).

21. Gutmann, B. *et al.* The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants. *Molecular Plant* **13**, (2020).

22. Oldenkott, B., Yang, Y., Lesch, E., Knoop, V. & Schallenberg-Rüdinger, M. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in Escherichia coli. *Communications Biology* **2**, (2019).

23. Qulsum, U., Azad, Md. T. A. & Tsukahara, T. Analysis of Tissue-specific RNA Editing Events of Genes Involved in RNA Editing in Arabidopsis thaliana. *Journal of Plant Biology* **62**, (2019).

24. Qulsum, U. & Tsukahara, T. Tissue-specific alternative splicing of pentatricopeptide repeat (PPR) family genes in Arabidopsis thaliana. *BioScience Trends* **12**, 569–579 (2018).

25. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, (2010).

26. Ramaswami, G. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nature Methods* **10**, (2013).

27. E. Yang, E. van N. M. Z. et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).

28. Drescher, A. *et al.* 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**, (2002).

29. Farré, J.-C., Aknin, C., Araya, A. & Castandet, B. RNA Editing in Mitochondrial Trans-Introns Is Required for Splicing. *PLoS ONE* **7**, (2012).

30. Wu, B. *et al.* Identification of Symmetrical RNA Editing Events in the Mitochondria of Salvia miltiorrhiza by Strand-specific RNA Sequencing. *Scientific Reports* **7**, (2017).

31. Meng, Y. *et al.* RNA editing of nuclear transcripts in Arabidopsis thaliana. *BMC Genomics* **11**, (2010).

Chapter 3

The U-to-C RNA editing affects the mRNA stability of nuclear genes in Arabidopsis thaliana

3.1.Introduction

The process by which genetic information flows from DNA to RNA to protein is defined as the central dogma in molecular biology [1]. Consequently, RNA follows the nucleotide base sequences of the gene from which it is transcribed. RNA plays a very significant role in gene expression, which is mediated by various steps including RNA cleavage, RNA splicing, translation into amino acids, post-transcriptional modifications, RNA editing, and mRNA stability[2,3]. RNA editing is defined as the change in RNA sequences by base deletion, insertion, and conversion [4–6]. There are two main types of substitutional RNA editing in plant mitochondria and chloroplasts. One of them is cytidine-to-uridine (C-to-U) conversion, which is observed in all land plants, while the other type is a reverse uridine-to-cytidine (U-to-C) alteration. The U-to-C RNA editing is rare in terrestrial plants, except in hornworts and ferns, and is therefore referred to as an occasional phenomenon [7]. Recently, genome-wide characterisation of U-to-C RNA editing events has been reported in the nuclear genes of *Arabidopsis thaliana* [8]. Because of their rare occurrence, negligible work has been done to investigate the effects and mechanisms involved in U-to-C RNA editing.

In seed plants, RNA editing comprises of the deaminase reaction of cytidines to uridines and is mainly found in chloroplasts and mitochondria [9] .The C target sites are recognised by the PLS subfamily of pentatricopeptide repeat (PPR) proteins [10,11], which bind to RNA molecules and induce plastid and mitochondrial RNA development in eukaryotic organisms [12]. Most of the PLS proteins are characterised by the presence of an extended C-terminal along with highly conserved protein domains of E1, E2, and DYW [13–15]. These PPR motif-containing proteins are described by the presence of a degenerate 35 amino acid tandem repeat. The *Arabidopsis* genome encodes for more than 400 PPR proteins, including the P and PLS subfamilies [16]. The RNA editing machinery, commonly known as the editosomes, comprises of four types of proteins: the pentatricopeptide

repeat (PPR) protein, organelle RNA recognition motif (ORRM1) protein, RNA editing factor interacting protein (RIP), multiple organellar RNA editing factor (MORF), and organelle zinc-finger protein (OZ1). PPRs constitute a large family of proteins which are responsible for RNA editing [13,17–20][17-21]. The C-terminal DYW domain of PPR proteins has a function similar to that of cytidine deaminase, which is responsible for C-to-U conversion [21].

In *Arabidopsis*, RNA editing usually occurs in seedlings and leaves. Recently, extensive uridine to cytidine and adenine to inosine (G) RNA editing events have been reported in 12-day-old whole seedlings and leaves of 21-day-old seedlings, respectively [22]. Additionally, among all RNA editing events specific to the untranslated regions (UTRs) of mature mRNAs, U-to-C conversion was found to be the most common, followed by uridine to guanine (U-to-G) editing [22]. However, in this study, direct comparison of gDNA and cDNA sequencing from the same sample, which is indispensable for eliminating the possibility of DNA mutations or sequencing errors, was not conducted. In the current study, a comparison of cDNA with the gDNA sequence is presented. In addition, RNA-sequencing-based analysis of 12-day-old and 20-day-old *Arabidopsis* seedlings has also revealed certain U-to-C RNA editing events [8], which prompted us to explore the genes involved in the U-to-C RNA editing in *Arabidopsis*.

Here, we have investigated the impact of U-to-C editing on the mRNA stability of the nuclear genes in *Arabidopsis*. In this study, we examined the U-to-C RNA editing-related genes in *Arabidopsis thaliana*, a model dicot flowering plant. We focused on U-to-C RNA editing events in the UTRs of mature mRNAs, which are known to affect their secondary structures. We also investigated the stability of mRNAs during callus culture. The effect of U-to-C RNA editing in *Arabidopsis* was analysed by adding the transcription inhibitor, actinomycin D (Act D), to cell

suspension cultures of transgenic *Arabidopsis* generated by *Agrobacterium*-mediated transformation.

3.2. Materials and Methods

3.2.1. Plant growth and sample collection

Colombia (Col-0) seeds of *Arabidopsis thaliana* were soaked in water and nurtured in the dark at low temperatures (4°C) for 3-4 days. Seeds were sown in small paper cups containing a 1:2:1 mixture of perlite, moss, and vermiculite and covered with a saran wrap to maintain the moisture content. The cups were then placed in a U-ING hydroponic grow box (Osaka, Japan) in a culture room at 22°C temperature, 45% relative humidity, and a 16h light/8h dark cycle. After germination, water and fertilisers were supplied to the seedlings according to the requirements. Seedlings were harvested at different days interval such as 4, 8, 12, and 20 days. The original seeds of the other ecotypes, including *Landsberg erecta (ler), nossen, Columbia-GD1 (col-gd1), Wassilewskija (WS),* and *Columbia-0* were obtained from Kyoto University (Kyoto, Japan).

3.2.2. RNA extraction and cDNA synthesis

Total RNA was isolated from different seedling samples at 4, 8, 12, and 20 days**,** using the Plant Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions given in the manual, and cDNA was synthesised subsequently using reverse transcriptase (Superscript III; Invitrogen, Carlsbad, CA, USA) and oligo dT primers. The sequences of the forward and reverse primers are listed in Table 1.

3.2.3. Gene transformation

Healthy *Arabidopsis* plants were grown until the flowering stage. All genes targeted for U-to-C RNA editing was cloned into a binary expression vector pRI-AN 101 digested with the restriction enzymes of *BamHI* and *EcoRI* and introduced into *Agrobacterium tumefaciens*, which was later transformed into *Arabidopsis* plants using the floral dip method [34]. Each plant transformed with edited gene was corresponded with the plants transformed with non-edited genes. The transformed Agrobacterium cells were grown at 30°C in liquid LB containing kanamycin (a selective marker). The culture was centrifuged, and cells were resuspended in sucrose solution (5%) containing 0.03–0.05% Silwet L-77 (optical density [OD] = 0.6). Flowers of *Arabidopsis* plants were dipped into Agrobacterium solution for 2–3s, with gentle agitation, and then covered for 16–24h to maintain high humidity. The dipping process was repeated twice (2–3 times per week). MS media containing kanamycin was used to confirm plant transformation. subsequently, the seeds were harvested from the transformed plants and incubated on MS medium for around a week.

3.2.4. Actinomycin D treatment

Col-0 seeds collected from the transformed plants were grown at 24°C in a growth chamber. A week-old seedling were treated with actinomycin D (Sigma-Aldrich; 100 mg/mL) for 0 (control), 1, 2, 4, 8, 12, 16, and 24hours. Samples were harvested at the designated time intervals and immediately frozen in liquid nitrogen. Total RNA was extracted from cell culture samples using the RNeasy Plant Mini Kit (Qiagen).

3.2.5. Gene expression analysis

cDNA was synthesised from 1µg of each RNA sample using the Superscript III cDNA Synthesis Kit. Then, RT-qPCR was performed using Stratagene Mx3000P and SYBR Green Supermix (Bio-Rad) under optimised conditions to minimise the primer-dimer formation and maximise the amplification efficiency. Here, we used CMV promoter of vector pRI-AN 101 for RT-PCR amplification to distinguish the transferred edited genes from endogenous Wild-type genes. The forward and reverse primers of the *ubiquitin C* (UBC) gene were used as housekeeping genes. The expression of each gene was analysed in duplicate, and transcript abundance at each time point was expressed relative to that at the 0 h time point. The glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) gene was used as a second reference gene using forward and reverse primers. The Ct value is the point at which the PCR curve crosses the threshold of the linear part of the curve. We used the Ct value to determine transcript abundance in this study. The higher the Ct value (30–35), the lower the mRNA quantity; this is because more cycles of amplification are needed to detect the fluorescence. A small Ct value (10–15) implies that the gene of interest is highly expressed.

3.3. Results

Total RNA was isolated from *Arabidopsis* ecotype Columbia (Col-0) seedlings harvested at 4, 8, 12, and 20 days, followed by cDNA synthesis and PCR. A previous study has identified various RNA editing events, including U-to-C, in targeted genes. The highest number of editing events were found in the seedlings and leaves as compared to those observed in stem tissues [22]. In another study, RNA-sequencing-based analysis of 12- and 20-day-old seedlings revealed the occurrence of U-to-C RNA editing in *Arabidopsis thaliana*. A total of seven genes, namely *AT2G16586*, *AT5G42320*, *AT1G05670*, *AT3G41768*, *AT4G32430*, *AT3G47965*, and *AT5G52530*, were

identified as targets for U-to-C RNA editing [8]. In the present study, we investigated the effects of such RNA editing on mRNA abundance of all above mentioned target genes. Recent discovery of novel types of DYW domains of PPR proteins, which are limited to species with U-to-C editing, imply that the domains are somehow involved in the amination of uridines in plant organelles [23]. Therefore, this study was focused on the PPR genes. In addition, we, herein, also report a new PPR gene (*AT2G19280*) targeted for U-to-C RNA editing.

To quantify the mRNA abundance of the target genes in Arabidopsis seedlings, we performed quantitative real-time PCR. The results showed that the amount of mRNA was degraded in seedling samples showing U-to-C RNA editing, compared to that observed for the seedlings showing no editing. The transcript abundances of *AT2G16586*, *AT5G42320*, *AT1G05670*, *AT3G41768*, *AT4G32430*, *AT3G47965*, and *AT5G52530* were analysed at different seedling stages such as 4-, 8-, 12-, and 20-day-old seedlings (Figure 3.1). The mRNAs were found to be significantly reduced in the seedling samples with U-to-C RNA edited genes. This supports the evidence for the effect of these edited nuclear genes on total mRNA abundance. The average transcript abundance of genes was calculated based on the difference between the expression of the gene of interest and the control gene, *ubiquitin C* (UBC), using the delta-delta Ct method. The ΔΔCt Ct value was found to be the lowest in those seedlings in which the U-to-C RNA editing was observed, indicating the least amount of mRNA in such seedlings.

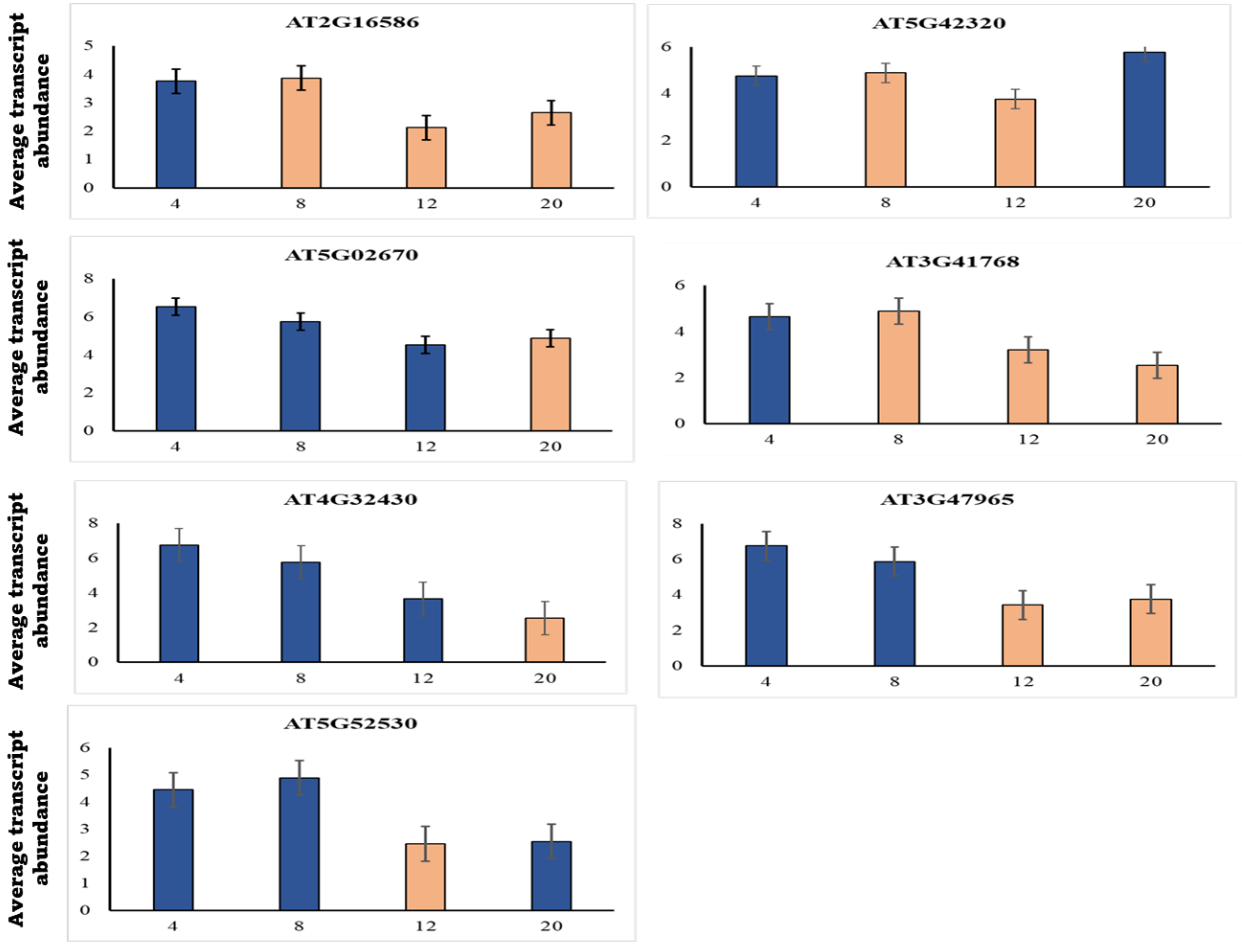


Figure 3.1: Quantitative real-time PCR (RT-qPCR) analysis of average transcript abundance in *Arabidopsis* seedlings at different seedling stages. Orange colour shows the samples with edited gene while blue shows the seedling samples with no editing. Average mRNA transcript abundances in 4-, 8-, 12-, and 20-day-old seedlings for the genes *AT2G16586*, *AT5G42320*, *AT1G05670*, *AT3G41768*, *AT4G32430*, *AT3G47965*, and *AT5G52530* are shown (n = 3)

The PPR gene *AT2G19280* was identified as the target for U-to-C RNA editing, specifically in 12-day-old seedlings. To confirm this further, this PPR gene was sequenced in 12- and 20-day-old *Arabidopsis* seedlings using both forward and reverse primers. The Sanger sequencing results revealed the presence of U-to-C RNA editing in 12-day-old seedlings but not in 20-day-old

seedlings (Figure 3.2 A and B). The U-to-C RNA editing site was detected at the position of 2972 nt in the 3′-UTR of mature PPR mRNA. The percentage of RNA editing was calculated based on double peak ratios. The U-to-C RNA editing was detected with 50%–60% editing efficiency in 12-day-old seedlings; however, no editing was detected in 20-day-old seedlings (Figure 3.2C). Furthermore, cDNA sequencing was also performed around 12-days, and U-to-C editing was observed in the 11–13-day-old seedlings. Sanger sequencing-based chromatograms showing U-to-C RNA editing in 11-, 12-, and 13-day-old *Arabidopsis* seedlings are shown in Figure 3.3, however, the highest editing efficiency peak was detected in 12-day-old seedlings, and no editing was detected in 8-, 9-, 10-, 14-, and 15-day-old seedlings (Figure 3.2D).

To quantify the amount of *AT2G19280* mRNA in *Arabidopsis* seedlings, we performed RT-qPCR. The results showed that the high amount of *AT2G19280* mRNA was degraded in 12-day-old seedlings as compared to that observed for 4-, 8-, and 20-day-old seedlings (Figure 3.2E). The transcript abundance of *AT2G19280* was also analysed in 9-, 10-, 11-, 13-, and 14-day-old seedlings (Figure 3.2F). This supports the evidence for the effect of the edited PPR gene *AT2G19280* in the total mRNA abundance. The average transcript abundance of *AT2G19280* was calculated based on the difference between the gene of interest and the control gene, *ubiquitin C* (UBC), using the delta-delta Ct method. The delta Ct value was found to be the lowest in 12-day-old seedlings, indicating the lowest amount of mRNA.

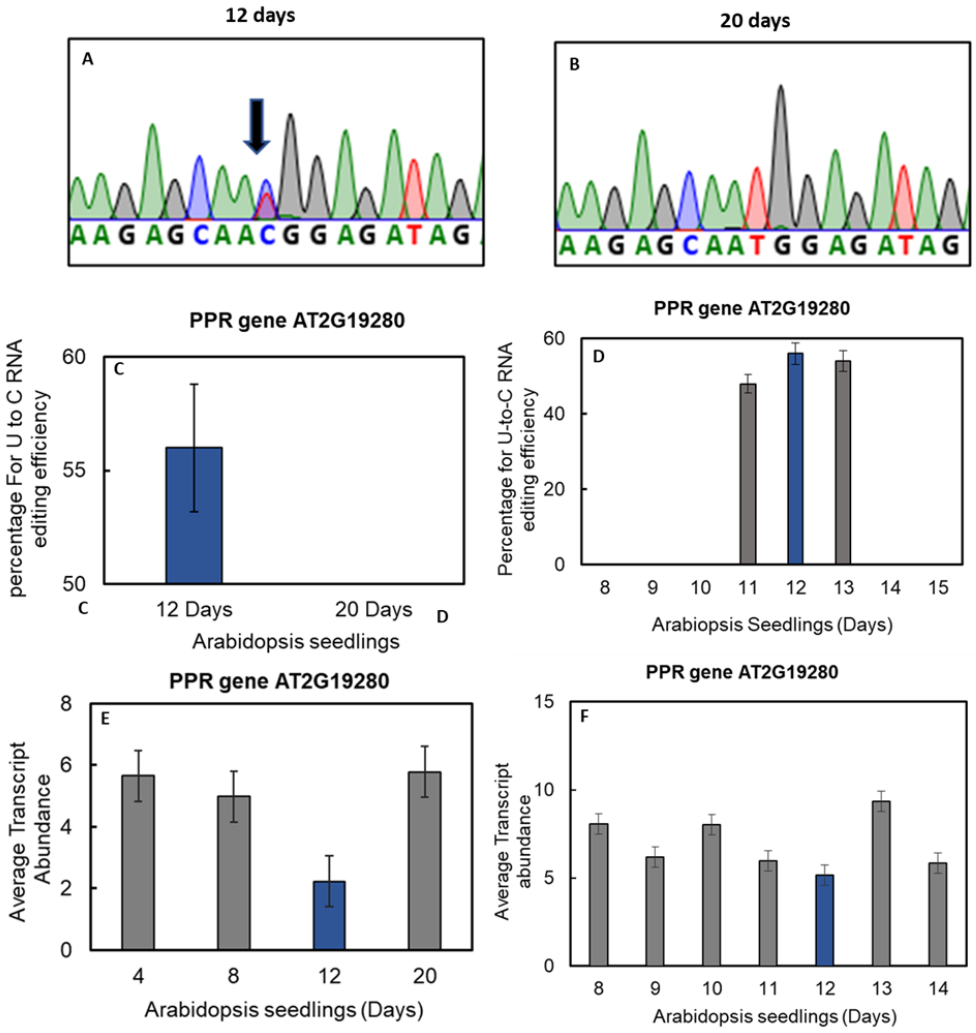


Figure 3.2: Analysis of U-to-C editing in the PPR gene *AT2G19280*. (A and B) Sanger sequencing-derived chromatograms showing U-to-C RNA editing sites in 12-day-old (A) and 20-day-old (B) seedlings. (C) Efficiency of U-to-C RNA editing (%) in 12- and 20-day-old *Arabidopsis* seedlings. (D) Efficiency of U-to-C RNA editing (%) in 8–15-day-old *Arabidopsis* seedlings. (E) Average mRNA transcript abundance in 4-, 8-, 12-, and 20-day-old seedlings. (F) Quantitative real-time PCR (RT-qPCR) analysis of average transcript abundance in *Arabidopsis* seedlings at different seedling stages.

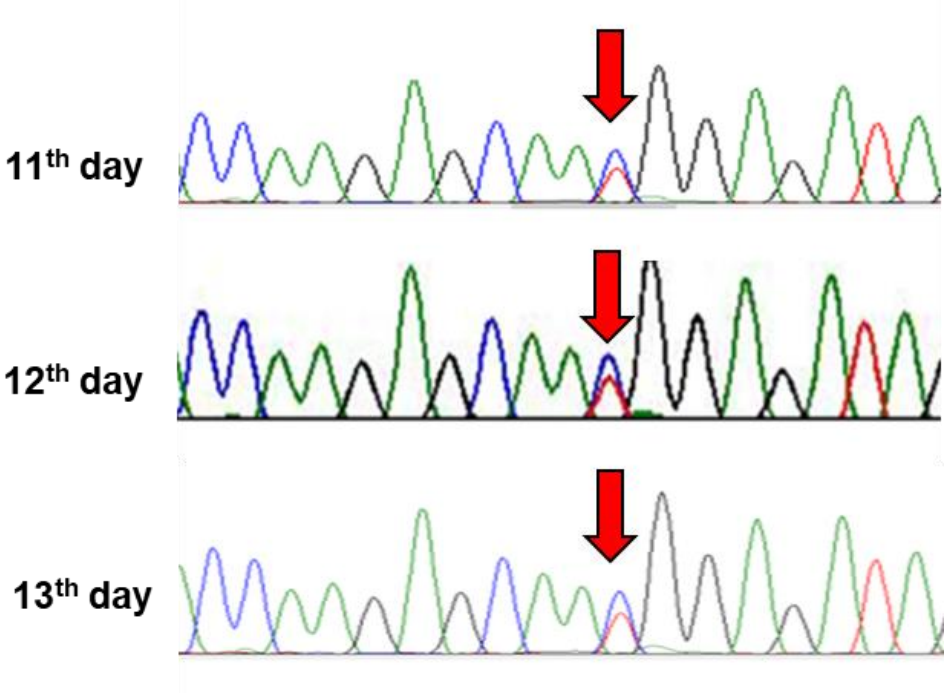


Figure 3.3. Sanger sequencing-based chromatograms showing U-to-C RNA editing in 11-, 12-, and 13-d-old Arabidopsis seedlings.

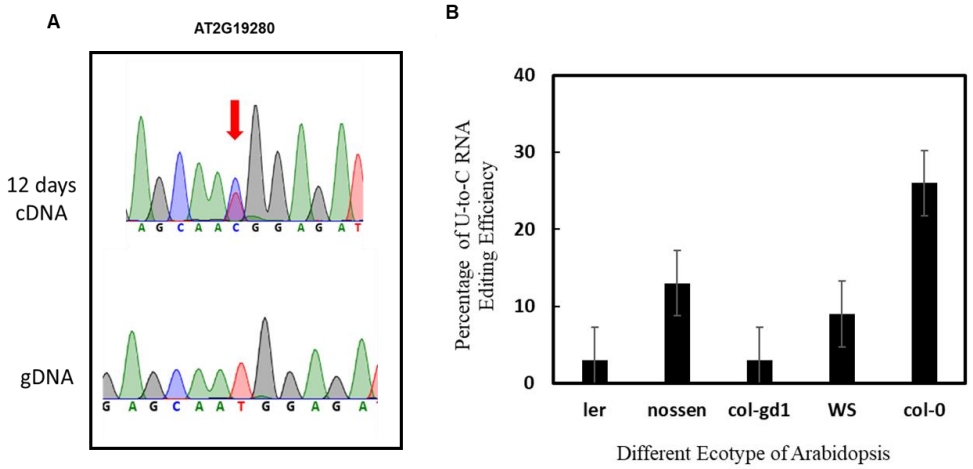


Figure 3.4. (A) Validation of RNA editing by comparing the cDNA and gDNA sequences from 12-day-old seedlings. (B) The U-to-C RNA editing efficiency (%) in different *Arabidopsis* ecotypes.

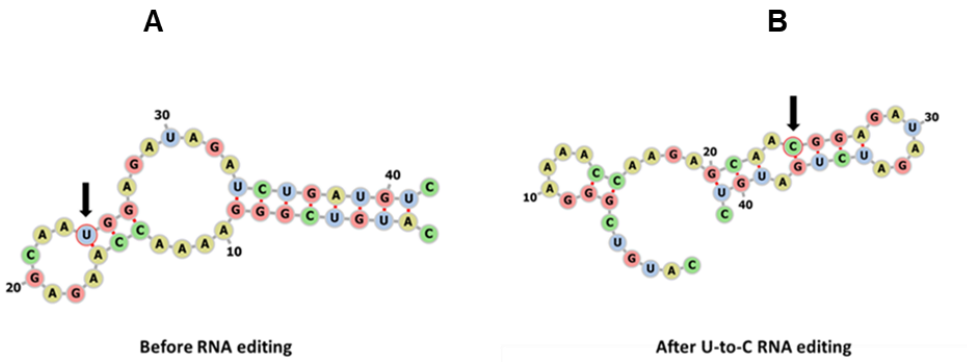


Figure 3.5. Effect of U to C RNA editing on the secondary structure of mRNA. (A and B) Secondary structure of mRNA before (A) and after (B) U to C RNA editing.

Further, we validated the RNA editing site experimentally by comparing the Sanger sequencing carried out for both the genomic DNA (gDNA) and synthesised cDNA. We extracted DNA and mRNA from the same aliquot obtained from the sample of 12-day-old seedlings. By sequencing the paired DNA and cDNA samples and carrying out independent analysis of each chromatogram by two individuals, we confirmed the U-to-C RNA editing site (Figure 3.4A). The cDNA showed a double peak, representing T and edited C nucleotides, whereas no double peak was observed in gDNA sequencing. Sequencing was performed using a sense primer that targeted the editing site. No editing was identified at the gDNA level, indicating that the event was a post-transcriptional modification. The different ecotypes, including *Landsberg erecta (ler), nossen, Columbia-GD1 (col-gd1), Wassilewskija (WS), and Columbia-0* from different geographical ecosystems were also examined, and the percentage of RNA editing efficiency was calculated (Kyoto University, Kyoto). The editing efficiency was observed to be the highest for ecotype Columbia-0 (Figure 3.4 B). To determine the effects of U-to-C editing on the function of the PPR gene *AT2G19280*, we predicted the secondary structure of its mRNA using an online software tool, mFold. The secondary structures of *AT2G19280* mRNA before and after U-to-C RNA editing in the 3′-UTR are shown in Figure 3.5. It was observed that before editing, ‘U’ was located at the

bulge loop of *AT2G19280* mRNA; however, the edited 'C' was found to be placed in the hairpin stem and complementary paired with 'G', after the editing.

Flowering plants were transformed separately with the edited and wild-type genes. To determine the transcript decay rate in *Arabidopsis*, the transcription inhibitor actinomycin D (Act D) was added to the seedling samples of transformed *Arabidopsis*. A schematic representation of the treatment with Act D for analysing the mRNA stability using Agrobacterium-mediated transformed plants is shown in Figure 3.6. MS media containing kanamycin were used to confirm the gene transformation. After the confirmation, seeds were harvested from transformed plants and grown in MS culture plates for a week. After adding Act D, the mRNA abundance decreased significantly in the plants transformed with the edited genes. The overall transcript mRNA decay rate of the edited genes was higher than that of the wild-type gene (Figure 3.7). Effective transcription inhibition using Act D was further validated by the reduction in cDNA synthesis from total RNA (Figure 3.8A). Analysis of cDNA synthesised from samples treated with Act D for 0 (control), 1, 2, 4, 8, 12, 16, and 24 h using the polyacrylamide gel electrophoresis (PAGE) indicated that bands were intact in samples treated with Act D for up to 16h; however, no bands were detected at later time points (Figure 3.8 B).

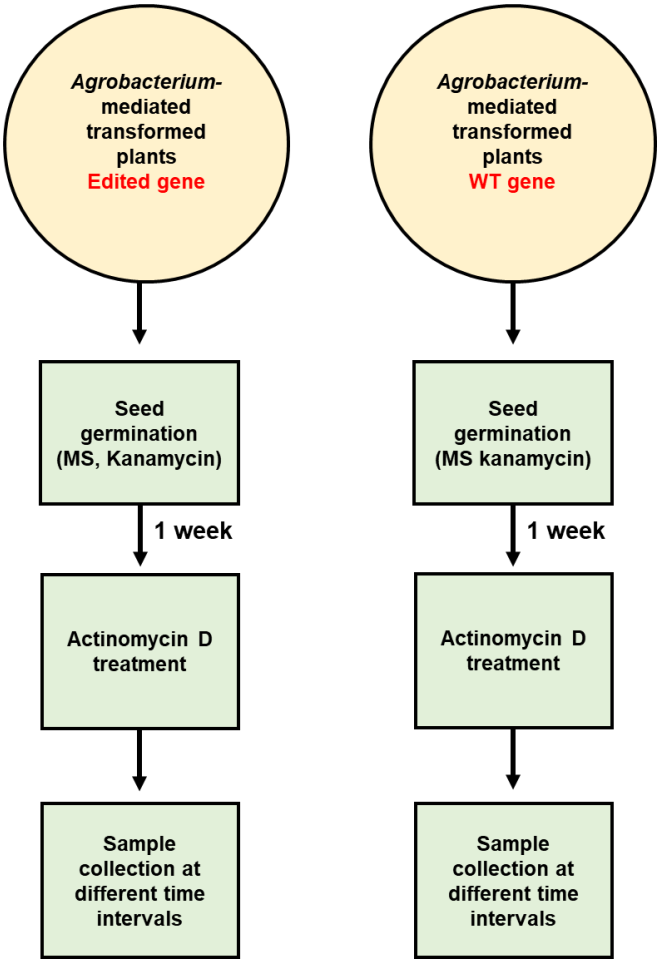


Figure 3.6. Schematic representation of the treatment with the transcription inhibitor, Actinomycin D (Act D) for mRNA stability analysis using *Agrobacterium*-mediated transformed plants. Growth of Agrobacterium-mediated transformed plants on MS media containing selective marker kanamycin indicates the insertion of the gene of interest into the plant cells.

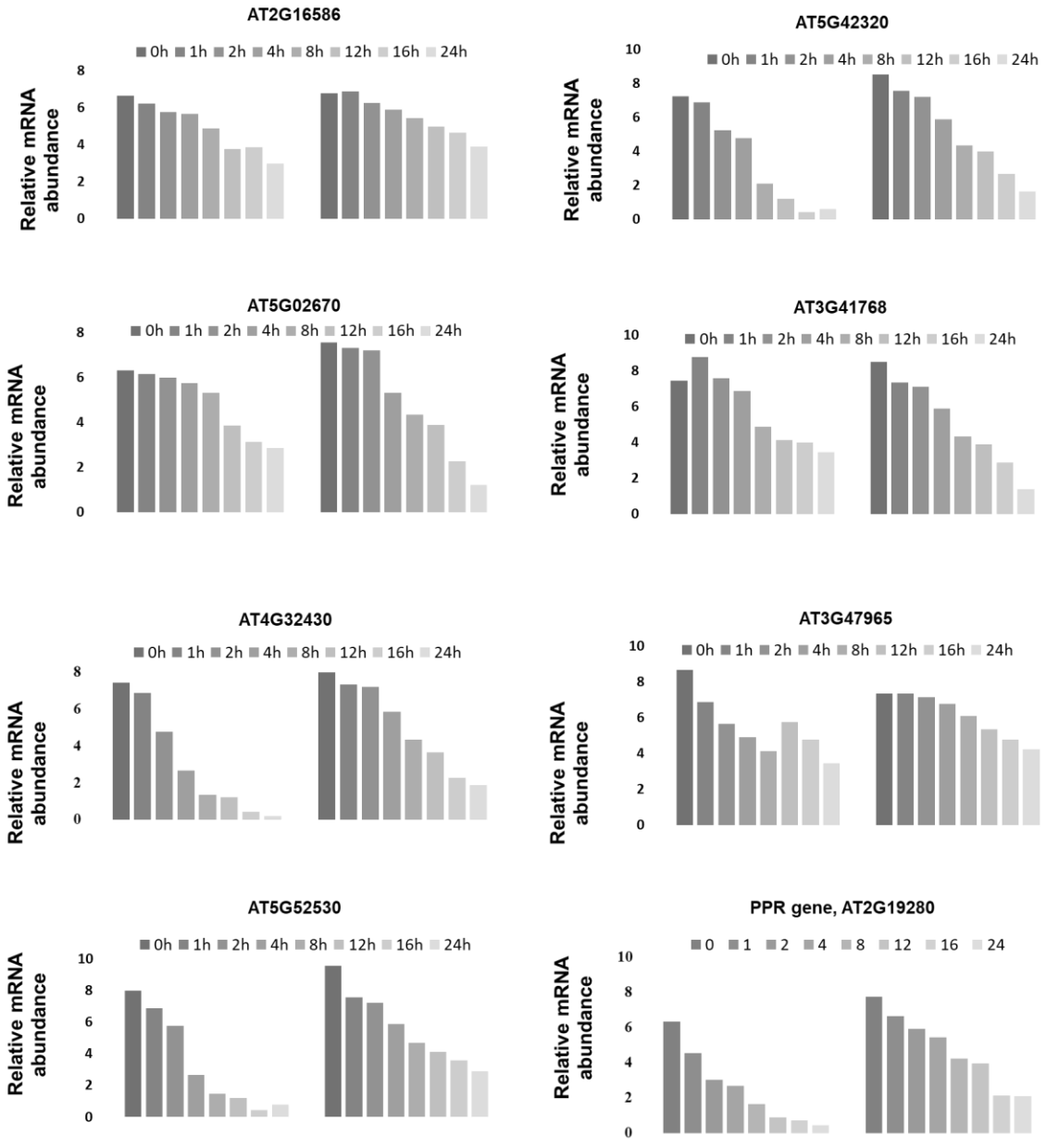


Figure 3.7. Effect of U-to-C RNA editing on mRNA stability (A-H). Transcript decay rates of wild-type and edited genes in *Arabidopsis* seedlings transformed using the floral dip method. Transcript abundances were analysed by RT-qPCR the relative mRNA expression was calculated for each studied gene (n = 3).

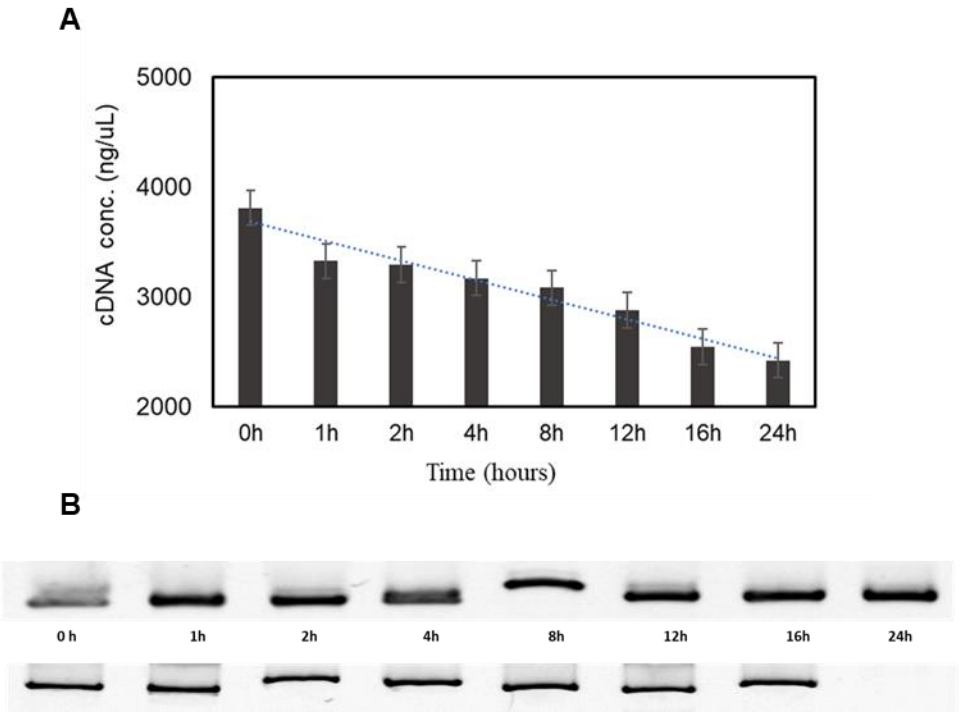


Figure 3.8. Analysis of mRNA stability after adding the Actinomycin D. A. cDNA concentration (ug/uL) after adding Actinomycin D. **B.** Analysis of cDNA by gel electrophoresis before ActD treatment (0h) and 1,2,4, 8, 12,16, and 24h after ActD treatment (Upper line, wild type gene and lower line, edited gene).

3.4. Discussion

To our knowledge, this is the first report of the effect of U-to-C RNA editing on the mRNA stability of nuclear genes in *Arabidopsis*. In our previous study, we have identified seven genes, namely *AT2G16586*, *AT5G42320*, *AT1G05670*, *AT3G41768*, *AT4G32430*, *AT3G47965*, and *AT5G52530*, as targets for U-to-C RNA editing. The UTRs of genes encoding proteins involved in RNA metabolism and RNA binding, including PPR proteins, Zn-finger (ZnF)-related proteins, ribosomal protein L2, transmembrane proteins, and two hypothetical proteins, were identified as targets of U-to-C editing. Since approximately 50% of the ribosomal RNA, *AT3G41768*, were affected by editing, it might have a significant effect on their functions. Similarly, the mRNA for transmembrane protein AT2G16586 was identified with 77.3% of U-to-C RNA editing efficiency that may affect its general biophysical characteristics. In addition, the PPR gene, *AT4G32430*, was also identified to have 20.43% U-to-C RNA editing. Thus, in the present work, we demonstrated their effects on mRNA abundance and a decrease in the stability was observed in most of the target genes.

We also confirmed the rare occurrence of U-to-C RNA editing events in the gene, AT2G19280, by comparing the Sanger sequencing of gDNA and cDNA in 12-day-old seedlings of *Arabidopsis* plants. The U-to-C RNA editing (50%) was detected at 2972 nt in the 3′-UTR of the mRNA transcript obtained from 12-day-old seedlings of *Arabidopsis thaliana*. RNA editing events have been observed to be more common in seedlings than in any other plant tissues [22]. Seedlings are considered very important in plant physiology.

In other studies, based on *Arabidopsis*, bioinformatics approaches have revealed various RNA editing events, which has also been described in nuclear transcripts [24,25]. The tissue-based analysis of RNA editing indicated that these events were post-transcriptional alterations, not genomic DNA

mutations. Previous findings suggest that RNA editing sites are essentially recognised by a neighbouring cis-regulatory element of RNA-binding PPR-associated enzyme in plants [26]. Future studies are required to better understand the mechanisms involved in RNA editing, such as isolation of editing enzymes, identification of cis-regulatory elements, and RNA-sequencing-based validation of the editing sites. RNA editing in the UTRs affects mRNA stability due to changes in their secondary structures [25,27–29]. Moreover, RNA editing at some sites has an unfavourable effect on plant development, growth, and fertility [30]. The PPR gene, *AT4G32430,* was identified with 20.43% U-to-C RNA editing in the 3′-UTR of 20-day-old seedlings [8]. The mRNA stability of the PPR gene, *AT4G32430,* was also found to be affected by U-to-C RNA editing (Figure 4E). We identified RNA editing events in the UTR regions of the PPR genes, which affected mRNA stabilities. The editing in the PPR gene *AT2G19280* was not reported in RNA-seq data [8], as the gene was expressed in very low amount to be counted as successful reads. However, to confirm whether the identified RNA editing site was truly positive, we searched for evidence of the identified RNA editing site in *Arabidopsis* RNA-seq data generated by public laboratories, using an online software (http://signal.salk.edu/atg1001/3.0/gebrowser.php). The PPR gene *AT2G19280* was aligned against publicly available RNA-seq databases to confirm our findings. The target T sites were identified as edited C sites in various databases (Figure S6). In this study, we examined U-to-C RNA editing in *Arabidopsis* seedlings at different days. Sanger sequencing was performed to identify the sites and efficiency of U-to-C editing. U-to-C RNA editing, which is rare, occurs in the 3′-UTR of mature mRNAs. This region of the mRNA also serves as the site for other types of RNA editing, including C-to-U and U-to-G editing.

RNA editing also affects the gene expression levels. Transcript abundance is determined by the rates of mRNA synthesis and degradation caused by polymerase and nuclease, respectively [31]. A

comprehensive study of mRNA decay rates in *Saccharomyces cerevisiae* demonstrated that changes in mRNA abundance following the inhibition of transcription were found to be equal to those following a heat shock [32]. The study also revealed that factors controlling ribosome biogenesis are regulated at the post-transcriptional level [32]. These findings confirmed that mRNA transcript decay rates may provide insights into different biological activities that are not identified based only on transcript abundance. In this study, mRNA stability was estimated by Agrobacterium-mediated transformation of *Arabidopsis* culture. Transcript abundance was determined using qRT-PCR. The average transcript abundance was calculated using the delta-delta Ct method, based on the differences in the expression levels of the test and control genes. The delta Ct value was found to be the lowest in around 12-day-old seedlings, indicating the presence of the least amount of mRNA. We also compared the decay rate of the edited mRNA with that of wild-type mRNA. The results showed that U-to-C RNA editing greatly affected the mRNA transcript abundance due to modification of the secondary structure of mRNA in the edited PPR gene, *AT2G19280 and AT4G32430*. The fact that mRNA abundance decreased in plants, suggests a decrease in the stability of the edited RNA. Thus, our results suggest that U-to-C RNA editing of nuclear genes, particularly of the PPR gene, has a considerable impact on plant stability.

3.5. Conclusions

We, herein, have confirmed the effect of U-to-C RNA editing on mRNA abundance in *Arabidopsis*, a flowering plant. A rare occurrence of U-to-C editing was identified exclusively in young seedlings, indicating that the enzyme essential for this editing event might be expressed around 12 days only. This U-to-C RNA editing sites were mostly found in the untranslated region (3' UTR) of the mature mRNA and affected its secondary structure. We demonstrated the correlation between the U-to-C RNA editing-related genes, especially the pentatricopeptide repeat (PPR) gene, and their roles, such as alteration in secondary structure and mRNA abundance of edited genes through seed culture of transgenic Arabidopsis generated by *Agrobacterium*-mediated transformation. Thus, we concluded that U-to-C RNA editing adversely affects the plant stability.

3.6. References

1. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, (1970).
2. Stern, D. B., Goldschmidt-Clermont, M. & Hanson, M. R. Chloroplast RNA Metabolism. *Annual Review of Plant Biology* **61**, (2010).
3. del Campo, E. M. Post-Transcriptional Control of Chloroplast Gene Expression. *Gene Regulation and Systems Biology* **3**, (2009).
4. Stuart, K. D., Schnaufer, A., Ernst, N. L. & Panigrahi, A. K. Complex management: RNA editing in trypanosomes. *Trends in Biochemical Sciences* **30**, (2005).
5. Knoop, V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cellular and Molecular Life Sciences* **68**, (2011).
6. Maas, S. Posttranscriptional recoding by RNA editing. in (2012). doi:10.1016/B978-0-12-386497-0.00006-2.
7. Gerke, P. *et al.* Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*. *New Phytologist* **225**, (2020).
8. Ruchika, Okudaira, C., Sakari, M. & Tsukahara, T. Genome-Wide Identification of U-To-C RNA Editing Events for Nuclear Genes in Arabidopsis thaliana. *Cells* **10**, (2021).
9. Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. & Brennicke, A. RNA Editing in Plants and Its Evolution. *Annual Review of Genetics* **47**, (2013).
10. Kotera, E., Tasaka, M. & Shikanai, T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* **433**, (2005).

11. Zehrmann, A., Verbitskiy, D., van der Merwe, J. A., Brennicke, A. & Takenaka, M. A DYW Domain–Containing Pentatricopeptide Repeat Protein Is Required for RNA Editing at Multiple Sites in Mitochondria of *Arabidopsis thaliana* . *The Plant Cell* **21**, (2009).

12. Barkan, A. & Small, I. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology* **65**, (2014).

13. Lurin, C. *et al.* Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis[W]. *The Plant Cell* **16**, (2004).

14. Cheng, S. *et al.* Redefining the structural motifs that determine <scp>RNA</scp> binding and <scp>RNA</scp> editing by pentatricopeptide repeat proteins in land plants. *The Plant Journal* **85**, (2016).

15. Xing, H. *et al.* Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. *Scientific Reports* **8**, (2018).

16. SCHMITZLINNEWEBER, C. & SMALL, I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science* **13**, (2008).

17. Doniwa, Y. *et al.* The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. *Gene* **454**, (2010).

18. Schallenberg-Rüdinger, M., Lenz, H., Polsakiewicz, M., Gott, J. M. & Knoop, V. A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. *RNA Biology* **10**, (2013).

19. Chateigner-Boutin, A.-L. *et al.* The E domains of pentatricopeptide repeat proteins from different organelles are not functionally equivalent for RNA editing. *The Plant Journal* **74**, (2013).

20. G.H. Luo, X. H. L. Z. J. H. Z. C. Z. Q. Y. H. F. G. J. C. F. Transition and transversion mutations are biased towards GC in transposons of Chilo suppressalis. *Genes* **7**, 1–12 (2016).

21. Shikanai, T. RNA editing in plants: Machinery and flexibility of site recognition. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1847**, (2015).

22. Qulsum, U., Azad, Md. T. A. & Tsukahara, T. Analysis of Tissue-specific RNA Editing Events of Genes Involved in RNA Editing in Arabidopsis thaliana. *Journal of Plant Biology* **62**, (2019).

23. Oldenkott, B., Yang, Y., Lesch, E., Knoop, V. & Schallenberg-Rüdinger, M. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in Escherichia coli. *Communications Biology* **2**, (2019).

24. Meng, Y. *et al.* RNA editing of nuclear transcripts in Arabidopsis thaliana. *BMC Genomics* **11**, (2010).

25. Wu, B. *et al.* Identification of Symmetrical RNA Editing Events in the Mitochondria of Salvia miltiorrhiza by Strand-specific RNA Sequencing. *Scientific Reports* **7**, (2017).

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

27. Drescher, A. *et al.* 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**, (2002).

28. Zeng, W.-H., Liao, S.-C. & Chang, C.-C. 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**, (2007).

29. Farré, J.-C., Aknin, C., Araya, A. & Castandet, B. RNA Editing in Mitochondrial Trans-Introns Is Required for Splicing. *PLoS ONE* **7**, (2012).

30. Hammani, K. & Giegé, P. RNA metabolism in plant mitochondria. *Trends in Plant Science* **19**, (2014).

31. E. Yang, E. van N. M. Z. et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).

32. Narsai, R. *et al*. Genome-Wide Analysis of mRNA Decay Rates and Their Determinants in *Arabidopsis thaliana*. *The Plant Cell* **19**, (2007).

Chapter 4

Development of RNA editing system by expressing Ho rnworts specific "DYW-type" PPR proteins

4.1. Introduction

RNA editing describes as a post-transcriptional modification that change bases in mRNA sequences relative to their corresponding genomic DNA sequences[1]. There are different types of RNA editing events including the deamination of adenosine (A) to inosine (I) residues in animals [2], bacteria [3,4], or fungi [5], insertions or deletions of uridine (U) in mRNA sequences in the mitochondria of protozoa [6], the conversion of specific cytidines (C) to uridine [7], or the insertion of guanosine (G) residues in some viruses [8,9]. In plants, RNA editing was first discovered in the mitochondria of following plants [10–12], as well as in chloroplasts [13–15]. There are mainly two types of substitution RNA editing in plant mitochondria and chloroplast; C to U editing, which is generally found in all land plants while the other type is "reverse" U to C alteration, that are considered to be restricted in some ferns and hornworts. Recently, genome-wide identification and characterization of U-to-C RNA editing events has been reported in the nuclear genes of *Arabidopsis thaliana* [16,17]. In animal, A-to-I and C-to-U RNA editing are generally caused by ADARs and APOBEC-AID deaminase family, respectively [18–22]. However, the enzymes responsible for U-to-C editing was not discovered yet, although it is the abundant phenomenon in lower plant species [23]. In genetics, a nonsense mutation is a point mutation in a sequence of DNA that results in a premature stop codon (UAA, UAG, UGA), or a nonsense codon in the transcribed mRNA, and in a truncated, incomplete, and usually nonfunctional protein product. If U-to-C editing become possible, we can treat many diseases caused by nonsense mutations such as Duchenne muscular dystrophy, cystic fibrosis, and hemophilia, and also frequent diseases such as cancers, metabolic disorders, and neurological disorders. Therefore identification of enzymes responsible for U-to-C mutation could lead to therapies that treat genetic disease by restoring or removing the stop codon sequences at the mRNA level.

Genetical approaches with model plants e.g. *Arabidopsis*, Rice, *Physcomitrella patens* revealed that specific editing site recognition in plant organellar RNA editing is governed by PPR proteins with C-terminal E or EDYW domain extension[24]. The C-terminal DYW has been considered to work as a catalytic domain for C-to-U editing due to the high similarity with cytidine deaminases. This hypothesis was strongly supported by the recently reported in *E. coli* C to U RNA editing with expression of a single DYW containing PPR protein from *P. patens* [25] as well as in vitro RNA editing [26]. Furthermore, structural analysis of DYW domain verified its cytidine deaminase conformation with a Zn ion associated catalytic center [27,28]. Molecular mechanism of U to C editing is, however, not yet clear since there are no suitable model plants having a system for U-to-C RNA editing in organelles. Recent studies on sequence-based analysis of the hornworts discovered that the PPR proteins in this species comprise unique C-terminal DYW-like domains with distinguished signatures. These domains are believed to be the strongest candidates for the U-to-C RNA editing, since such domains were not observed in other model plants having only C-to-U RNA editing [29,30]. The high number of proteins with deviant DYW domains is intriguing in the light of the high amount of reverse U-to-C RNA editing that could identify. Naturally, the characteristic "DRH" and "GRP" DYW domain variants could be the attractive candidates to represent factors for reverse U-to-C RNA editing. The U-to-C RNA editing has originated from the more ancient and widespread C-to-U editing, using the same mechanisms for RNA target recognition linked to a biochemical enzyme variant, possibly converting a deaminase into a transaminase. Given the likely earlier evolutionary origin of plant C-to-U RNA editing among land plants it is suggestive that PPR proteins remain at the core of target recognition also for sites of U-to-C editing.

In our present work, we have investigated the RNA editing factors by cloning the hornwort specific PPR proteins into bacterial cells. In E. coli, expression of the two PPR proteins called PPR65 and PPR56 from the moss *Physcomitrella patens* leads to the editing of co-expressed respective targets from *ccmFC* and *nad3/nad4* transcripts [31]. So far only these two moss PPR proteins are available for in E.coli RNA editing. We cloned the three hornworts PPR proteins C-terminal variants GRP, DRH, and DYW fused with the *P. patens* pentatricopeptide repeat (PPR) protein PPR56 and evaluated its cytidine deaminase activity and opening a new research area for investigating the U-to-C RNA editing related DYW-type domains in the molecular biology field.

4.2. Methods and materials

4.2.1. Plasmid construction

For bacterial expression system, the *P. patens* PPR56 coding sequence with modified C-terminal DYW variants from hornworts is inserted into the pETG_41K vector system resulting in the fusion to a His6-tagged maltose binding protein, and the RNA editing target sequence sites is cloned downstream by in-fusion cloning technique. In animal cells, to enable directing the enzymes to a target editing sites, we cloned the deaminase/transaminase domain of PPR into pCS2+only expression vector using the *BamHI* and *XhoI* (Takara, Shiga, Japan) restriction enzymes to yield pCS2-PPR56-DYW. The domain was amplified by PCR from complementary DNA of hornworts using forward and reverse primers harboring the appropriate restriction sites. Primers were designed using the software DNADynamo. Lists of selected genes are given in Table 4.1.

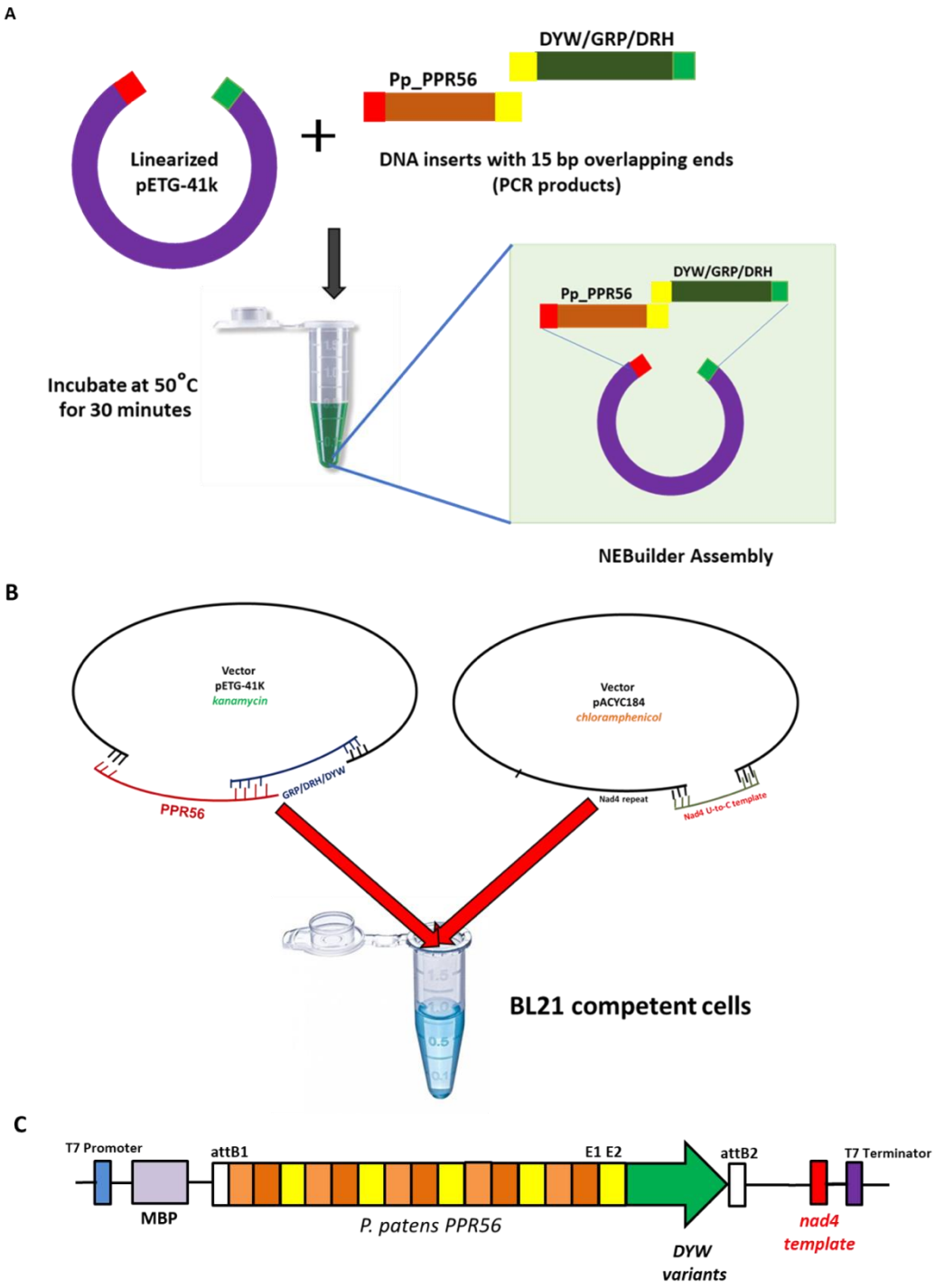


Figure 4.1. Schematic representation for the experimental methodology. **A.** NEBuilder cloning for the Plasmid construction. **B.** Co-transformation of editing factor PPR56 and the target nad4 template into BL21 competent cells, expressed by IPTG induction. **C.** Diagramatic

illustration for cloning the PPR56, editing factor and fusion to a His6-tagged maltose binding protein, and the RNA editing target sequence sites is cloned downstream by in-fusion cloning technique

Table 1: List of Hornworts specific PPR- DYW-type candidates studied genes.

S.No.	Given gene name	Annotation	Primer sequences	
			Forward	**Reverse**
1.	g63	PPR-GRP	AAGAAACCTGCAGTGGCG	AGGTCGACCCTTACAACTGC
2.	g18202	PPR-GRP	AAGAAACCTGCAGTGGCAAC	CGGTCGACCCTTACAACTG
3.	g16507	PPR-GRP	AAGAAACCTGCAGTTACGACCA	CGGTCGACCCTTGCAACT
4.	g17021	PPR-DRH	AAGAAGGCTGCCAAGGCAT	ATGACGATCCCTACATGAAC
5.	g17318	PPR-DRH	AAGGCTGCAAAGGCGTG	ATGTCTATCGCCACATGAAC
6.	g18409	PPR-DRH	AAGCCCGCCAAAGCGTGC	ATGCCTATCTCTGCAAGAAC
7.	g18545	PPR-DRH	AAGAAACCTGCCAAGGCTTG	GTGCCGATCCCTACATGAAC
8.	g10823	PPR-DRH	AAGAAGCCAGCCAAGGCAT	GTGGCGATCCCCACAAGAGC
9.	g16688	PPR-DYW	AAGAAGAATCCAGGGTGCAG	CCAATAATCTCTGCAAGAGC
10.	g10204	PPR-DYW	GGGTACACGTGGATTGAGCT	CCAATAATCACCGCATGTGC

Annealing temperature 58degree, 35 thermal cycles. Forward adapter AATACCATATTTTTGAAA and reverse adapter AAGTTGCGGCCGCAC

4.2.2. Target sequence:*nad4* gene:

RNA editing target sequences of *nad4* U-to-C template (including at least 33bp upstream and 5bp downstream of the editing site) flanked by appropriate restriction sites *sphI* and *SalI* were generated by primer annealing at 50°C with two thermal cycles and cloned into the vector. This sequence is original PPR56 target sequence but target site is changed from C to T. The target

sequence includes the target sites for both C-to-U and U-to-C types of RNA editing. The target sequence is cloned in pACYC184 vector.

For U to C

NNNNNNNNNNNNNNNNNNNNTCGGGCTCNTGAGCGCTTGTTTCGGCGTGGGTATG

GTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTGCATGCTCAAACA

TCAATTTTTATATAGGTATAGACGGTATCTCTTTATTTTGTCGACCGATGCCCTTGAG

AGCCTTCAACCCAGTCACNCCTTCCGGNTN

Primer sequence
>nad4FSphIF
GGCGCCATCTCCTTGCATGCTCAAACATCAATTTTTATATAGGTATAGACGGTATCT

pink : extended adapter from PPR56 end (15 bp)
Yellow : Sph1 restriction site
Grey: forward primer

>pACYC184_SalI_R
ACCGGAAGGAGCTGACTG

For C to U target

Targeted C is highlighted in red

GGCCTCTTGCGGGATATCTCAAACATCAATTTTTATATAGGTATAGACGGTATCTCTT

CATTTTGACGTCTCAAACATCAATTTTTATATAGGTATAGACGGTATCTCTTCCTTTT

AGATCTTCAAACATCAATTTTTATATAGGTATAGACGGTATCTCTTCGTTTTTGTACA

TCAAACATCAATTTTTATATAGGTATAGACGGTATCTCTTCTTTTTATGCATTCAAAC

ATCAATTTTTATATAGGTATAGACGGTATCTCTTCATTTTGGTACCTCAAACATCAAT

TTTTATATAGGTATAGACGGTATCTCTACATTTTCTGCAGTCAAACATCAATTTTTAT

ATAGGTATAGACGGTATCTCTCCATTTTCGATCGTCAAACATCAATTTTTATATAGGT

ATAGACGGTATCTCTGCATTTTTACGTATCAAACATCAATTTTTATATAGGTATAGA

GTATCTCTTCATTTTGGATCCTCTACGCCGG

I use below forward primer sequence for the sequencing.
>PACYC184_EB_F
AATCTAACAATGCGCTCAT

>pACYC184_SalI_R
ACCGGAAGGAGCTGACTG

4.2.3. Co-transformation

Constructs containing editing factors and downstream targets were co-transformed into BL21, Rosetta 2 (DE3) competent cells. 1µL of each plasmid were added into 50 µL bacterial cells. Mixed well, and incubated on ice for 30 minutes. Then heat shock at 42°C for 40 seconds and again on ice for 2 mins. Added 400 µL of LB media and incubated for 60 minutes at 37 °C and 300 rpm. Plating the cells on agar plates of Luria Broth with 50µM kanamycin and 30µM chloramphenicol), incubated at 37 °C overnight. Both constructs were amplified with respective primer sets and observed through gel electrophoresis.

4.2.4. IPTG induction

5mL E. coli starter cultures (Luria Broth with 50µM kanamycin and 30µM chloramphenicol) were grown overnight. 40 µL of the pre-culture were used to inoculate 4mL of the same media. Cultures were grown at 37°C until an OD600 of 0.4–0.6 was reached. Cultures were cooled on ice for a minimum of 5min before adding 0.4mM IPTG for induction of construct expression supplemented with 0.4mM ZnSO4. Cells were incubated at 16 °C and 180rpm for 20h before harvesting 2mL samples. Samples were frozen in liquid nitrogen and stored at −80°C until further use.

4.2.5. Cell culture and transfection

About 3 × 105 cells per well were seeded in 12-well culture plates (Costar, Corning, NY, USA), grown for 24 h to 50–70% confluence, and then subjected to transfection. Cell culture medium was Dulbecco's Modified Eagle Medium with high glucose (WAKO, Tokyo, Japan) supplemented with 10% fetal bovine serum (GIBCO, Invitrogen, Canada). Lipofectamine 3000 (Invitrogen, Carlsbad, CA, USA) was used for transfections.

4.2.6. Detection of RNA editing in E. coli.

Total RNA was extracted from *E. coli*. cells using a kit system (purelink, Invitrogen). RNA was eluted with RNase-Free water and complementary DNA (cDNA) was synthesized with a random primer with ReverTra Ace qPCR RT Master Mix. A reverse primer upstream of the T7 terminator stem-loop sequence and a forward primer binding in the PPR protein coding region were used for reverse transcription-PCR amplification. PCR amplification assays contained 1µL template of cDNA, 0.4µM of each primer, 1× recommended PCR GoTaq buffer master mix, (Takara), and double distilled water in a final volume of 25µL. Amplification assays included 5min initial denaturation at 95°C followed by 35 cycles each with 30s denaturation at 95°C, 30s annealing at 61°C, 2min synthesis at 72°C, and a final step of synthesis for 7min at 72°C. For purification of

PCR products, 2U ExoI (TAKARA) and 0.5U Shrimp Alkaline Phosphatase (TAKARA) were added and incubated at 37 ℃ for 1h followed by 15min at 80 ℃ and sequenced by europhin genomics.

4.2.7. Quantitative RT-PCR Analysis

cDNA was prepared from 1 ug of each RNA sample using the superscript III cDNA synthesis kit, and quantitative RT-PCR analysis was performed using an bioanalyzer instrument and SYBR Green Supermix (Bio-Rad) under conditions optimized to maximize amplification efficiency and minimize primer-dimer formation, using GAPDH primers as housekeeping genes. For every transcript, each cDNA sample was analyzed in duplicate, and transcript abundance was expressed as a ratio relative to control sample.

4.3. Results

4.3.1. Hornworts specific PPR proteins

I studied the conserved sequences for thousands of PPR proteins and found that the PPR proteins with C-terminal DYW-like domains in the Hornworts, *Anthoceros* are with the characteristic alterations in their DYW domain signatures ("DYW-type", "DRH-type" and "GRP-type")[29,30,32]. We have identified that, out of thousands of studied PPR genes, about 50% were identified with C-terminal motifs, the GRP-type genes dominate in number (445), which is followed by the DRH-type genes (175), and very small number of the DYW-type (6). The percentage division of each PPR proteins type DYW variants are shown in Figure 4.2. Numerous characteristic alterations in conserved positions along the entire DYW domain are observed among the KPAxA_DRH and the KPAxA_GRP-type DYW proteins identified in *Anthoceros agrestis* [30].

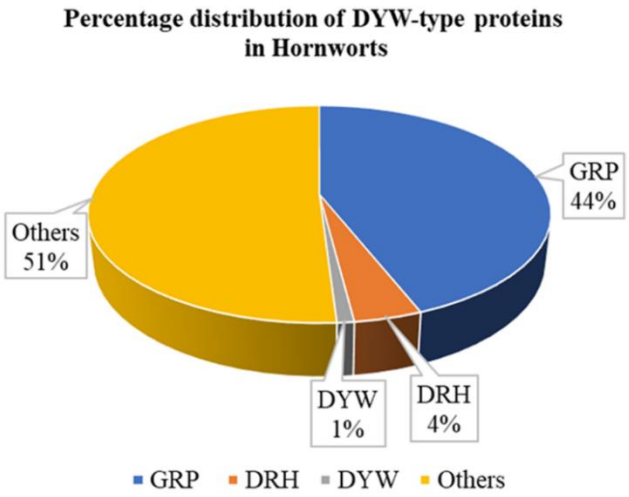


Figure 4.2. Percentage distribution of expansion and diversification of RNA Editing gene family of "DYW-type" PPR proteins in the Hornworts.

Recently study presented structures and functional data of a DYW domain in an inactive ground state and activated. DYW domains harbour a cytidine deaminase fold and a C-terminal DYW motif, with catalytic and structural zinc atoms, respectively. A conserved gating domain within the deaminase fold regulates the active site sterically and mechanistically in a process that we termed gated zinc shutter [33]. Amino acid sequence of these DYW, DRH and GRP variants show about 75-85% similarity. While catalytic site with Zn1 is relatively highly conserved, C-terminal DYW motif is not. Indeed, when we have predicted the theoretical structure for these variants using an online software AlphaFold2 [34], analyzed by PyMOL, https://pymol.org/2/, they seem to fold different way especially at the C-terminal DYW-motif (figure 4.3). Therefore, detailed structural configuration and comparative studies investigated by the protein NMR or X-ray crystallography will be also necessary for understanding the difference between canonical

DYW domain and DYW:KP domains. The sequence logo for amino acid for the variants are generated by online platform Weblog (Crooks GE, 2004)

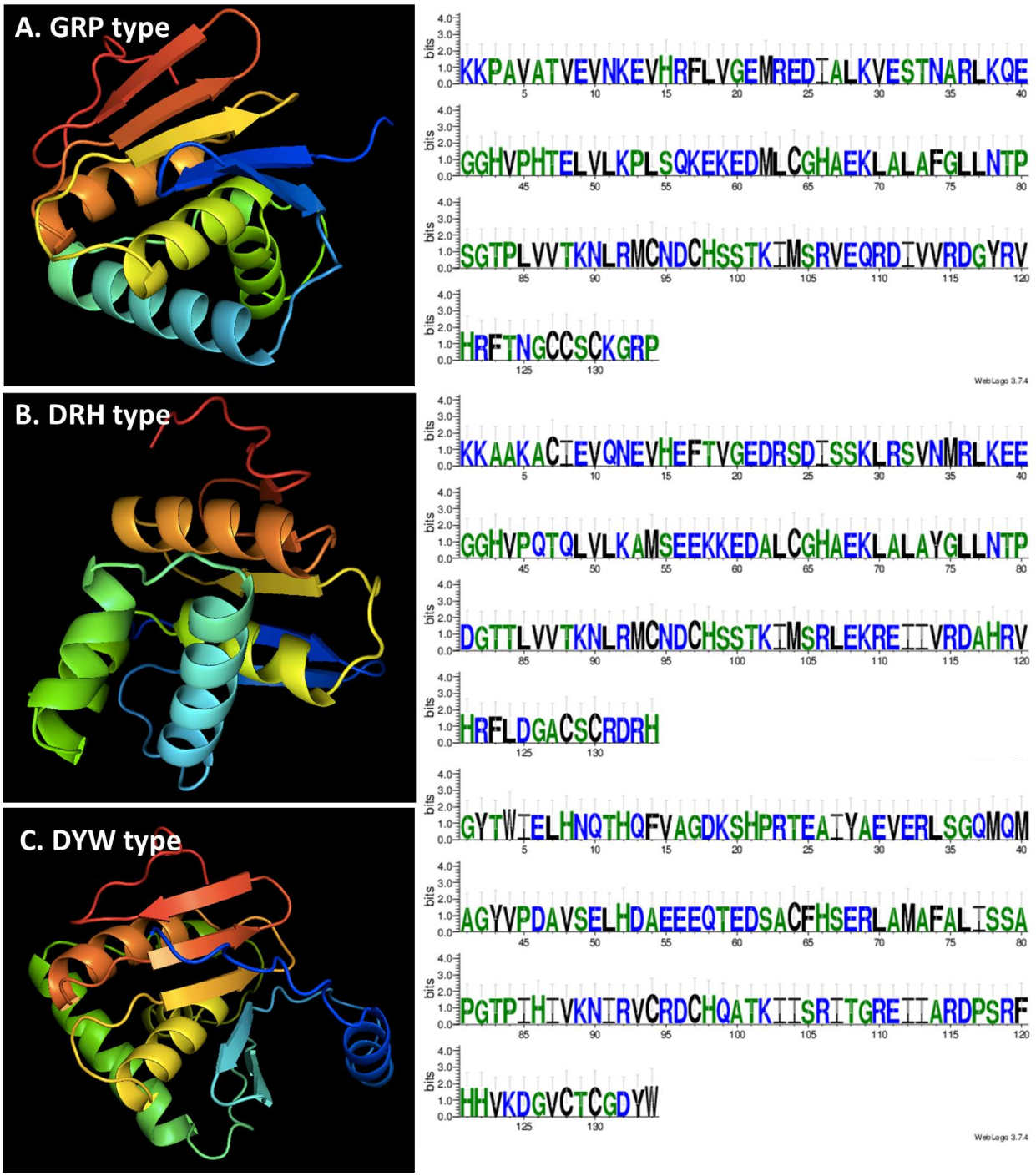


Figure 4.3: The three-dimensional structural prediction for three variants, A. GRP-type, B, DRH-type, C. DYW-type. Sequencing Weblog generator showed the three different variants of DYW type PPR proteins.

4.3.2. DYW domain from hornwort PPR protein has cytidine deaminase activity in E.coli

Bacterial expression system is used for studying the C-to-U and U-to-C editing catalysis of DYW-like domains in hornwort by fusion cloning with the PPRE1E2 domain of the *P. patens* pentatricopeptide repeat (PPR) protein PPR56. PPRE1E2 domain of the PPR56 is fused with the hornworts PPR proteins C-terminal variants GRP, DRH or DYW domain, followed by the target sequence site, nad4 with original cytidine target (nad4-UCA) or nad4 with U at the target site (nad4-UUA) was cloned into pETG41K and expressed in E. coli (Figure 4).

We have observed C-to-U RNA editing for the two fusion PPR proteins with DYW domain, PPR56-DYW-g16688 and PPR56-DYW-g10204 (Figure 4.5A). While PPR56-DYW-g16688 showed 50% of C-to-U RNA editing efficiency, the editing efficiency for PPR56-DYW-g10204 is 100% (Figure 4.5B). We also co-transformed repeated nad4 targets with one nucleotide substitution at -1 or +1 position. PPR56-DYW-g10204 showed edited 100% all target sites except for U<u>C</u>G, which show 80%. Editing efficiency with PPR56-DYW-g16688 are only 20-50% in all repeated nad4 sites. Sanger sequencing analysis in *E. coli*. for C-to-U RNA editing by Hornworts pentatricopeptide repeat protein PPR with DYW type domains, Pp_PPR56_DYW-g10204 for its nad4_9 repeat (Figure 4.6A). While GRP and DRH were observed for neither C to U nor U to C RNA editing. The sequencing data for all the clones from GRP, DRH and DYW type domains analyzed for U-to-C RNA editing are shown in Figure 4.6 B

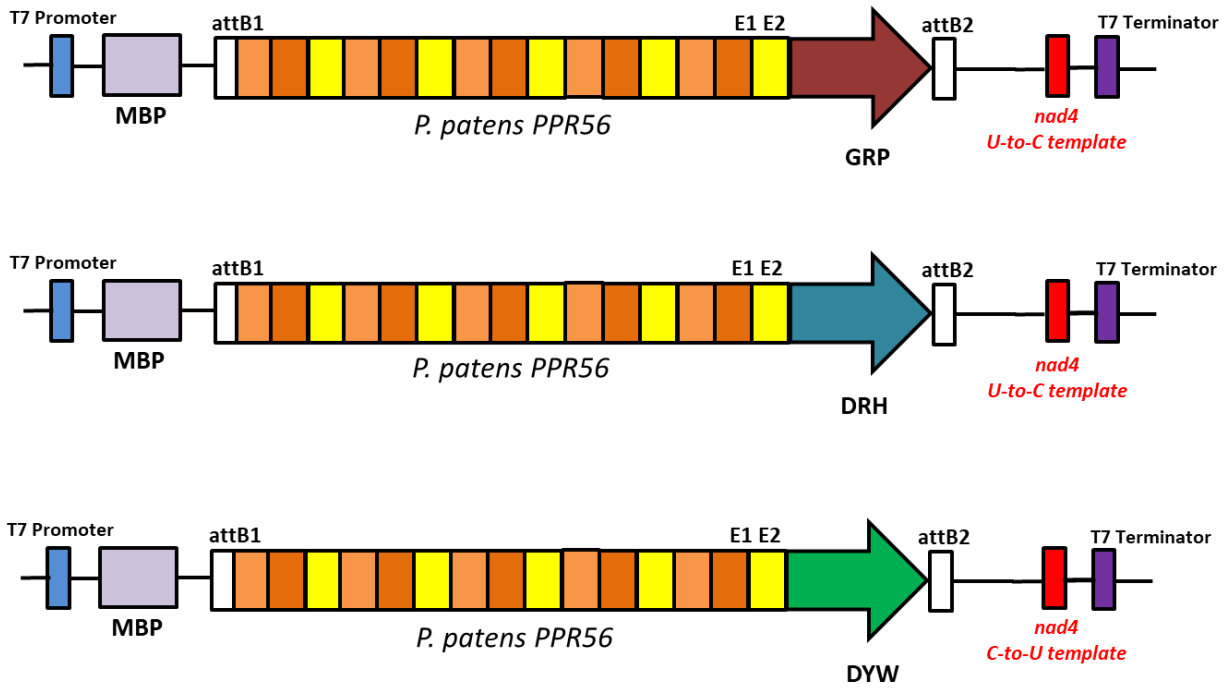


Figure 4.4: *Escherichia coli* expression system for RNA editing. *P. patens* pentatricopeptide repeat (PPR) protein PPR56 with modified C-terminals and nad4 editing template. The PPR56 coding sequence is inserted into the pETG_41K vector system resulting in the fusion to a His6-tagged maltose binding protein. The hornworts PPR proteins C- terminal variants GRP, DRH and DYW domains is cloned downstream, followed by the target sequence site. Expression is driven by a T7 promoter inducible by isopropyl β-D-1-thiogalactopyranoside (IPTG).

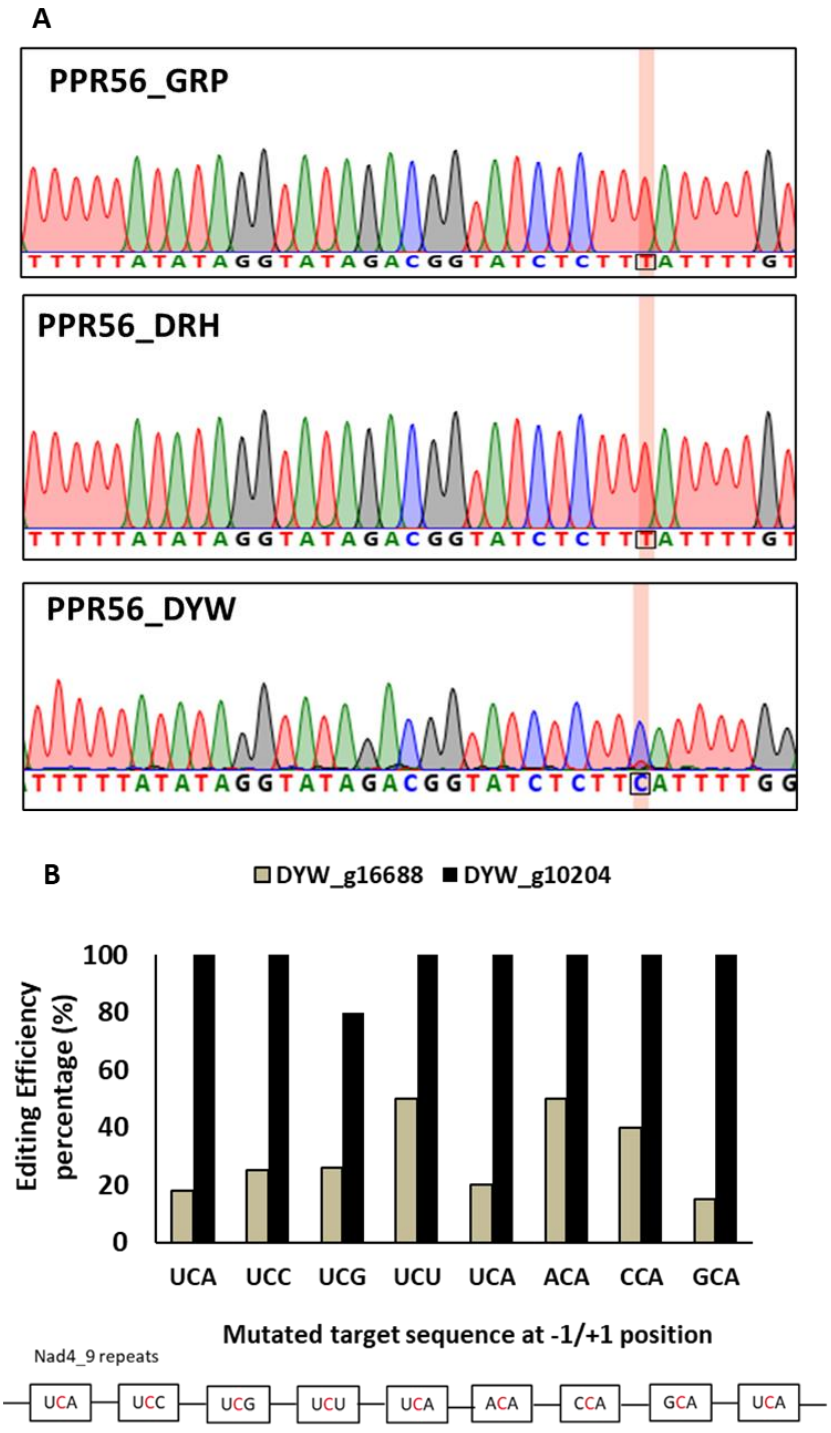
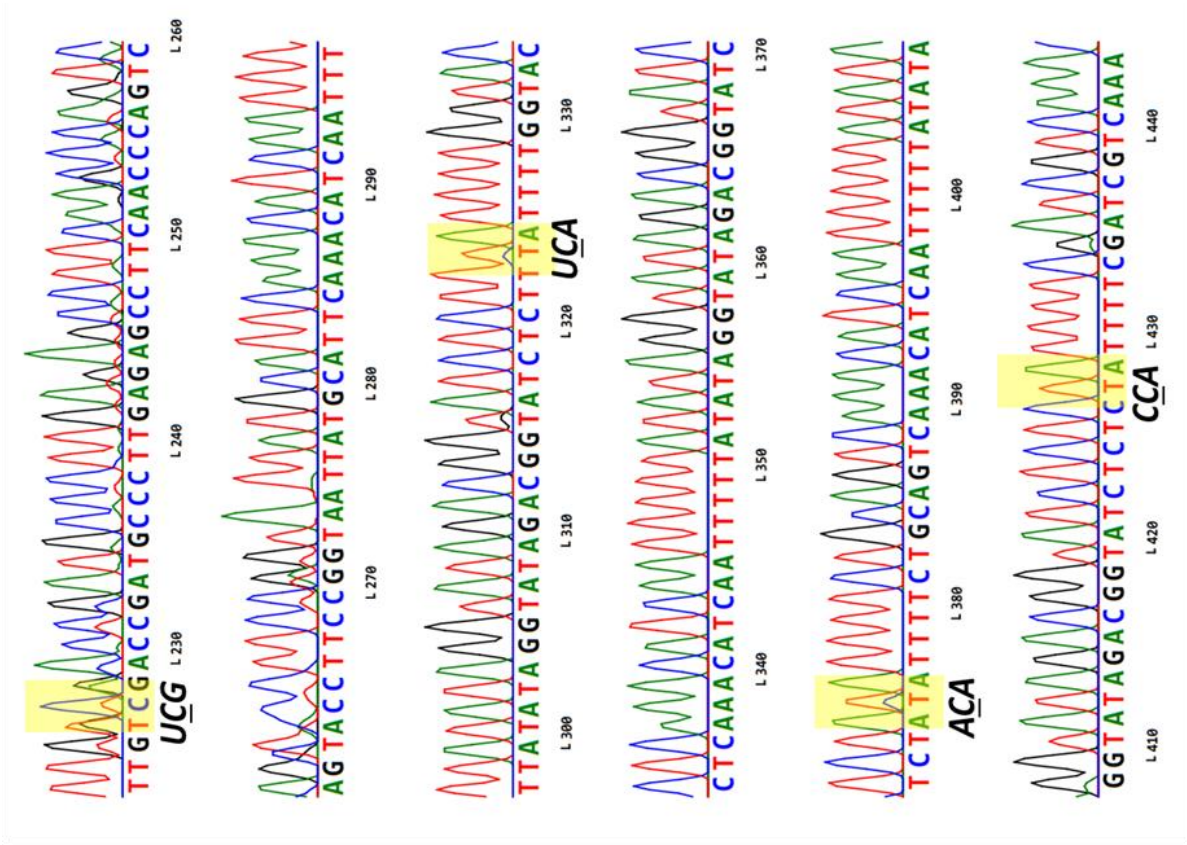
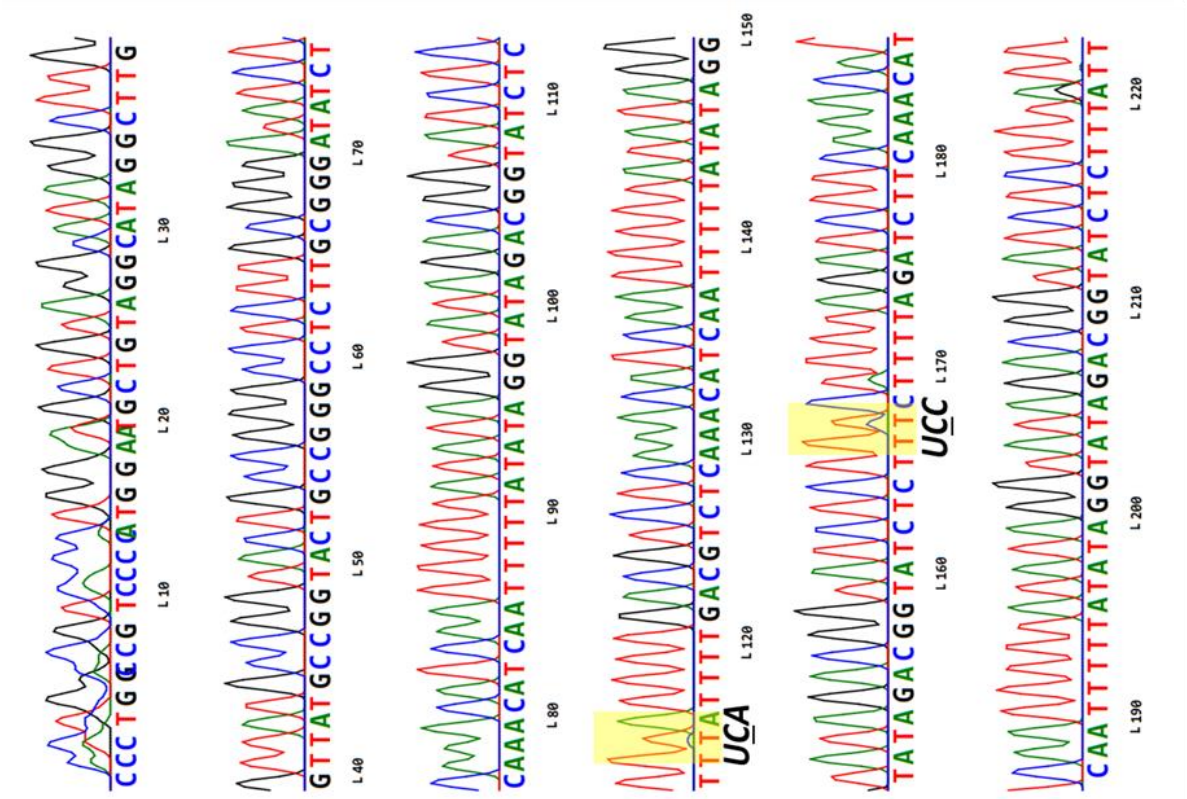


Figure 4.5 A. Sanger sequencing analysis for RNA editing in *E. coli*. **B.** Percentage for C-to-U RNA editing by Hornworts pentatricopeptide repeat protein Pp_PPR56_DYW-g16688 and Pp_PPR56_DYW-g10204 for its nad4_9 repeat. (n:3**)**


UCG
UCA
ACA
CCA


UCA
UCC

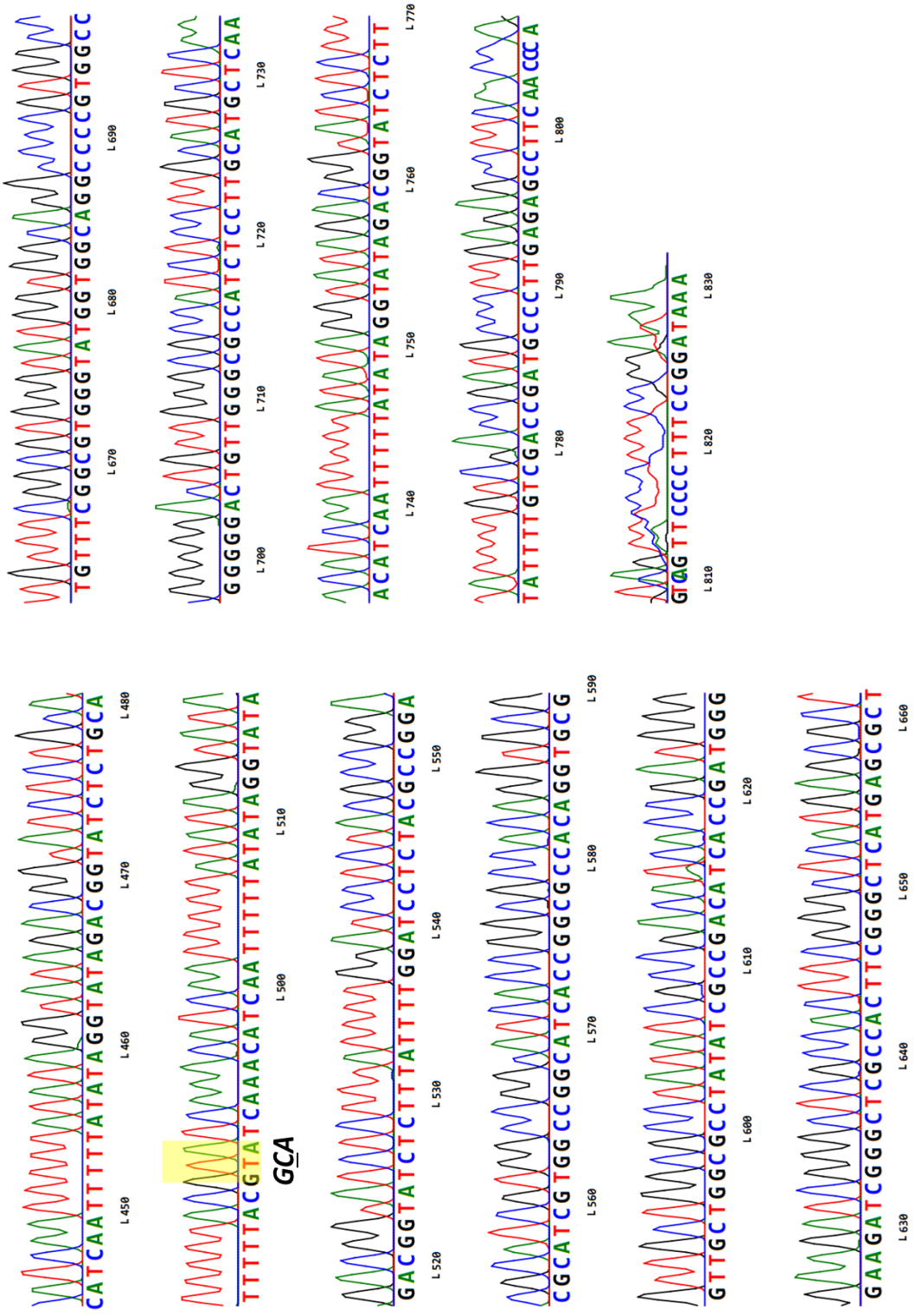


Figure 4.6:A. Sanger sequencing analysis for C-to-U RNA editing in E.coli .

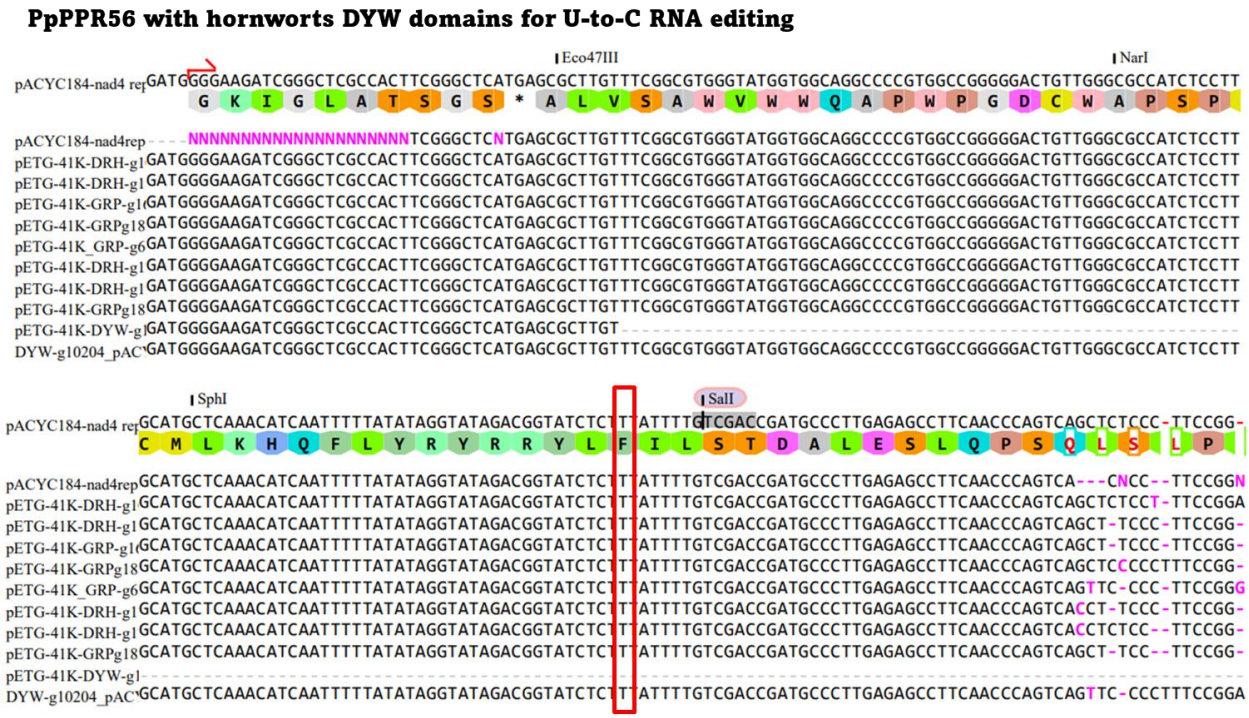


Figure 4.6: B. Sanger sequencing analysis PPR56 with hornworts DYW domains for U-to-C RNA editing. (n:3**)**

4.3.3. Gene Expression of Hornworts DYW-type domains into Bacterial cells

Failed RNA editing with DRH and GRP type DYW domain can be very low expression in E.coli and HEK293 cells. Thus, transcript abundance of each fusion genes was calculated by quantitative-PCR. Although, DYW type fusion genes were expressed the best, DRH and GRP types also showed 40-50 % expression of the DYW type, suggesting that the GRP and DRH variants are unable to induce RNA editing because of the lower expression in E coli. The Ct is the value where the PCR curve crosses the threshold, in the linear part of the curve. The higher the Ct (30-35), the less the mRNA detected is present, because it need more cycles of amplification to detect the fluorescence. If the Ct has a small value (10-15), the gene is highly expressed. From the amplification plots. It is clear that DYW domains required least number of cycles to get fluorescence. Therefore, it is highly expressed into the cells as compared to the other domains

(Figure 4.7 A-B). Average Transcript abundance was calculated by delta Ct, difference between the test gene and control gene. Delta Ct value was found highest for DYW domains, indicating the most mRNA is present. GAPDH was used as a housekeeping gene. Expression of Hornworts specific DYW domain variants were also quantified into animal, HEK cells. Gene expression level was measured by relative abundance of transcript present. Amplification plot showing the maximum expression of DYW domains to get fluorescence within minimum number of cycles. Gel electrophoresis bands after RT-PCR showed the expression of DYW domain only Figure 7 F-I.

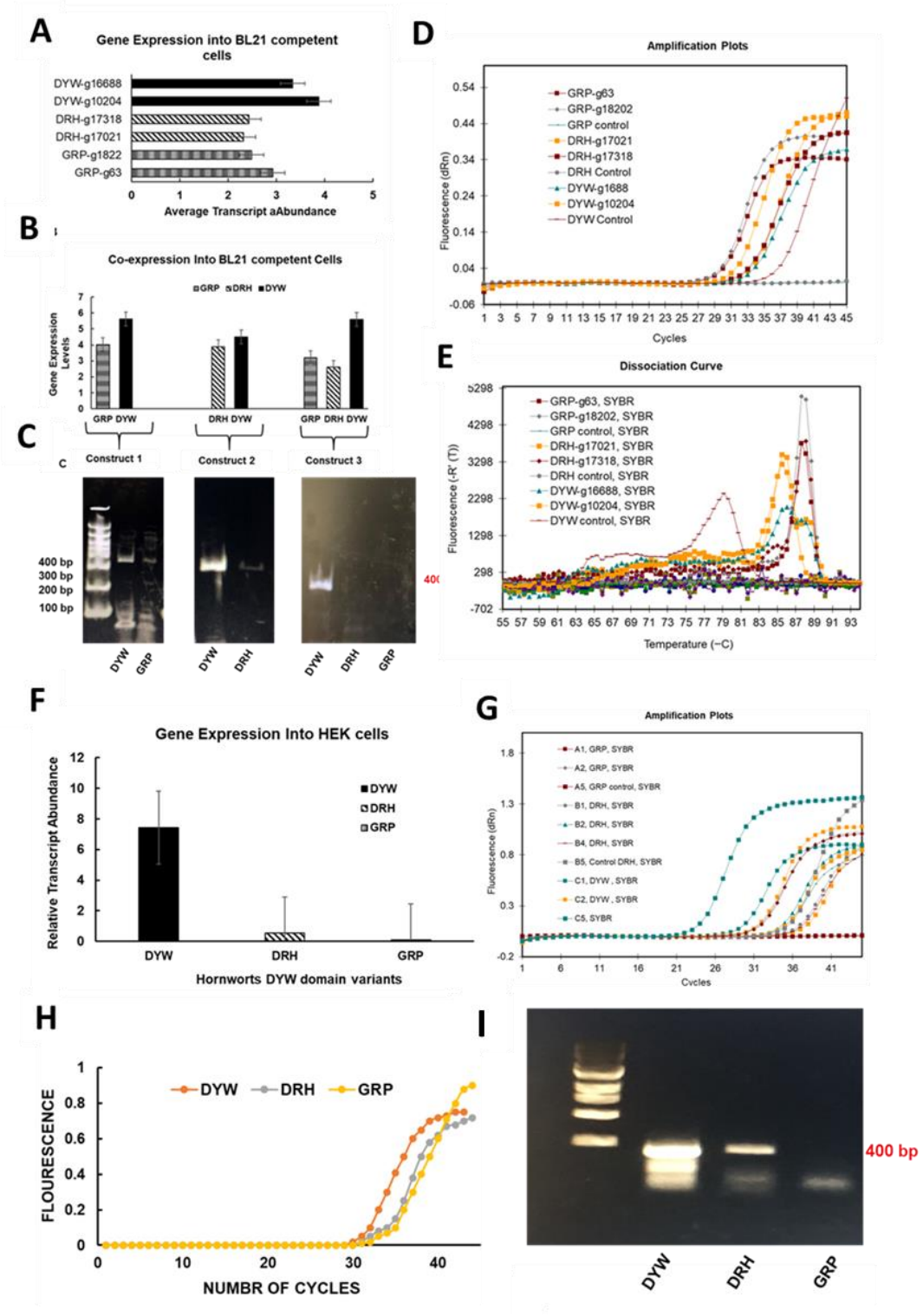


Figure 4.7. Expression of Hornworts specific DYW domain variants into the E. coli, BL21 competent cells. A. Comparison of Average transcript abundance of DYW-type PPR proteins. **B**. Comparison of co-expression levels of DYW-types into Bacterial cells. **C.** Electrophoresis gel

showing the band patterns observed for RT-PCR. **D.** Amplification plots. **E.** Dissociation curve. **Expression of Hornworts specific DYW domain variants into animal, HEK cells. F.** Gene expression level was measured by relative abundance of transcript present. **G.** Amplification plot showing the maximum expression of DYW domains to get fluorescence within minimum number of cycles (**G** and **H**). Gel electrophoresis bands after RT-PCR (**I**).

4.3.4. Expression of Hornworts specific DYW-type domains into HEK293 cells

In case that the fusion proteins are not stable in E.coli, we tried human cell expression system. The *P. patens* PPR56 PPR domain with modified C-terminal DYW variants from hornworts was cloned into the pCS2+ vector system and their C-to-U and U-to-C editing activity was analyzed. PPR56 fused with DYW successfully edited co-expressed tareget cytidine into uridine, though the efficiency was much lower (about 51 %) than in E.coli system. DRP and GRH type DYW variants did not show C-to-U as well as U-to-C conversion like in E.coli (figure 4.8).

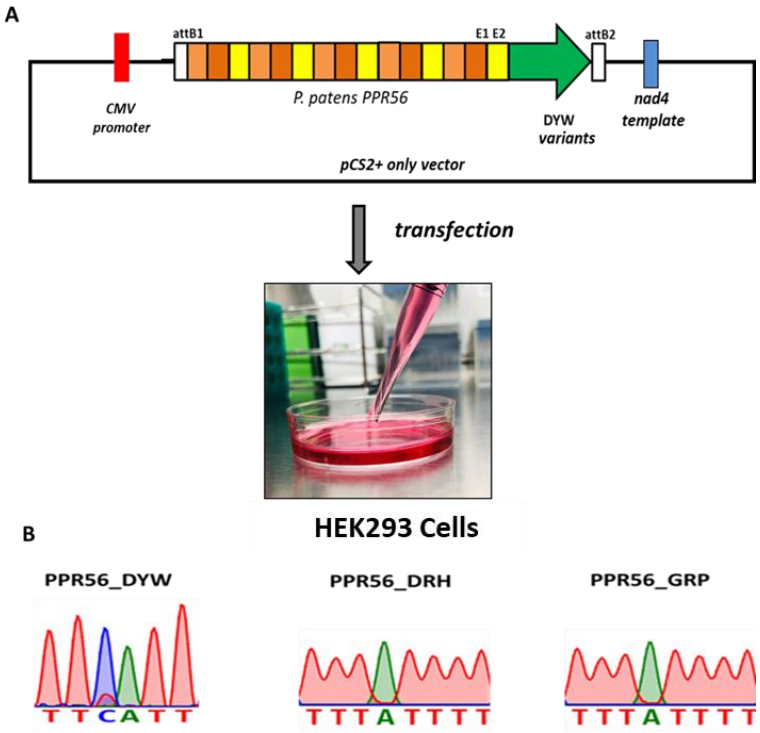


Figure 4.8: Expression of Hornworts specific DYW-type domains into HEK293 cells

Since, Members of the apolipoprotein B mRNA editing enzyme, catalytic polypeptide (APOBEC), a member of the APOBEC family can also perform editing on its own in vitro as well as in vivo in the absence of cofactors for catalyzing the deamination of C-to U RNA editing in mammalian cells and tissues [36–38]. After confirming the DYW domains for the C to U RNA editing, further, analyzed the RNA editing patterns by substituting the DYW domains with APOBEC1 deaminase enzymes from mammalian system. We cloned the APOBEC1 enzymes downstream of PPR56, targeted the nad4 template for C to U RNA editing, and expressed into BL21 competent cells. After RT-PCR analysis of nad4 genes, C to U RNA editing was identified up to 28%. The schematic for all the three successfully developed artificial system for C to U RNA editing in both E coli and HEK293 cells are shown in (Figure 4.9 A-D).

Considering the hypothesis that the C-to-U and the U-to-C editing sites come together, indicating that an amino-group, dissociated from C which further converts to U, could be integrated with the neighboring U that subsequently converts to C [39] co-expression of DYW and GRP domains were also investigated in the Bacterial system (Figure 4.9 E and F). However, no editing was reported for either C-to-U or U-to-C.

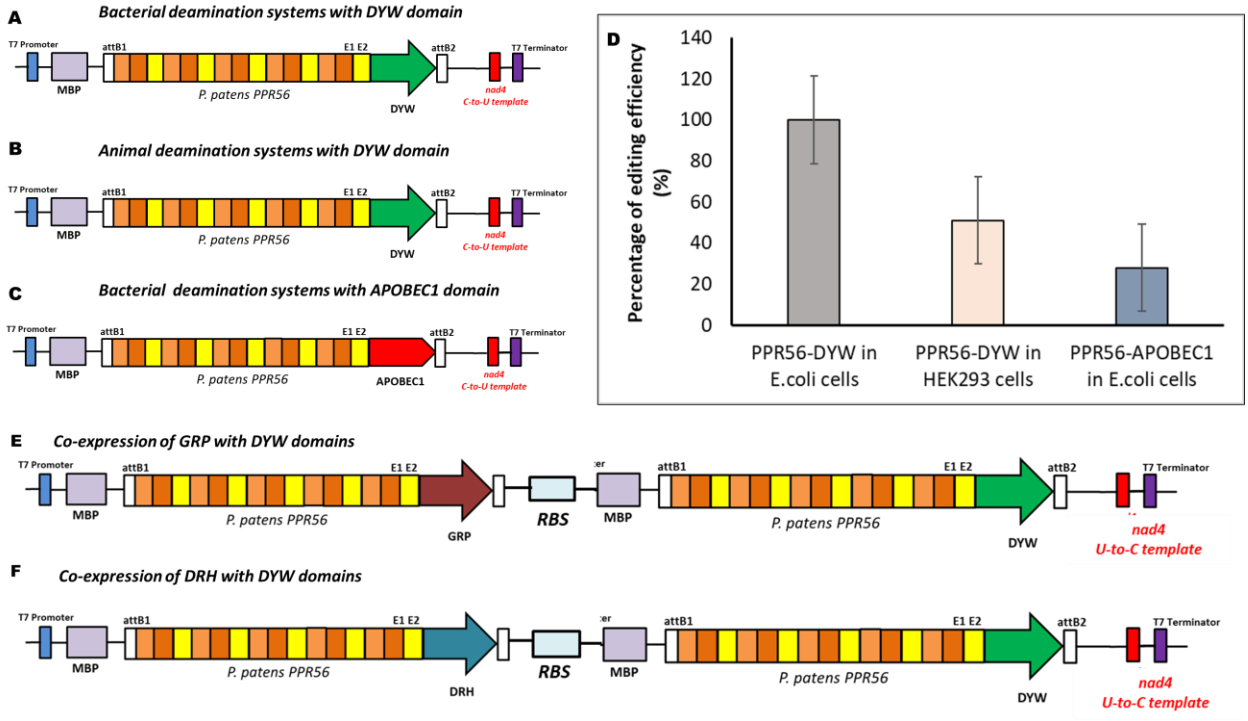
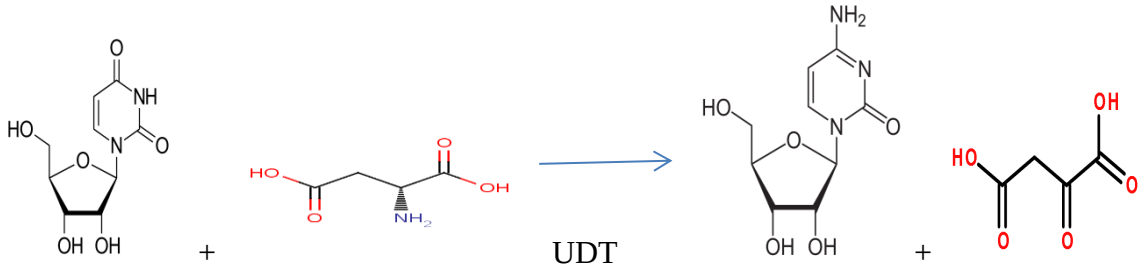


Figure 4.9: schematic representation for the artificial system developed for C to U RNA editing. **A**. Bacterial deamination system developed with PPR56-DYW. **B.** Artificial deamination system developed with PPR56-DYW in mammalian HEK293 cells. **C.** Bacterial deamination system developed with PPR56-APOBEC1. **D.** The average percentage of editing efficiencies observed in the three systems (n:3). The schematic for the co-expression of GRP and DRH with DYW domain (**E** and **F**).

4.3.5. In vitro Uridine Aminotransferase Assay

Transaminases or aminotransferases are enzymes that catalyze a transamination reaction between an amino acid and an α-keto acid. In this method, Uridine aminotransferase catalyzes the transamination of Uridine and aspartate, forming Cytidine and oxaloacetate. The oxaloacetate is then reduced to malate by the Malate dehydrogenase, while NADH is simultaneously converted to NAD. The change in absorbance due to the consumption of NADH is measured at 340 nm and is proportional to the Uridine aminotransferase activity in the sample (Fgure 4.10). Samples were

also observed for the aminase activity using 1H-NMR. Appearance of NH_2 peak after 20 hours of GRP treatment indicates the possibitlity of aminase activity in this domain (figure 4.11).



Oxalacetate + NADH + H+ MDH L-Malate + NAD+

Methodology:

Uridine aminotransferase may be assayed spectrophotometrically in a coupled reaction with Malate dehydrogenase in the presence of NADH. One unit oxidizes one micromole of NADH per minute at 25°C under the specified conditions.

Reagents:

Prepare reagent mixture containing:

For 1 mL,

Uridine (MM 244.0 g/mol); 0.24 g

Aspartate; 0.1g

Malate dehydrogenase; 10 uL

Enzyme:

GRP domains of PPR proteins dissolved in Tris buffer; 10 uL

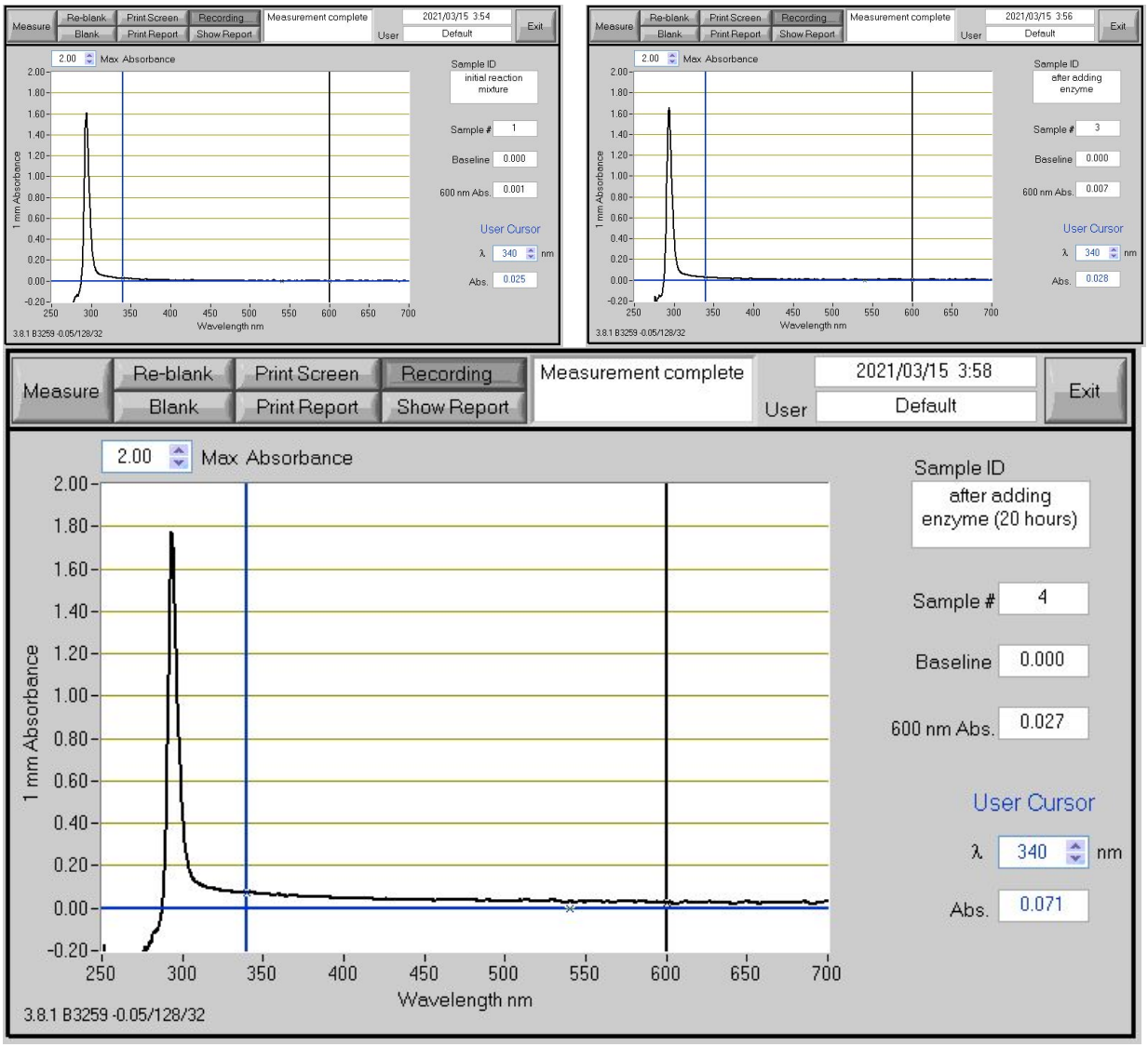


Figure 4.10. Absorbance for GRP proein

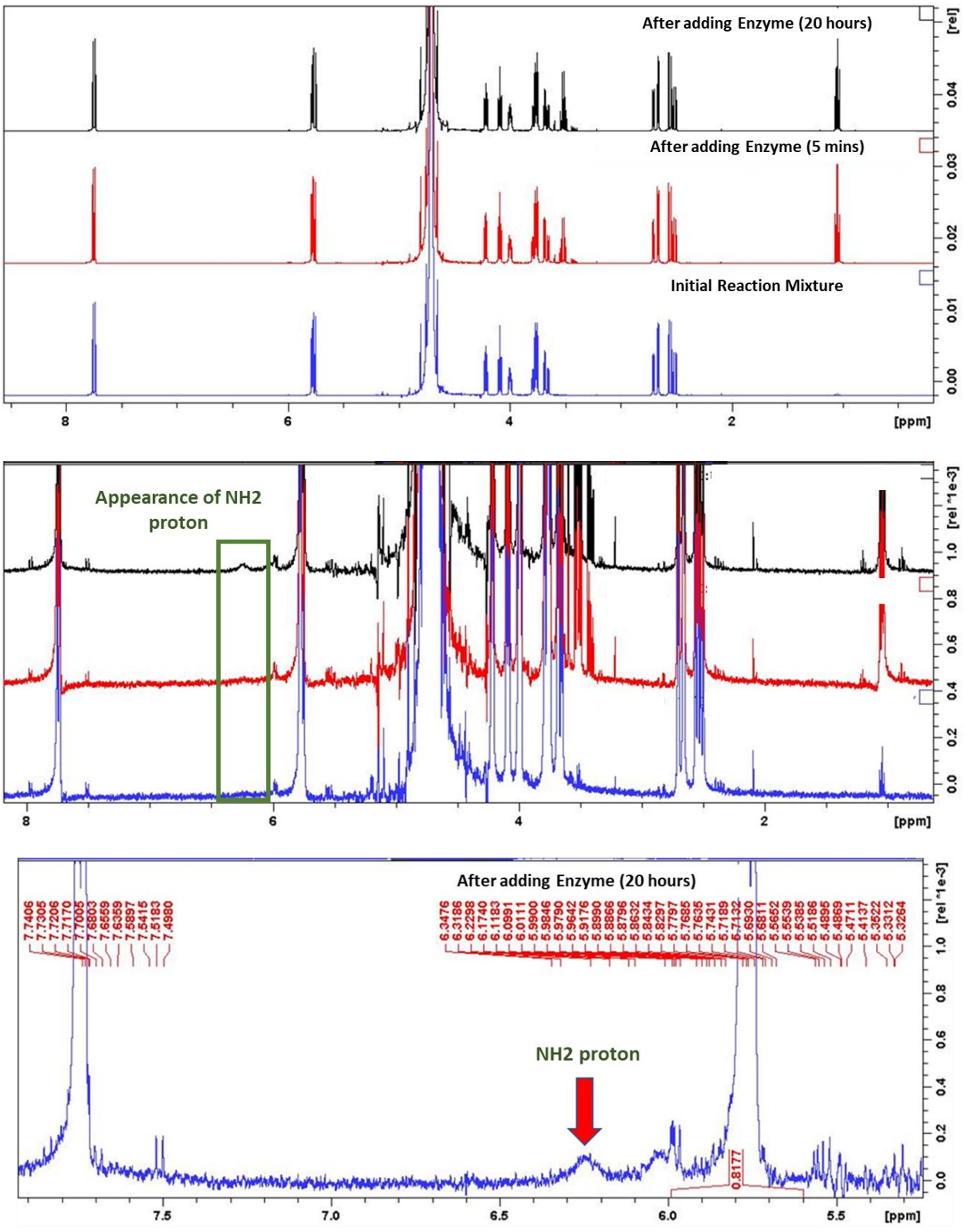


Figure 4.11. ^{1}H-NMR analysis for GRP domain

4.3.6. Protein expression (Attempt 1)

The modified C-terminal DYW variants from hornworts PPR genes are cloned into the pColdGSTDNA vector system resulting in the fusion to a GST-tagged protein. Expression is driven by *cspA* promoter. The domain was amplified by PCR from complementary DNA of hornworts using forward and reverse primers harboring the appropriate restriction sites. Expression was inducible by isopropyl β-D-1-thiogalactopyranoside (IPTG).

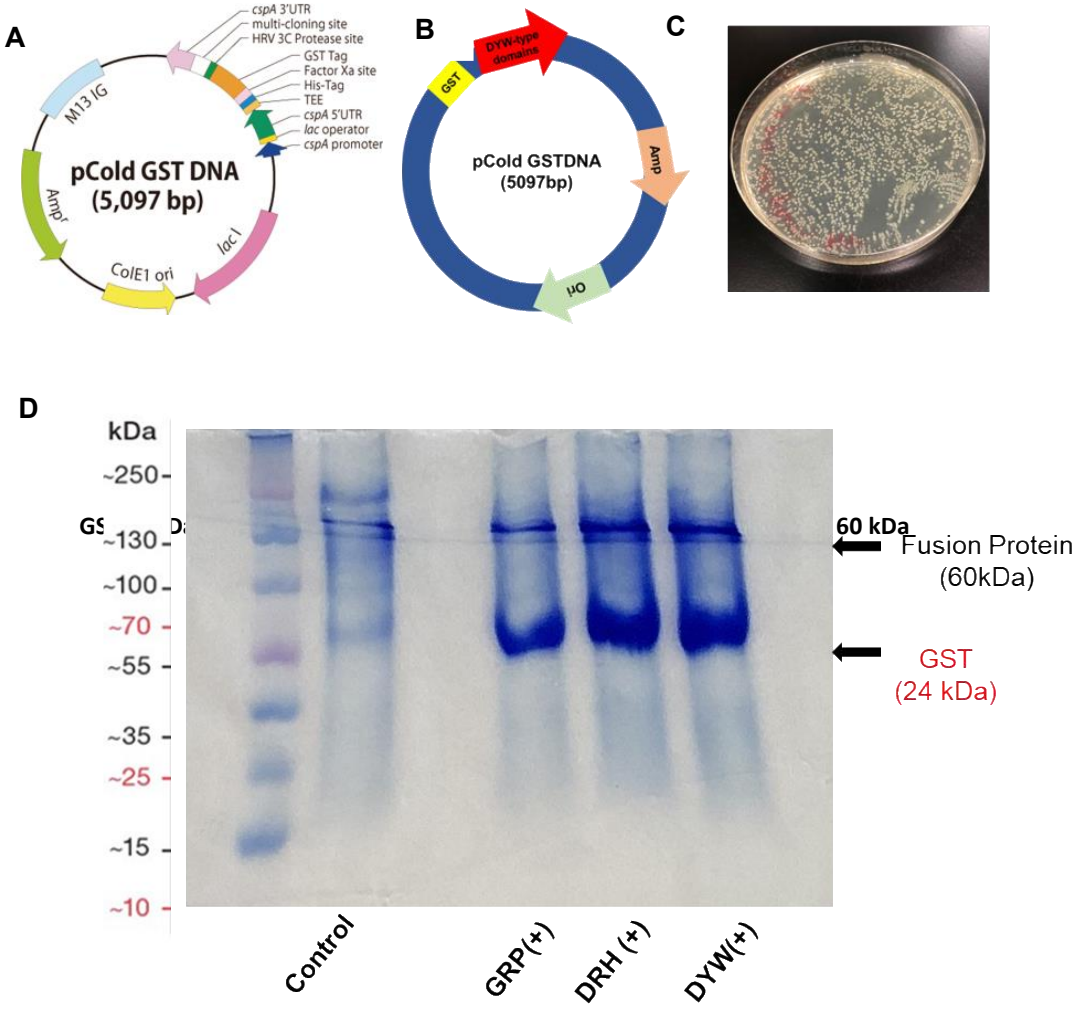
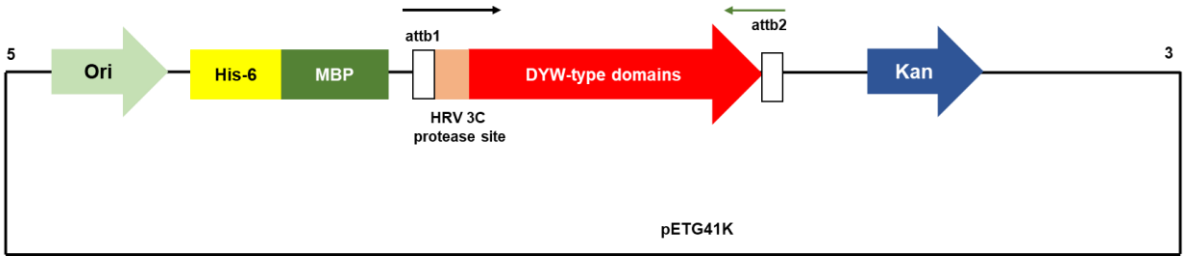


Figure 4.12. (A) pColdGST DNA vector, **(B)** DYW-type domains were cloned into pColdGST plasmid. **(C)** colony formation. **(D)** SDS-PAGE: Comparison of expression of proteins before and after IPTG induction. **GST:24kDa, HRV Protease site: 22kDa, Target protein: 12-14kDa, total: 60 kDa**

(Attempt 2)

In addition, the hornworts PPR proteins C- terminal variants GRP, DRH and DYW domains is cloned into the pETG_41K vector system resulting in the fusion to a His6-tagged maltose binding protein. The HRV 3C cleavage protease site, *Leu-Phe-Gln-Pro*, was inserted upstream of DYW domain as shown below. Expression is driven by a T7 promoter inducible by isopropyl β-D-1-thiogalactopyranoside (IPTG) for 20h. Expressed proteins were extracted by amylose resin and the bacterial lysates were examined on a denaturing sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) before and after IPTG induction (Figure 4.13A). followed by western blotting with respective antibodies (Figure 4.13 B). lastly, the fusion protein was purified and digested with HRV 3C protease enzyme (Figure 4.13 C).



TGACAAGTTTGTACAAA Forward Adaptor (attb1)

CTGGAAGTTCTGTTCCAGGGCCCG HRV 3C protease site
Leu-Phe-Gln-Gly-Pro

BsrGI

TGACAAGTTTGTACAAACTGGAAGTTCTGTTCCAGGGCCCGAAGAAACCTGCAGTGGCG

Forward Adaptor | Protease site | Gene seq

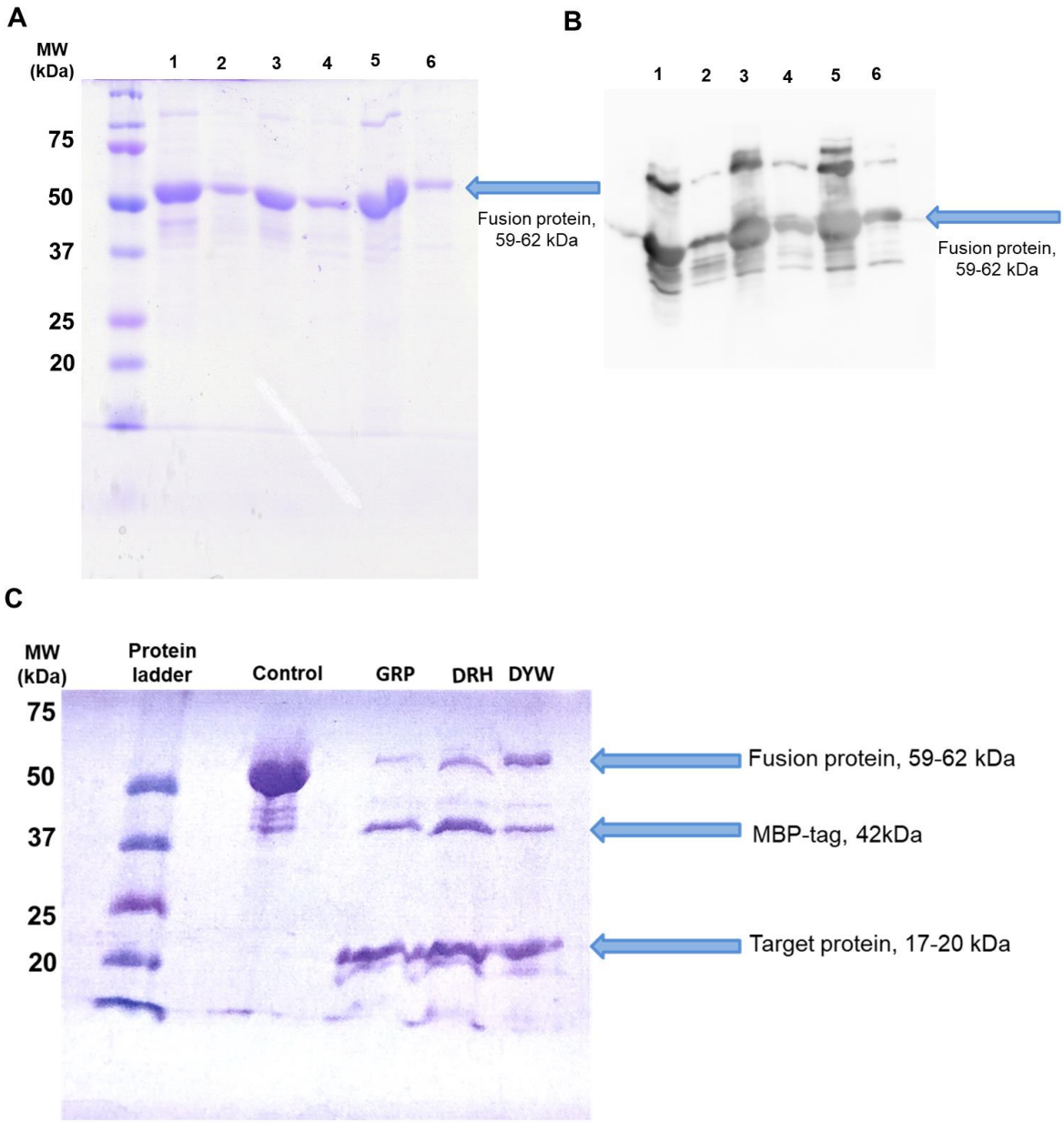


Figure 4.13: Expression of hornworts specific PPR proteins by pETG41k vector. (A) SDS-PAGE, **(B)** western-blotting, **(C)** HRV 3C protease digestion. **Lane 1: GRP, Lane2: GRP (non-induced), Lane 3: DRH, Lane4: DRH (non- induced), Lane 5: DYW, Lane6: DYW (non-induced).**

4.3.7. Protein sequence alignment

```
DYW      MSRIGALESHMKVKVRILMKSNLSPYIIMEEWFRRLS----------------------      37
GRP      ----------MA-----TAKAEASFLLFAVHWLSG-GTQQIDVCKGILRGAPANLISGSR      44
DRH      ----------MQVELSLVAQDKQSTIHPMRNLLAAKSSSLSMKCGCVVRNGPSGKIAGVA      50
                   *       : : *      . :   .

DYW      ----------------------TLRSNTKGQEDELFRQEALLGYPRLR--KGDKNYTGSV      73
GRP      LLASKASRCSSVRSPPLPLLRDGFSSNGRCQEQEHLWQQFFLDPTMWWDKRRDKKSSRSP      104
DRH      Q-------------DSSP----AVKSMTHKERQELLWQQFFLDPLDWWDNRADKRSPRHP      93
                            . *  : :.:* : *: :*.       : **.

DYW      E---MSRSDGLWI---------DWEGVSSLHSVHYGGELESSGSVKATKEGKHLQ----V      117
GRP      DFRRKGSGKVLWLNHMFKPAWVDEQLVSLDMQMGHGASPSKSSSPDIGEFGRRVGSEMGS      164
DRH      DFRHKTSEEPLWLNSPRKPAWVDSQLARWDVEMRGFPAL--------G-----------A      134
         :       . **:         * : .    .:

DYW      S--TFE----TESEASVCVESSLMGISSKRVGRPELW----MTPADVSSLCEEGRLRESV      167
GRP      GCSSVRSPALPHPMGHGASPSQSISQDTPEPGRRAGWEIGVVTAVTVSTLCQQGKISEAV      224
DRH      G----------------NVED------------------ENSVMKSLTKRCEEAELGEAV      160
         .                  .                     .   ::. *::..: *:*

DYW      QALNVLEHQGMQAHPDTYYHLLQQCIKRKAWQDGLQVHAHMKRSGLEVNSLLAGSLVRMY      227
GRP      GALESLAETDSGVSQSLFYTVLKLCTDKKDLVVGRRVHDLSVKLGYESNAYLGNHILRMY      284
DRH      EALNVLVQRGYRISHSLFYQILKRCMAERDLGCGRRIHALVQTGGYGSNAFVANHIIGMY      220
          **: * . .     . :* :*: *  .:    * ::*      *   *: :.. :: **

DYW      ASSGKILTARQIFDQMSRHDVFTFTALMKGYLSCGQADKVLDLYKNMKDEGVKPDKFVFT      287
GRP      ACGGRLEEAMDVLTRVGKPDAFMWFSIISAYAKYGKPGEAIQLYHEMRRSGVKANAHVFV      344
DRH      ASHGELEDAVQVFRNVTAPSVHMWSSIILAHVRHGQPAQAIQLYQQMREASVKPDNRIFV      280
         *. *.:  * ::: .:   ... : ::: .:   *:  :.::**::*:  .** :  :*.

DYW      VVLNACTSLGNIQEGRQTHAEIIKVWSETDVIVDNCLLDLYAKGGSMEEASSVFDRMHER      347
GRP      AALQACAFAADLESGRKVHRDVEQSGVPSNLFIGSCLVDMYAKCGSMKDARQVFDSLPSK      404
DRH      VALKACASAGDLVSGKQVHADIPQGAVQEDGFVATSLVNMYAKCGSLVEARKVFDSLRKK      340
         ..*:**:  .:: .*::.* :: :     : :: ..*:::*** **: :* .*** : .:

DYW      DVVSWNTMIAGYTRSGEGSKALRLYRQMQEAKV-NADGSTFVAVLNACASMADGGFGKQV      406
GRP      DVVTWNTLIAGYAQQGLGQEALVVYGQMEEAGLILADHVTFSCVLKACGSTGALQKGRQL      464
DRH      NVVTWNAMIAGYSQHGLGREALALYESMQHEGITPADHITFVCLLQACASAGALQLGRQL      400
         :**:**::****:: * * :** :* .*:.  :  **  ** .:*:**.* .    *:*:

DYW      HAQIARSRWESDMYVCNALVDMYSKCGNLQAARTVFDNLQTRDVVSWNAMISGYSRHGYN      466
GRP      HSQIVERGLLHDVVVGNCVVDMYAKCGKLEDARQVFDSMATKNVVTWTALIAGYSQQGLG      524
DRH      HLQIRARGLEGD-------------------------------------------------      412
         * **       *

DYW      EEVLHVFRQMLHKGVK-PTSITFVGILNACASLGALEDGEQLHAHVRNCGLESQIFVGSA      525
GRP      QEALVLYGQMEEAGLILANHVTFACLVKACGSTGALQKGRQLHSQIVKRGLLDNVVVGNC      584
DRH      ------------------------------------------------------VRMGNC      418
                                                                : :*..

DYW      LVDMYCKCGFIAGARAVFDRMPKRDVVSWTSMFSGYADHGLAKEALCLFEEMQQEDVKPN      585
GRP      LVDMYAKCGELEDARQVFDSMTTKDVVTWTTLIAGYAQQGLGQEALVVYGQMEEAGLN--      642
DRH      LVDMYAKCGSLDDARTVFDSLPTKDVVSWNAMIAGYVQQGLGHHALTLYASMRKEGIT--      476
         *****.*** : .** *** : .:***:*.::::**.::**.:.** :: .*.: .:.

DYW      SGTIVSVLNACGNAAALAEGKEMHTHASRCGFASDLCVCTALISMYGKCESMEDALEVFN      645
GRP      ------------------------------------------------------------      642
DRH      ------------------------------------------------------------      476
```

```
DYW      GISDKDAVAWTAMLTAYIHNRRERDALQVFKEMLQKNVQPTDVTFVCVLNACANLAVLGE       705
GRP      ------------------------------------LADHVTFACLVKACGSTGALQK       664
DRH      ------------------------------------VANLVTFSCLLQACASVGALQQ       498
                                                *** *:::**.. ..* :

DYW      GKRIHAMIDRSGHRSRIFVNNALVDMYAKCGRLDIARMVFDKMPRRDIISWNALMAAYTQ       765
GRP      GRQLHSQIVKRGLLDNVVVGTCLVDMYSKCGELEDARQVFDSMATKDVVTWTALIAGYGA       724
DRH      GKQLHSEIRERGLEADVFISSCLVDLYSKCGTLEDARKVFDSFPTRDVVTWTALLNGYAE       558
         *:::*: * . *    :.:...***:*:*** *: ** ***.:  :*:::*.**: .*

DYW      HGRSKEALIIFEHLLQMEVELDYITFVTVLAACSHAGLVKEGCSYYRSMILDHGISPTEG       825
GRP      CGQGHKALACFEEMLVAGVHPNDVTFTCLLVACSHEGLVHEGRKLLDSMAEQHGIKPTIE       784
DRH      HSDGHKAIQCFEEMLKQGIKPNDTTFLCLLVACSHAGLVQEGQKYFNQMVEDHGIAPTDY       618
          . .::*:  **.:*   :. :  **  :*.**** ***:** .   .*  :*** **

DYW      HHVCMVDLLGRAGRLDEAEEFITNLSAQPGVAVWMALLGACRLHGNVGIAEHAAERVLQL       885
GRP      HYTCMIDLLGRAGQLDEAEKMLLSCEGQINIVGLTSLLNACKSHGDIEKAAWCFESIVRL       844
DRH      HYSCMVDLLARSGQLDEAEHMLRTSAFVNDVVGWKALLSACKNHGDAERGKRCFDHLVQL       678
         *: **:***.*:*:*****.:: .     .:.   :**.**: **:   .  . : :::*

DYW      DPSHDAAHVLLANTYAAAGMWREKLAVRRLLKDKGLKKNPGCSWTEIKNEIHLFFAEDKR       945
GRP      DPQMASAYVLMANAYADAGRWTDVDRIETRRRAAGANKKPAVASIEVNTEVHRFLVGERR       904
DRH      DPEDASAYVLLGNVYANAGRWDDVTRIESLRKSVGAWKKAAKACIEVRNEVHEFTVGEDR       738
         **.  :*:**:.*.** ** * :   :.   :  *  *: . :  *:..*:* * . : *

DYW      HPQTDIIYDTLDKLIGTIKEAGYVPNTSFVLHDVDEKEKERCLRYHSEKLTIAFALINTA       1005
GRP      KDVALKV----STNTRVKQEGGHVPHTKLVLKPLCEKEKEDELCGHAEKLALAFGLLNTP       961
DRH      SDISSKL----SVNTRLKEEGGHVPRTELVLKTVSEQEKEDALCGHAEKLALAYGLLNTP       795
            :  :    .      :*.*:**.*.:**: : *:***  *  *:***::*:.*:**

DYW      PGTPLLILKNLRMCGDCHNAAKYISKVVGREIVTRDVSRFHHFKDGECSCRDYW     1059
GRP      SGTPLVVTKNLRMCSDCHSSTEIMSRLEQRDIVVRDGYRVHRFANGSCSCKGRP     1015
DRH      DGETLLVTKNLRMCNDCHSSTKIMSRLESREIIVRDAHRVHRFLDGSCSCGDRH     849
          *  *:: ******.***.::: :*::  *:*:.**  *.*:* :*.*** .
```

4.3.8. Amino acid sequence of the studied Hornworts PPR proteins

GRP type

1. g63.t1 gene=g63

MGTIWPPDPLWTQQVGIGVCTRILPGAPAQLMLGSAPLANKASRCSSVRSPALARPVGHGASPFESSSDKPEPGRRAGSE
MGVVTAVTVSTLQQGKLSEAVGALESLAQSDSGVSQCLFYTVLKLCTAQKDLVVGRRVHDLTVKLGYESDAYLGNHI
LRMYSCGGRLKEAMDVFTRVGKPDVFMWVSIISAHAKHGKPKEAIQLYHMRRSGVKANGYIFVAALQACAFAADLES
GRTVHRDVLHSGIRTDLFVNNCLVDMYAKCGKMKDARQVFMDTKDVVTWNALIAGYAQQGLGQEALVVYGQMEEA
GLYLANDVTFACLLKACAITGALQKGRQLHSQIVRGPLDNVVVGNCLVDMYAKCGQLKDARQVFDSMASKDVVTWN
ALIAGYAQQGLGQEALVVYGQMEEAGLNANDVTFACLLKACASTGALQKGRQLHAQIVKRGLLDDVVVGNCLVDM
YAKCGQLKDARQVFDSMASKDVVTWTALISGHARYGELEDARQIFDSMASKDVVTWNALIAGYAQQGLGQEALVVY
GQMEEAGLIFANDVTFACLLKACASSGALQKGRQLHSQILKRGLLDDVMVGTCLVDMYAKCGELEGARQVFDNMAS
KNVVTWNALIAGYGACDEGRKALACFEEMLGAGVHPDDVTFTCLLVACSHEGLAQEGRKHFDSMAEQHGIKPSIEHY
NCMIDLLGRAGQLDEAEKMLLSCEGQSNVVGLTSLLNACKNHGDIERAVWCFENIVRLDPQMASAYVLMANAYADA
GRWTDVDRIETQRRSAGANKKPAVATVEVNKEVHRFLVGEMREDIALKVESTNARLKQEGGHVPHTELVLKPLSQKE
KEDMLCGHAEKLALAFGLLNTPSGTPLVVTKNLRMCNDCHSSTKIMSRVEQRDIVVRDGYRVHRFTNGCCSCKGRP

2. g18202.t1 gene=g18202

MATLWQPDPLWTQQQIGVCTGIPRGAPAHLASGSPPQENNQRSRCSSIRSPAIPVPVKSVAGNGSSSDGRRQELEHLWQ
QFFLDPAMWWDNRLDKKRSRSPDFRHKGSGEVLWLNNLLKPAWVHEQLVSLDMKMGRGASPDKAEPGRRAGSGMG
VVTDGTVSMLCQQGKLSEAVGALESLAQTDSAISQFLFYTVLKLCTAKKDLVRGRRVHDLIVKLGYDSNAYMGNHILR
MYACGGRLQEAMDVFTRVGRPDAFMWSSIISAYAKHGKPKEALQLYHEMRRSGVKADAHVYVAALQACAFAADLES
GREVHCDVEQSGISTNLFVGNCLVDMYAKCGKLKDARQVFDSMATKDVVTWNALISGYAKRGELKDARQVFDSMNT
KNVVTWNALIAGYSEQGLGHEALVVYGQMEQEGLILADHVTFACLLKACASTGALQKGRQLHSLIVERGLLDTAVLG
NSLVDMYAKCGQLEDARQIFDSLTTKNVVTWNALIAGYAQQGLGQEALVLYEKMEHEGLILADHVTFACLLKACACT
GALQKGRHLQSLIVERGVMDNVVVGTCLVDMYAKCGELEDARQVFDSLATKNLVTWNALIAGYAEQGLGHEALVVY
GQMKQGGLISADHVTFACLLKACASTGALQKGRQLHSLIVERGLLDDIVLRTCLVDMYAKCGELEDARQVFDSLATKD
VVTWNALIAGYTQQGLGQEALVLYGRMEQEGLIFANHVTFACLLKACGSTGALQKGRQLHSQIVERGLLDHVVLGNC
LVDMYAKCGELEDARHVVDSMKTTDVVTWNALISGYGTSDDGHKALACFDEMLVAGIRPDGATFTCLLVACSHEGL
VTEGRKHFDAMAEQHGIKPTIEHYTCMVDLLGRAGQLDEAENMLLSCEGQTDVVALTSLLNACKSHRDLERALWCFE
SIVRLDPQWASAYVLMANTYADAGRWSDVDRIETLRMAAGANKKPAVATIEVNMEVHRFLVGERREDIALKLESTNA
RLKQEGGHVPHTKLVLKPLSEKEKEDALCGHAEKLALAFGLLNTPSGTPLVVTKNLRMCSDCHSSTEIMSRLEQRDIVV
RDGYRVHRFANGCCSCKGRP

3. g16507.t1 gene=g16507

MATVWQPDPLWTQHVGVCTGILLGAPAHLSLGSQLLANKASRCSRVRSPAVPLPRKCVGRDGFSSDGRCQEEQEQEHL
WEQFFLDPTMWWDNRRDKKSSRSPDFRRKGSGEVLWLNNMSKPAWVDEQLLCLDMQMGHGDSPSQSSSPDKLESGR
LEGSEMGSRCSSSVRSPALPLPIKCVGGDGFPNEGLRREHEHLWQQLFLEPTMWWDNRRDKKSSGAPDFRRKGSGEVL
WLNNLLKPAWVDEQLVSLDMQMGPGASPSKSRSPDKLEPRRRKGSEIGVVTAGTVSMLCQQGKLSEAVGALESLVQA
DSAISQSLFYSVLKLCTAKKDMVVGRRVHDLTVKLGYDLNAYLGNHILRMYACGGRLREAMDVFTRVAKPDAFLWSS
IISAYAKHGKPNEAIQLYHEMRRSGVKADAHVFVAALQACALAADLESGRKVHCDVMQSAFSTNLFVGNCLVDMYA
KCGELKDARQVFDNMATKDVVTWTVLIAGYSQQGLGHEALVVYGQMEQEGLVLANHVTFACLLNACASTGALQKG
RQLHSQIVERGLLDNLLVGNCLVDMYAKCGELEDARQVFDSMNTKDVVTWTTLIAGYAERGLGHEVLVLYGQMEQE
GLILANHVTFGCLLKACASTGALREGRQLHTQIVERGLLDNVVLGTCLVDMYARCGELDDARQVFDSVNTKDVVTWN
ALIAGYAQQELGHEALALYGQMEQDGILANDITFACLLKACASTGALQKGRQLHSQIVKKGMLKNAVLGTCLVDMYA
KYGELEDGRKVFDCMDRKDVVTWTALISGYGTCDEGHKALACFEEMLIAGIQPDRATFTSLLAACSHEGLVHEGRNLF
DSMAEQHGIKPTLEHYNCMIDLFGRAGQLGEAEKMLLSYEGQGNIVGLTSLLSASKSHGDIERAMWCFERIVRLDPQNA
SAHVLMANAYADAGRWTDVDRIETQRMVSGANKKPAVTTIEVNKQVHRFLVGERREDIALKVESTNARLKQEGGHLP
HTKLVLKPLSEKEKEDALCGHAEKLALAFGLLNTPSGTPLVVTKNLRMCSDCHSSTEIMSRLEQRDIVVRDGYRVHRFA
AGSCSCKGRP

DRH type

4. g17021.t1 gene=g17021

MATQVMGSNLWVQFPQLANSPMLAASAYKQPCSTTGWGVSSSAMAFASTRNSSCTTLDDRLTQKERQELLWQHFFLD
PVEWWDNRAENRSSASPRQQHPDFRHKTSRLSLWLDSSHKPDWVDSQLARWDQGRRPEELLRLSEAVEALDVLVQRG
WGRHVTPALFCRLLKRGLAQNDLAFTRRLHSLAVASGHGSNAFLANHMIGMYAAHGKLEDAMQAFSQVPVPDVFMW
SSIILAHARHGQPAEGIKLYREMQELAEVEPDNRLYVAALTACAAAQDLPAGKQVHADIQASAGKADIFLGNSLVNMY
AKCGSVVDAREVFDTMPRRDVVTWTAMIKGCTQQGLGRDALALYTQMQQEGVTPANRVTYMCLLQACANVGALQQ
GKELHLKIREKGLEADAAIGNSLVNMYAKCGSLEDACAVFDSLPRRDVVTWTAMIAGYSQRGLGQDALALYSRMQLE
GITPANHVTFVCLLQACASVAALQQGKQLHSQIRERGLEADEVIGNCLVNMYAKCGSLEDARRVFDRLPTRDLVTWTA

LLTGYAEHGDGHMAIRCFEGMLQQGVEPDDTTFTCLFVACSHAGLVHQGQRYFRMMVEDYGIVPTDKHYSCMVDLL
GRSGQLEEAESMLRSRSFVQDVVGWTALLTACKSYGDVERGRRCFDHLLELEADDASSYVLLENIYAKAGRWDDVDR
IENLRKMAGITKKAAKACIEVQNEVHEFTVGEDRSDISSKLRSVNMRLKEEGGHVPQTQLVLKAMSEEKKEDALCGHA
EKLALAYGLLNTPDGTTLVVTKNLRMCNDCHSSTKIMSRLEKREIIVRDAHRVHRFLDGACSCRDRH

5. g17318.t1 gene=g17318

MAAWRVARHSHRATTASLDPTKTNWWDNLRRESQEALGFYKPPWVGAQLADLDMPAGYQRGVTAETVSQLCQHGQ
LTDGVEALGVMVQTQSCVPQSLFYTVLKQCAVQRDLVLGRRVHALTVKGGYESDTFLCNHLLRMYATHARLQEAMD
VFAVVSEPDAFTWSAIISAQVKYGQDKQAIQLYHQMCRSGVKPDGHVFVAALQACARGADLETGQQVHAHVLASGVE
PDMFVCNCLVDMYAKCGSLEDAHRVFDGLGQKNVVTWTSLIAAYSQQGMWQEALGLYRAMQQEGMVAPDSVTFVY
LLQACTSVGAPALETGKQLHAQIRDRGLQSDMFVATSLVNMYAKCGSLDDAQTMFDGLPQKDVVMWNAMIAGYSQ
QGMGQEALQLYSAMQEEGITPDNRVTFVSLLQACASVGDLEQGKELHLLIRERGFEADAVIVSSLVNMYAKCGSLGDA
RKVFDTLPRKDVVTWNAMIAGYAQQGHGREALTLYGNMQHEGINPANHVTFLCLLQACASVGALQQGKQLHSEIRER
GLEGDAFISSCLVDMYSKCGDLEDAGKLFDSLPRRNLVTWNALLNGYAQHNDGHMAIRCFEDMLQQGVQPDETTFTC
LLVACSHAGLVQEGQRYFRMMVEEHGIVPSDYPLSCMVDLLGRSGQLDEAEHLLRSASFVNDVVGWTALLTACKGYG
DVERGRRCFDHLVKLAPEDASAYVLLGTIYANAGRWDDVDRIESMRKSAGAMKKAAKACIEVTNEVHEFTVGEERSDI
SPTLMILNSRLKKEGGHVPQTQLVLKAVSELEKEDALCGHAEKLALAYGLLHTPAGTTLLVTKNLRMCNDCHSSTKIM
SRLEKRDIIVRDAHRVHRFVDGSCSCGDRH

6. g18409.t1 gene=g18409

MVRQLELALVPDCRLWDSGKLVTQPRSAAFHMPKLLSAHPSSSGLRGGFRFRNRPEQYNASTERISEKHRQELLWHHF
FSHPSEWWDNRADKESPRHPDFRHKTSDEPIWLSNSQKPPWVDSQLAEWDAQMRGARASMEHSMAQALRKRCQEAD
LNEAVGALDLLVQGGCHVTTAVFYRLLQRCMAHKNLACGRLVHALALKYGYEGNAFLATHIIRMYASLGKLDDAGL
VFTKVSRPNLHMWSCTILAHARQGQPTQAIDLYRQFRVTGVEPNNHIFVAVLKACASAKDLVSGKEVHGDIQACAMM
DDVYIGNSLVNMYAKCGGLNDARQVFDNLPCKNVVTWTAMIAGYTQHGLGQQALALYGTMQQEGISPADHVTFVCL
LQACTSSRSLQQGKQLHLQLREQGLEADAVIGNSLVNMYAKCGSLEDARNLFDRLPTRDVVTWNSMIAGYTQQGLGQ
EALDLYASMQHEGKTPADHVTFVCLLQACANVGALQQGRDLHSKIRATGLEAHAVTGNGLVNMYAKCGRLEEALTV
FHNLPTKDVVTWTAMLNGYAEHGDAQLAIQCFQQMLQHGIQPNETTFLCLLVACSHAGMVHEGQKYFRIMVEDYGIP
ATDSHYNCMVDLLGRSGRLAEAEHVLQTRSCEQDEVGWKSLLTACRSYGDVERGQRCFDRVVKLVPEHASSYVLLEN
IYADAGRWDDVCRIERLRKSAGAVKKPAKACIEVQKALHEFTVGEARSDLTPELTSAYARVKDEGGHIPHTDLVLKAL
SEQEKELALCGHAEKLALAYGLLNTPEGTTLLVTKNLRMCTDCHSSTKIMSRLEKREIIVRDAHRVHRFADGSCSCRDR
H

7. g18545.t1 gene=g18545

MERLQSRWAIQCDSRSSVWSRIGGHREQSGFRFYSRLGAKLRLGGNALVVARISEKERQELSCQHPPQRWDNRPAERH
VYSGRVAIDIGIGSTSGIANYGSMLETLPLREAVEALELLVQGAQDVHTSFFYTCLRRCMAHKDLGLGARVHALAVKSG
YESNAFLANHIMGMYASHGKLTDSVQVFAKVTAPNAHMWATIILAHARHGQATQAIQLYRQMMDSAVRPDDRIFVA
VLKACAAAQDSVFGKEVHAHVLASDLEGDVFVANSLVTMYAKCGDLEAARMVFDSLGRRDVVTWTAMIAGYMEKG
LGQDALALYASMQQEGTIPADSVTFVCLLKACASVGALQQGKQLHAEIRERGLESHVCVGSSLVDMYSKCGSLEDARK
LFDRLRTRNLLTWTAMLNGYAEHSEGHRAIRFFEDMLEQDILPDDTTFVCLFVACSHAGLVHQGQRYFKTMVEEYGIV
PTDHHYSCMVDLLGRSGHLDEAERMLQSTSFVNDVVGWKALLSACKTHGDVQRGQRCFDHVVQSEPGNASSYVLLG
NIYANAGRWDEVDRLENLRKSAGAAKKPAKACIEVKNEVHEFTVGEERSDVSPKLRSVNTRLKEEGGHVPQVQLVLK
AVSEQEKEDSLCGHAEKLALAFGLMNTPDGTTLLVTKNLRMCNDCHSSTKIMSRLEKREIIVRDAHRVHRFLDGSCSCR
DRH

8. g10823.t1 gene=g10823

MLSASQAWFVDGRHVGNRICQLGATAVGTVSVSRGAAEAGQGQKEWQEVLWQDFFSDPWQWWDYRADAKSPQYP
DFRHKTSKQALWLNNVYKPEWVDEELARRDVGLSKFTTSMVETLSRHCENGAVGEAVEVLELLVQRGCSVSLSVFRL
VLKNCGSARDVGLGKRVHALAVRSGYESNAFLANHVICMYACHGELEEAVQVFGKVPVPDAYMWSSMILAFARQGK
PAEAIRFYLQMRESGVEPDNHIFVAVLKACANAADLVSGKQVHADIMTSGVRPNVFVGNSLVNMYAKCGSLDDAREV
FEGLETRDSVTWNAIITGYTQQGMGGESLALYSRMLQEGIAPADQATFVCLLKACASAGALQEGRQLHELIQKRRLEG
DVVVGSCLIDLYSQWGSVEDARRVFDGLPTKDVVAWTAMITGYAQQGMAEEALFLYESMLQEGITVPNRVTFLSLFQ
ACATVGAVQQGMQLHAQIRERGLETDVLVGNCLVDMYAKCGRLEDARSVFDALPIRDVVTWSALLNGYAEHSDGHM
AFQCFREMLQQGVKPNGTTFSSLLVACSHAGLVLEGLQYFRMMVDDYGIVPDDHHYSCMVDLLGRAGRLDEAENILL
TMSLDSYIVGWTSLLSACKSYGDVERGERCFQRLVEIEPEVATPYVLLCSMYANAARWEELDHIESLRKEAGAWKKPA
KACIEVKNRVHEFTVGEERSDVASMLRDVSTKLKLGGHVPETWLVLKAASEQEKEDALCGHAEKLALAYGLLNTPDG
TTLLVTKNLRMCHDCHSSTKIMSHVENREIIVRDVHRVHRFLNGACSCGDRH

DYW type

9. g16688.t1 gene=g16688

MSRIGALESHMKVKVRILMKSNLSPYIIMEEWFRRLSTLRSNTKGQEDELFRQEALLGYPRLRKGDKNYTGSVEMSRSD
GLWIDWEGVSSLHSVHYGGELESSGSVKATKEGKHLQVSTFETESEASVCVESSLMGISSKRVGRPELWMTPADVSSLC
EEGRLRESVQALNVLEHQGMQAHPDTYYHLLQQCIKRKAWQDGLQVHAHMKRSGLEVNSLLAGSLVRMYASSGKIL
TARQIFDQMSRHDVFTFTALMKGYLSCGQADKVLDLYKNMKDEGVKPDKFVFTVVLNACTSLGNIQEGRQTHAEIIKV

WSETDVIVDNCLLDLYAKGGSMEEASSVFDRMHERDVVSWNTMIAGYTRSGEGSKALRLYRQMQEAKVNADGSTFV
AVLNACASMADGGFGKQVHAQIARSRWESDMYVCNALVDMYSKCGNLQAARTVFDNLQTRDVVSWNAMISGYSRH
GYNEEVLHVFRQMLHKGVKPTSITFVGILNACASLGALEDGEQLHAHVRNCGLESQIFVGSALVDMYCKCGFIAGARA
VFDRMPKRDVVSWTSMFSGYADHGLAKEALCLFEEMQQEDVKPNSGTIVSVLNACGNAAALAEGKEMHTHASRCGF
ASDLCVCTALISMYGKCESMEDALEVFNGISDKDAVAWTAMLTAYIHNRRERDALQVFKEMLQKNVQPTDVTFVCVL
NACANLAVLGEGKRIHAMIDRSGHRSRIFVNNALVDMYAKCGRLDIARMVFDKMPRRDIISWNALMAAYTQHGRSKE
ALIIFEHLLQMEVELDYITFVTVLAACSHAGLVKEGCSYYRSMILDHGISPTEGHHVCMVDLLGRAGRLDEAEEFITNLS
AQPGVAVWMALLGACRLHGNVGIAEHAAERVLQLDPSHDAAHVLLANTYAAAGMWREKLAVRRLLKDKGLKKNPG
CSWTEIKNEIHLFFAEDKRHPQTDIIYDTLDKLIGTIKEAGYVPNTSFVLHDVDEKEKERCLRYHSEKLTIAFALINTAPGT
PLLILKNLRMCGDCHNAAKYISKVVGREIVTRDVSRFHHFKDGECSCRDYW

10. g10204.t1 gene=g10204

MLGRLPTLCVSQSVMAYKISRCLAVSYRVPSPTWRMPRNWVFAFEVERHLGHFAGENDTVAERGWNKLSRDGLEKV
VDWLFRSWLNPFCPTSELAVEGNAAHVLLADCEASGEQHERVTAYHPPFPSVGDIESLCREGRLTEAIGALDALEQRGV
RLHSELIAQVLQECALKKSLADGKHIHSYILRSGLESSPFLAGHVIRMYASCGRMLEARKVFDNLPKQKVFAWTALMK
GYVSHGQAIETLKLFKQMDTAGVKPDKFIFVTIVNACARLVDLEEGKRVHAFIHNSGSELDIYVENALIDMYAKCGSIED
ARQVFDKMQRRDVVSWNAIISGYARSGLVEESLKLYQQMQQANVRPNAVTFACVLNGCASRAALKEGTEVHAQTRK
NGMDSDVFVGTALVDMYTKCGMLEEAFKAFCTVPVRGVFTWNAMLRGCVEHGQGKEALRLFHQMQLSSLQPDALTF
VYVLKACASIPALDEGQNVHRLLSAKGFASDVFVGNALMDMYARFGKVKEAQDVFSMLPREDVASWNGMLKVYLSC
GYNDMVLRLFHNMQQEGVEPDQASFVLVSNACGNAATLEDGRRLHMQITQNGLDSDVLVGTALADMYGKCGRMEV
ARQIFDKLPKQDVVLWNVMIKGYAQQGLCKDALMLYKAMLKENVEPDSVTFLAVLGACVILGDFDEGRRLHVLLTER
GYGTNIVLANALIDMYARCGSMEDAYLVFDSMPTRDRITWNSMIKGYAQLGQGKKALENFDNMQQAGLKPDSLTFVS
VLNGCASIAALEEGKQIHDQINASGNESDISVGNCLLDMYVKCGDLFAARHTFEKLVKKDVVSWSTLLVGYAQHGCAQ
ECLQLFEQMQEDGVKPSMIAVVSALNACASRVALKQGKQIHDLIREGDFHANLFVGNALVDMYAKCGKPEVARQVFD
LMPERDVVSWNTMIAGYGHNGLGQDALNLFEQMKLQNVKLNHITFLSVLSACNHKGLTDEGCKHFESMTREYGICPT
NEHYACLVDLFGRAGRLDEAEQIIDQMPVQPSMAVWMALLAASRIHNNIKLAQRAAKHVIELEPQKASAYVLLANTYA
ALGQKEEESKVRNLMTERGVRKMPGYTWIELHNQTHQFVAGDKSHPRTEAIYAEVERLSGQMQMAGYVPDAVSELH
DAEEEQTEDSACFHSERLAMAFALISSAPGTPIHIVKNIRVCRDCHQATKIISRITGREIIARDPSRFHHVKDGVCTCGDY
W

4.4. Discussion

To our knowledge, this is the first experimental report for RNA editing by cloning the Hornworts specific DYW-type domain variants of PPR genes. One main question concerning RNA editing in plants is the nature of the factors responsible for so-called ‘reverse’ editing (U-to-C editing). The U-to-C RNA editing in plant organelles is frequently referred to as “occasional”. It was already clear that C-to-U and U-to-C editing are not correlated in any way, even in plants capable of both processes, suggesting independent mechanisms and presumably different trans-acting factors. U-to-C RNA editing has originated from the more ancient and widespread C-to-U editing, using the same mechanisms for RNA target recognition linked to a biochemical enzyme variant, possibly converting a deaminase into a transaminase. Given the likely earlier evolutionary origin of plant C-to-U RNA editing among land plants it is suggestive that PPR proteins remain at

the core of target recognition also for sites of U-to-C editing. We developed a bacterial expression system, in which we cloned the Hornworts specific PPR proteins with DYW domain truncated PPR56, *Physcomitrella patens* (moss) editing factor. The bacterial assay system, allowed to study RNA editing by hornworts specific PPR genes with its potential target sequences. We studied three different variants of C-terminal PPR proteins of hornworts GRP-type, DRH- type and DYW-types.

Out of these, variants of GRP and DRH-type, having no homologous in other taxa, are considered to be the strong candidate for Uridine to Cytidine RNA editing enzyme. In this research, we developed a bacterial expression system for C-to-U RNA editing by Hornworts DYW domains with moss PPR56. Similarly, we have also studied the RNA editing events by expressing the Hornworts DYW variants with PPR56 into animal HEK 293 cells. We observed the C-to-U RNA editing (30-40%) by potential PPR-DYW domain expression system. However, no RNA editing event could be observed for the other PPR derived variants of hornworts. In this way, we can conclude that DYW domains of Hornworts PPR proteins can be enough for efficient C-to-U RNA editing.

Gene expression levels in the transfected HEK cells and the transformed bacterial BL21 competent cells revealed that the expression of GRP-type and DRH-type domains were comparatively lower than the DYW-type domains. The DYW:KP might be something toxic (unwanted) for cells, thus the chimeric PPR56 proteins expressed less effective. In this way, expressed proteins might not be enough for enzymatic (transaminase) activity to cause the U-to-C RNA editing.

Alternative explanations for no U-to-C editing with the DYW:KP proteins can be missing of some other unknown essential hornworts specific RNA editing factors. DYW:KP may need some other unknown co-factors, likewise we have MORF, ORRM for DYW domains in

angiosperms [40–42]. Thus, an in vitro pull-down assay can be used to investigate the co-factors of the DYW:KP domain. We may be able to reproduce U-to-C RNA editing with the DYW:KP type PPR protein and the newly isolated co-factors.

This ineffective U-to-C editing with PPR56-GRP or -DRH proteins in *E.coli* may suggest that some other unknown hornworts specific RNA editing factors are essential. GRP and DRH type proteins may need some other unknown co-factors, likewise MORF and ORRM in Arabidopsis Thus, co-immunoprecipitation assay can be used to identify the other co-factors involved in hornworts. Interaction with identified candidate proteins will be verified by Yeast-2-hybrid and CoIP. Furthermore, reconstructing the U-to-C RNA editosome by co-expression of multiple candidate proteins in *E.coli* may also be suggested. To search for fusion partners of PPR56 as U-to-C RNA editing enzymes, we may analyze the mutant lines for the GRP and DRH type PPR genes in hornworts. The RNA interference as well as CRISPR based gene knock-out or knock-down lined for specific GRP or DRH type genes [43] need to be analyzed.

4.5. Reference

1. Knoop V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cellular and Molecular Life Sciences*. 2011;68(4). doi:10.1007/s00018-010-0538-9
2. Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. *Genome Biology*. 2017;18(1). doi:10.1186/s13059-017-1347-3
3. Wolf J. tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *The EMBO Journal*. 2002;21(14). doi:10.1093/emboj/cdf362

4. Bar-Yaacov D, Pilpel Y, Dahan O. RNA editing in bacteria: occurrence, regulation and significance. *RNA Biology*. 2018;15(7). doi:10.1080/15476286.2018.1481698

5. Liu H, Wang Q, He Y, et al. Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes. *Genome Research*. 2016;26(4). doi:10.1101/gr.199877.115

6. Benne R. RNA-editing in trypanosome mitochondria. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*. 1989;1007(2). doi:10.1016/0167-4781(89)90031-6

7. Chen S, Habib G, Yang C, et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science*. 1987;238(4825). doi:10.1126/science.3659919

8. Thomas SM, Lamb RA, Paterson RG. Two mRNAs that differ by two nontemplated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5. *Cell*. 1988;54(6). doi:10.1016/S0092-8674(88)91285-8

9. Cattaneo R, Kaelin K, Baczko K, Billeter MA. Measles virus editing provides an additional cysteine-rich protein. *Cell*. 1989;56(5). doi:10.1016/0092-8674(89)90679-X

10. Covello PS, Gray MW. RNA editing in plant mitochondria. *Nature*. 1989;341(6243). doi:10.1038/341662a0

11. Gualberto JM, Lamattina L, Bonnard G, Weil J-H, Grienenberger J-M. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature*. 1989;341(6243). doi:10.1038/341660a0

12. Hiesel R, Wissinger B, Schuster W, Brennicke A. RNA editing in plant mitochondria. *Science*. 1989;246(4937). doi:10.1126/science.2480644

13. Tillich M, Lehwark P, Morton BR, Maier UG. The Evolution of Chloroplast RNA Editing. *Molecular Biology and Evolution*. 2006;23(10). doi:10.1093/molbev/msl054

14. del Campo EM. Post-Transcriptional Control of Chloroplast Gene Expression. *Gene Regulation and Systems Biology*. 2009;3. doi:10.4137/GRSB.S2080

15. Stern DB, Goldschmidt-Clermont M, Hanson MR. Chloroplast RNA Metabolism. *Annual Review of Plant Biology*. 2010;61(1). doi:10.1146/annurev-arplant-042809-112242

16. Ruchika, Okudaira C, Sakari M, Tsukahara T. Genome-Wide Identification of U-To-C RNA Editing Events for Nuclear Genes in Arabidopsis thaliana. *Cells*. 2021;10(3). doi:10.3390/cells10030635

17. Ruchika, Tsukahara T. The U-to-C RNA editing affects the mRNA stability of nuclear genes in Arabidopsis thaliana. *Biochemical and Biophysical Research Communications*. 2021;571. doi:10.1016/j.bbrc.2021.06.098

18. Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*. 2014;157(6). doi:10.1016/j.cell.2014.05.010

19. Heidenreich M, Zhang F. Applications of CRISPR–Cas systems in neuroscience. *Nature Reviews Neuroscience*. 2016;17(1). doi:10.1038/nrn.2015.2

20. Azad MTA, Bhakta S, Tsukahara T. Site-directed RNA editing by adenosine deaminase acting on RNA for correction of the genetic code in gene therapy. *Gene Therapy*. 2017;24(12). doi:10.1038/gt.2017.90

21. CHESTER A. Optimization of apolipoprotein B mRNA editing by APOBEC1 apoenzyme and the role of its auxiliary factor, ACF. *RNA*. 2004;10(9). doi:10.1261/rna.7490704

22. Bhakta S, Sakari M, Tsukahara T. RNA editing of BFP, a point mutant of GFP, using artificial APOBEC1 deaminase to restore the genetic code. *Scientific Reports*. 2020;10(1). doi:10.1038/s41598-020-74374-5

23. Knie N, Grewe F, Fischer S, Knoop V. 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(1). doi:10.1186/s12862-016-0707-z

24. Barkan A, Small I. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology*. 2014;65(1). doi:10.1146/annurev-arplant-050213-040159

25. Oldenkott B, Yang Y, Lesch E, Knoop V, Schallenberg-Rüdinger M. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in Escherichia coli. *Communications Biology*. 2019;2(1). doi:10.1038/s42003-019-0328-3

26. Sasaki T, Yukawa Y, Wakasugi T, Yamada K, Sugiura M. A simple *in vitro* RNA editing assay for chloroplast transcripts using fluorescent dideoxynucleotides: distinct types of sequence elements required for editing of *ndh* transcripts. *The Plant Journal*. 2006;47(5). doi:10.1111/j.1365-313X.2006.02825.x

27. Salone V, Rüdinger M, Polsakiewicz M, et al. A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Letters*. 2007;581(22). doi:10.1016/j.febslet.2007.07.075

28. Schallenberg-Rüdinger M, Lenz H, Polsakiewicz M, Gott JM, Knoop V. A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. *RNA Biology*. 2013;10(9). doi:10.4161/rna.25755

29. Gerke P, Szövényi P, Neubauer A, et al. Towards a plant model for enigmatic U-to-C RNA editing: the organelle genomes, transcriptomes, editomes and candidate RNA editing factors in the hornwort *Anthoceros agrestis*. *New Phytologist*. 2020;225(5). doi:10.1111/nph.16297

30. Gutmann B, Royan S, Schallenberg-Rüdinger M, et al. The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants. *Molecular Plant*. 2020;13(2). doi:10.1016/j.molp.2019.11.002

31. Ohtani S, Ichinose M, Tasaki E, Aoki Y, Komura Y, Sugita M. Targeted Gene Disruption Identifies Three PPR-DYW Proteins Involved in RNA Editing for Five Editing Sites of the Moss Mitochondrial Transcripts. *Plant and Cell Physiology*. 2010;51(11). doi:10.1093/pcp/pcq142

32. Zhang J, Fu X-X, Li R-Q, et al. The hornwort genome and early land plant evolution. *Nature Plants*. 2020;6(2). doi:10.1038/s41477-019-0588-4

33. Takenaka M, Takenaka S, Barthel T, et al. DYW domain structures imply an unusual regulation principle in plant organellar RNA editing catalysis. *Nature Catalysis*. 2021;4(6). doi:10.1038/s41929-021-00633-x

34. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873). doi:10.1038/s41586-021-03819-2

35. Crooks GE HGCJBS. WebLogo: A sequence logo generator, Genome Research, . Published online 2004.

36. Lei L, Chen H, Xue W, et al. APOBEC3 induces mutations during repair of CRISPR–Cas9-generated DNA breaks. *Nature Structural & Molecular Biology*. 2018;25(1). doi:10.1038/s41594-017-0004-6

37. Blanc V, Xie Y, Kennedy S, et al. Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. *RNA*. 2019;25(1). doi:10.1261/rna.068395.118

38. Grünewald J, Zhou R, Garcia SP, et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature*. 2019;569(7756). doi:10.1038/s41586-019-1161-z

39. Meng Y, Chen D, Jin Y, Mao C, Wu P, Chen M. RNA editing of nuclear transcripts in Arabidopsis thaliana. *BMC Genomics*. 2010;11(Suppl 4). doi:10.1186/1471-2164-11-S4-S12

40. Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Hartel B, Brennicke A. Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *Proceedings of the National Academy of Sciences*. 2012;109(13). doi:10.1073/pnas.1202452109

41. Takenaka M. How complex are the editosomes in plant organelles? *Molecular Plant*. 2014;7(4):582-585. doi:10.1093/mp/sst170

42. Bayer-Császár E, Haag S, Jörg A, et al. The conserved domain in MORF proteins has distinct affinities to the PPR and E elements in PPR RNA editing factors. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2017;1860(8). doi:10.1016/j.bbagrm.2017.05.004

43. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*. 2021;12(1). doi:10.1038/s41467-021-21194-4

Chapter 5

Development of MS2 system with plant derived "DYW" type PPR protein

5.1. Introduction

There are various genome editing techniques for manipulation of genomic information in a targeted manner. [1–4] These methods include zinc-finger nucleases (ZFNs), transcription activator–like effector (TALE) nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated protein 9 (Cas9) [5,6]. These enzymes can be introduced into cells, and the technique known as genome editing with engineered nucleases. The most popular enzyme for the genome editing is CRISPR/ Cas9.[7,8]

Base editing is also a form of genome editing that allows direct, irreversible conversion of one base pair to another at a target genomic sequence without a double-stranded DNA breaks. Although these techniques are expected to find applications in the treatment of diseases, it remains difficult to achieve accurate genome editing in all affected cells. Moreover, incorrect genome editing has the potential to cause cancers and other diseases.

Therefore, we think genome editing is not a suitable technique for gene therapy for the treatment of patients as genome editing can potentially cause mutations and delivery process is also not convenient. However, on the other hand, RNA is expressed in all tissues and cells, and if RNA repair errors occur due to incorrect RNA editing, the mutated RNAs are quickly degraded and do not affect the genome sequence. [9,10] Therefore, for the patient treatment, RNA editing is more preferable than genome editing. Accordingly, we developed an artificial RNA editing system, based on the deaminase enzymes, for restoration of the wild-type genetic code in genetic diseases caused by T-to-C mutations. Some examples of such disorders include: ADA deficiency, cystic fibrosis, elliptocytosis, antithrombin III deficiency, and others. [11,12]

In previous studies, artificial A-to-I RNA editing was done by using ADAR family enzymes tethered to gRNAs complementary to the target sequence.[13,14] For tethering, Stafforst and colleagues used the SNAP-tag. [15–17] Montiel-Gonzalez and colleagues used Lambda-N protein and box B element RNA, and our group used the MS2 system.

Recently, one study has been published involving MS2 system for C-to-U RNA editing. It was demonstrated that an artificial RNA editing enzyme for C-to-U conversion could be designed [18,19]similarly to the A-to-I editase. For the purpose apolipoprotein B-mRNA editing (APOBEC) family member was used. [20] The catalytic polypeptide (APOBEC) and activation-induced cytidine deaminase (AICDA/AID) families are single strand specific cytidine deaminases, expressed in multiple cells and tissues, which catalyze cytidine-to-uridine (C-to-U) base substitutions in RNA, viral DNA, and genomic DNA. [21–24]

Using an MS2-tagged system, we previously restored the original sequence of a G-to-A mutation via A-to-I editing with the ADAR1 deaminase. [18,19,25] And also, we sought to restore C-to-U in the context of a T-to-C mutation using our artificial enzyme system along with a specific gRNA. Tagging with MS2 is based on the natural interaction between the MS2 bacteriophage coat protein and a stem-loop structure from the phage genome[26], which has been used for biochemical purification of RNA–protein complexes and combined with green fluorescent protein (GFP) expression to enable detection of RNA in living cells.[27] By using the phenomena in bacteriophage regarding the coat protein and stem loop, APOBEC 1 was bound to the MS2 coat protein and the gRNA bound to MS2 stem loop with a view to initiate the binding of the coat protein and stem loop thus allowing the gRNA to guide the deaminase at the specific location/site and perform editing at the targeted nucleotide sequence. Using the GFP point mutant blue fluorescence protein

(BFP) as a model target RNA, our artificial RNA editing system could successfully convert up to 21% C-to-U at the mRNA level, restoring the wild-type sequence.[28]

In plants, C-terminal E or EDYW domains extensions of PPR proteins are best known to have a conserved residues for cytidine deaminase enzymatic activity for C-to-U RNA editing.[29–33] In our previous work, we have investigated the RNA editing factors by cloning the hornwort specific PPR proteins into bacterial cells. In E. coli, expression of the two PPR proteins called PPR65 and PPR56 from the moss *Physcomitrella patens* leads to the editing of co-expressed respective targets from *ccmFC* and *nad3/nad4* transcripts[34]. Therefore, in this study, we opt to develop an artificial deaminase system for human cells by incorporationg the MS2 coat proteins with the deaminase domains from plant origins, *Physcomitrella patens, Arabidopsis* and, *Anthoceros.*

5.2. Methods and Materials

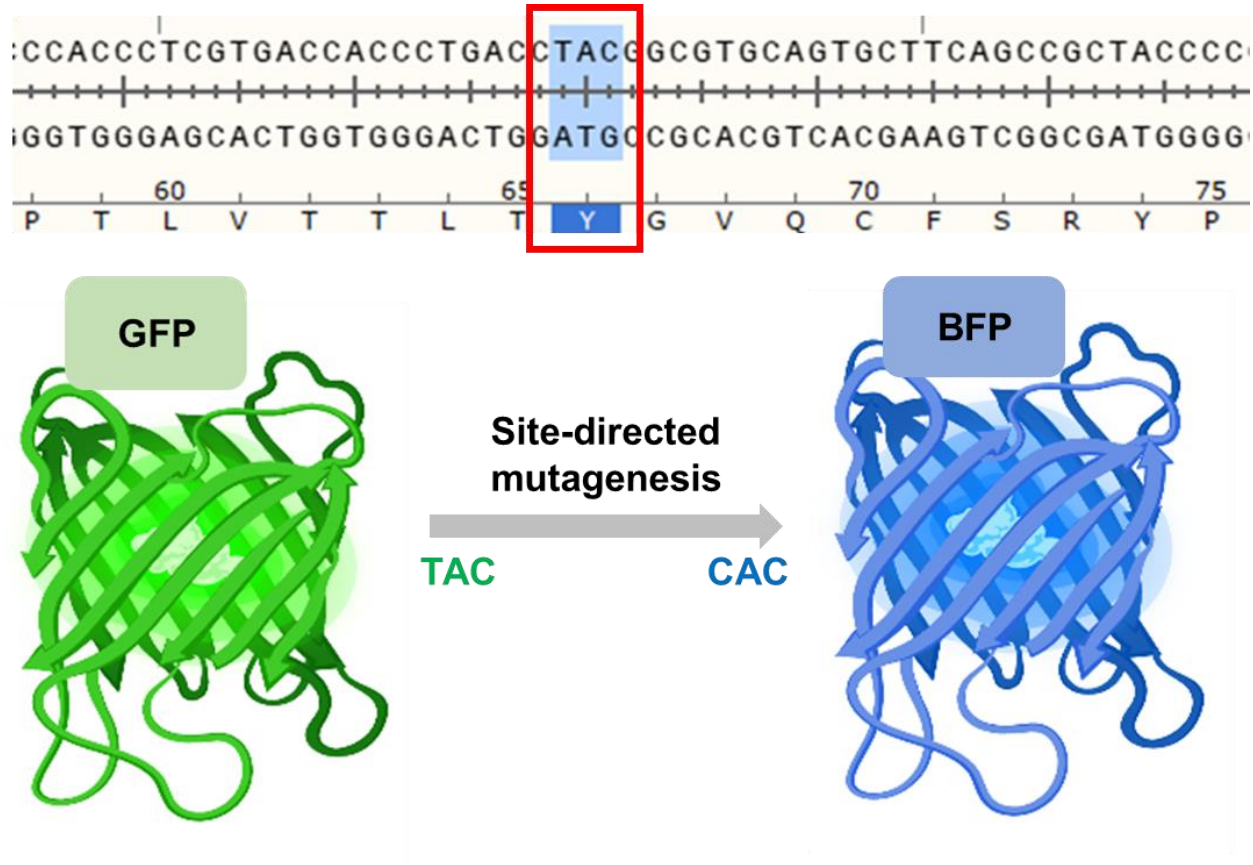


Fig 5.1. Preparation of the BFP target mRNA by point mutation at 199th position and stable transformant by transfection into HEK 293 cell line

5.2.1. Target plasmid (mutated EGFP or BFP) construct preparation

Using the pcDNA3-EGFP as backbone, performed the site-directed mutagenesis to convert the 66th amino acid codon TAC into CAC. For the site-directed mutagenesis, primer sets have been designed as[28] underlined letter indicates the site-directed mutagenesis position. KOD One PCR Master Mix (TOYOBO) was used to perform the site-directed mutagenesis PCR amplification consisting of the parameter of 15 cycles which included 10s for each denaturation at 98°C, 30s the annealing temperature of 60°C, and extension for 10s at 68°C. After PCR reaction, we performed

Dpn I (NEW ENGLAND Bio Labs) digestion of the amplification products. Added 0.5 µl of the *Dpn I* restriction enzyme (20 U/µl) directly to amplification reaction. Spined down the reaction mixtures in a microcentrifuge for 1 minute and immediately incubated each reaction at 37°C for 1 hour to digest the template plasmid. Transferred 5 µl of the *Dpn* I-treated plasmid DNA into 50 µl *DH5α* competent cells, mix the transformation mixture by pipetting the solution up and down several times. Then, incubating on ice for 10 minutes. Heat shocks the transformation reactions for 45 seconds at 42°C and then place the reactions on ice for 2 minutes. Added 400 µl SOC medium into transformation reaction and incubate the transformation reactions at 37°C for 1 hour at lab shaker. Plate the half volume of transformation reaction on agar plates containing ampicillin antibiotic for the plasmid vector.

5.2.2. Preparation of plant derived DYW deaminase plasmid construction

To target the enzyme to the BFP was chosen as a codon of interest; we cloned the *P.patens*, PPR56_DYW downstream of MS2 in pCS2+MT vector under the control of the pol II CMV IE-94 promoter, using the *XhoI* and *XbaI* (Takara, Shiga, Japan) restriction sites. The resultant plasmid was designated as pCS2+MT-MS2HB-PPR56_DYW. Plant derived PPR56_DYW were PCR-amplified for 5min initial denaturation at 95°C followed by 35 cycles each with 30s denaturation at 95°C, 30s the annealing temperature of 55°C, 2min synthesis at 72°C, and a final step of synthesis for 7min at 72°C, using primers with the appropriate restriction sites: *XhoI* catalytic PPR56_DYW Forward primer, and *XbaI* catalytic PPR56_DYW Reverse primer. Primers were designed using software DNADynamo https://www.bluetractorsoftware.com/DynamoDemo.htmPlasmid DNA was cloned into

Escherichia coli DH5∝ competent cells, and extracted using a QIAGEN Midikit (QIAGEN), with its identity confirmed by sequencing.

5.2.3. Preparation of the gRNA to direct the deaminase to target.

For site-directed deamination of the target RNA, the guide RNA was designed to anneal the sequence and mismatch the targeted cytidine. We used pCS2+ as backbone for guide construction. First, we designed primer for amplifying the MS2 stem loop sequence from pSL-MS2 6X (plasmid code: #27118) with guide RNA. The forward primer: ATCA**GAATTC**CACTGCACGCCGTTGGTCAGGGAATGGCCATG and reverse primer: ATTC**CTCGAG**CGCAAATTTAAAGCGCTGAT was used to perform MS2-gRNA PCR reaction. Underlined text represents the 21 base pair guide sequence. Highlighted text indicates the restriction enzymes sites such as *EcoR1* and *Xho1*. Next, the pCS2+ (addgene) was digested by *EcoR1* (TAKARA) and *Xho1* (TAKARA) restriction enzyme. After restriction digestion reaction, we performed agrose gel electrophoresis to separate digested plasmid and impurity. Then, extract the digested plasmid by NucleoSpin Gel and PCR Clean-up kit (TAKARA), confirming the concentration by NanoDrop-1000 and keep at -20℃.

5.2.4. Cell culture and growth

HEK293 cells from RIKEN BRC CELL BANK were maintained on 60 × 15 mm TrueLine Cell Culture Dishes (NIPPON Genetics) in Dulbecco's modified Eagle's medium (Nacalai Tesque, Kyoto, Japan) supplemented with fetal bovine serum (Thermo Fisher Scientific) at a volume ratio of 10:1 DMEM:FBS under 5% CO2 at 37 ◦C. Cells were used in experiments after at least three passages from frozen stocks. Site-directed mutagenesis pcDNA-EGFP had transfected into HEK293 cells and selected by 500ng/mL G418 (Geneticin) for 1 month. After selection, we

obtained HEK293 BFP stable cell line and maintained the cell line at 150ng/mL G148 for next step.

5.2.5. Transfection of plasmid DNAs into HEK cells

At 16–24 h before transfection, HEK293 BFP stable cells were seeded in glass bottom dish Cell Culture Plates (MATSUNAMI, Lot No. 190701) at a density of 5.5×10^5 cells/well. Before transfection, the medium was removed so that the volume per well was 0.5 mL, and 0.4 mL of fresh DMEM + FBS was added. Next, 50 µL of Opti-MEM (Thermo Fisher Scientific), 2.0 µL of 1.0 µg/µL PEI MAX (Polysciences, Illinois, USA), and 500 ng of plasmid (250ng Deaminase plasmid and 250ng of guide RNA) were mixed, incubated for 20 min, and added to each well. At 48h after transfection, total RNA from transfected cells was collected immediately after green fluorescence was observed using Juli light fluorescent microscopy .

5.2.6. Condition for the confocal microscopy

Transfected cells were observed for the GFP fluorescent on an FV1000D confocal laser-scanning microscope (Olympus, Shinjuku-ku, Tokyo, Japan) under optimized conditions. We designed our conditions to increase the effective resolution, dye selection, determination of the exposure time as well as the adjusted magnification. Setting the filter 1 is Alexa Fluor 405 for BFP, filter 2 is Alexa Fluor 488 for GFP and TD1 is for phase contrast as BFP was excited at 405 nm and, GFP at 488 nm wavelengths. The XY scanner is a vital point for taking the image as when its going from the upper level to the depth the intensity varied so at the best point it was fixed and the images were captured.

5.2.7. Total RNA was extracted from HEK293 cells using TRIZOL extraction method

Selected the cells having BFP expression from the transfected cells. Removed the media and wash by Ice cold PBS for 1 ml/well. For suspended cultured cells re-suspend the RNA in 1 ml of TRIZOL. Added 200ul of chloroform/ 1ml of TRIZOL, vortex for 15s and leave at room temperature for 2-3 minutes. Centrifuge the samples at 12000g for 15min at 2-8°C. Following the centrifugation there will be three phases visible within the tube. Transferred the aqueous phase (top) to a fresh tube. Added 500 uL of room temperature Isopropanol/1 ml TRIZOL to the new tube and incubate at room temperature for 10 minutes. Centrifuged the sample at 12000g for 10 minutes at 2-8°C. Following the centrifugation remove the supernatant. Washed the RNA pellet with 70-80% of 1 ml ethanol and vortex for 15s. Removed the supernatant and allow remaining Ethanol to air dry for 2-3 minutes. Re-dissolved the pellet in 10 ul of Autoclaved RNase-Free water. Further, complementary DNA (cDNA) was synthesized with a random primer with ReverTra Ace qPCR RT Master Mix.

5.2.8. RT-PCR analysis

PCR amplification assays contained 1μL template of cDNA, 0.4μM of each primer, 1× recommended PCR GoTaq buffer master mix, (Takara), and double distilled water in a final volume of 25μL. Amplification assays included 5min initial denaturation at 95°C followed by 35 cycles each with 30s denaturation at 95°C, 30s the annealing temperature of 55°C, 2min synthesis at 72°C, and a final step of synthesis for 7min at 72°C. For purification of PCR products, 2U ExoI (TAKARA) and 0.5U Shrimp Alkaline Phosphatase (TAKARA) were added and incubated in PCR machine at 37°C for 1h followed by 15min at 80°C and, then hold at 4°C. In a new PCR tube, add

11.5µL H20, 2.5 µL of forward primer, and 1 µL of PCR products (with ExoSAP). Sequenced by eurofins genomics. https://eurofinsgenomics.jp/jp/home/.

5.2.9. PCR-RFLP

For conforming the successful restoration of the GFP sequence, the above PCR products amplified using GoTaq buffer master mix (TAKARA) were digested by using a restriction enzyme that distinguished between the edited and nonedited DNA sequences. The PCR products were subjected to run in 6% PAGE, polyacrylamide gel electrophoresis followed by staining with SYBR Green dye (Invitrogen). Around 100 ng of cDNA was used for each PCR reaction, the total reaction volume was 20 ul. 8 ul of PCR product was used for restriction digestion, where the incubation was done at 37°C for 2-3 hours with *BtgI* (New England BioLabs) restriction enzyme, which cleaved the BFP sequence into two shorter fragments of 201 and 123 bp however, enabled to digest the restored GFP sequence. Equal volume (5 ul) of digested products were loaded into the 10 well comb. Imaging was done using the LAS 3000 image scanner. The presence of the intact 324 bp sequence confirmed restoration of C to U in the mRNA.

5.2.10. RNA-seq analysis

NGS data analysis was performed by GENEWIZ biotechnology co. LTD (Tokyo, Japan). Transcriptome sequencing experiments include RNA extraction and QC, library construction, purification, library QC and quantitation, as well as sequencing cluster generation and high through-put sequencing. Each step is important for data quality and quantity, which in turn affect the data analysis. To ensure the accuracy and reliability of the analysis results, every step is under strict monitoring and quality control. After mixing libraries based on their effective concentration and the required sequencing data volume, Illumina platform is used for high throughput

sequencing. The workflow for experimental steps involved in the library construction is shown in figure 5.2.

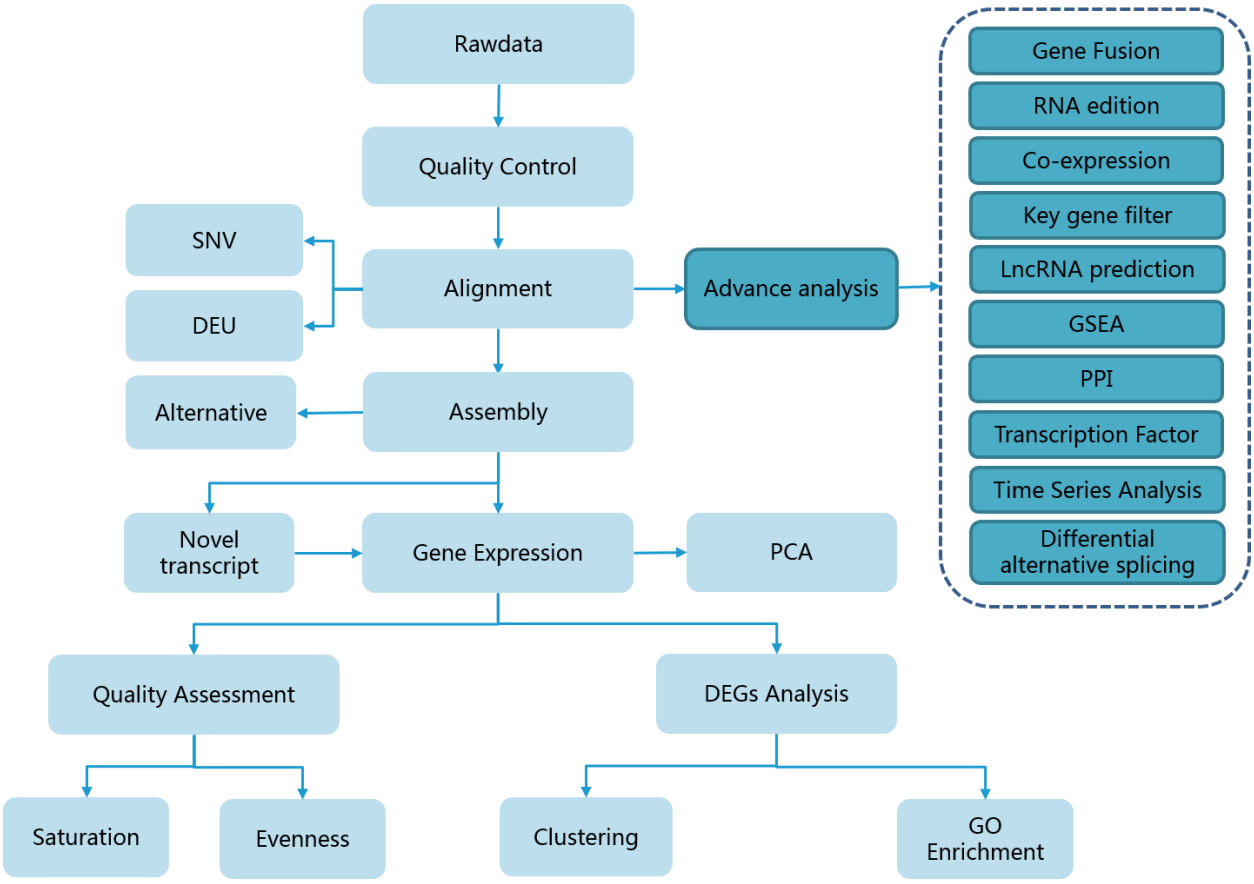


Figure 5.2. Flowchart showing the workflow of the RNA-seq analysis. From the RNA samples to the final data, each step, including sample test, library preparation, and sequencing, influences the quality of the data, and data quality directly

Data analysis

Quality Control: In order to remove technical sequences, including adapters, polymerase chain reaction (PCR) primers, or fragments thereof, and quality of bases lower than 20, pass filter data of fastq format were processed by Cutadapt (V1.9.1) to be high quality clean data.

Mapping: Firstly, reference genome sequences and gene model annotation files of relative species were downloaded from genome website, such as UCSC, NCBI, ENSEMBL. Secondly, Hisat2

(v2.0.1) was used to index reference genome sequence. Finally, clean data were aligned to reference genome via software Hisat2 (v2.0.1).

Expression analysis: In the beginning transcripts in fasta format are converted from known gff annotation file and indexed properly. Then, with the file as a reference gene file, HTSeq (v0.6.1) estimated gene and isoform expression levels from the pair-end clean data.

Differential expression analysis: Differential expression analysis used the DESeq2 Bioconductor package, a model based on the negative binomial distribution. the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions, Padj of genes were setted <0.05 to detect differential expressed ones.

SNV analysis: Samtools v0.1.19 with command mpileup and Bcftools v0.1.19 were used to do SNV calling.

5.3. Results and discussions

2.3.1. Development of MS2 system incorporated with *P. patens* PPR56-DYW

Genetical approaches with model plants e.g. *Arabidopsis, Rice, Physcomitrella patens* revealed that specific editing site recognition in plant RNA editing is governed by PPR proteins with C-terminal E1E2 or E1E2DYW domain extension[35].The C-terminal DYW domain has been considered to work as a catalytic domain for C-to-U editing due to the high similarity with cytidine deaminases. This hypothesis was strongly supported by the recently reported in *E. coli* and in vitro C-to-U RNA editing with a single DYW containing PPR protein from P. patens [36] In this study, we developed an animal expression system by incorporating the deaminase enzymes of plant organelles with the MS2 coat proteins. The *P. patens* PPR56 coding sequence including C-terminal E1E2DYW domain was inserted downstream of MS2 coat protein sequence in pCS2+MT vector

system under the control of the CMV IE-94 promoter, forming the recombinant plasmid as pCS2+MT-MS2HB-PPR56 (figure 5.3a). We also designed a gRNA with the MS2 stem loop (figure 5.3b), which in turn guided the PPR56 to reach the specific target nucleotide. As a target RNA, we used RNA encoding blue fluorescent protein (BFP) which was derived from the gene encoding GFP by a single nucleotide T-to-C substitution. Then, the two factors (MS2-PPR56 and gRNA) were co-transfected into the BFP stable HEK293 cells (figure 5.3c). the schematic for the interaction in between the MS2 proteins with the stem loop RNA guided by the gRNA is shown in (figure 5.3d).

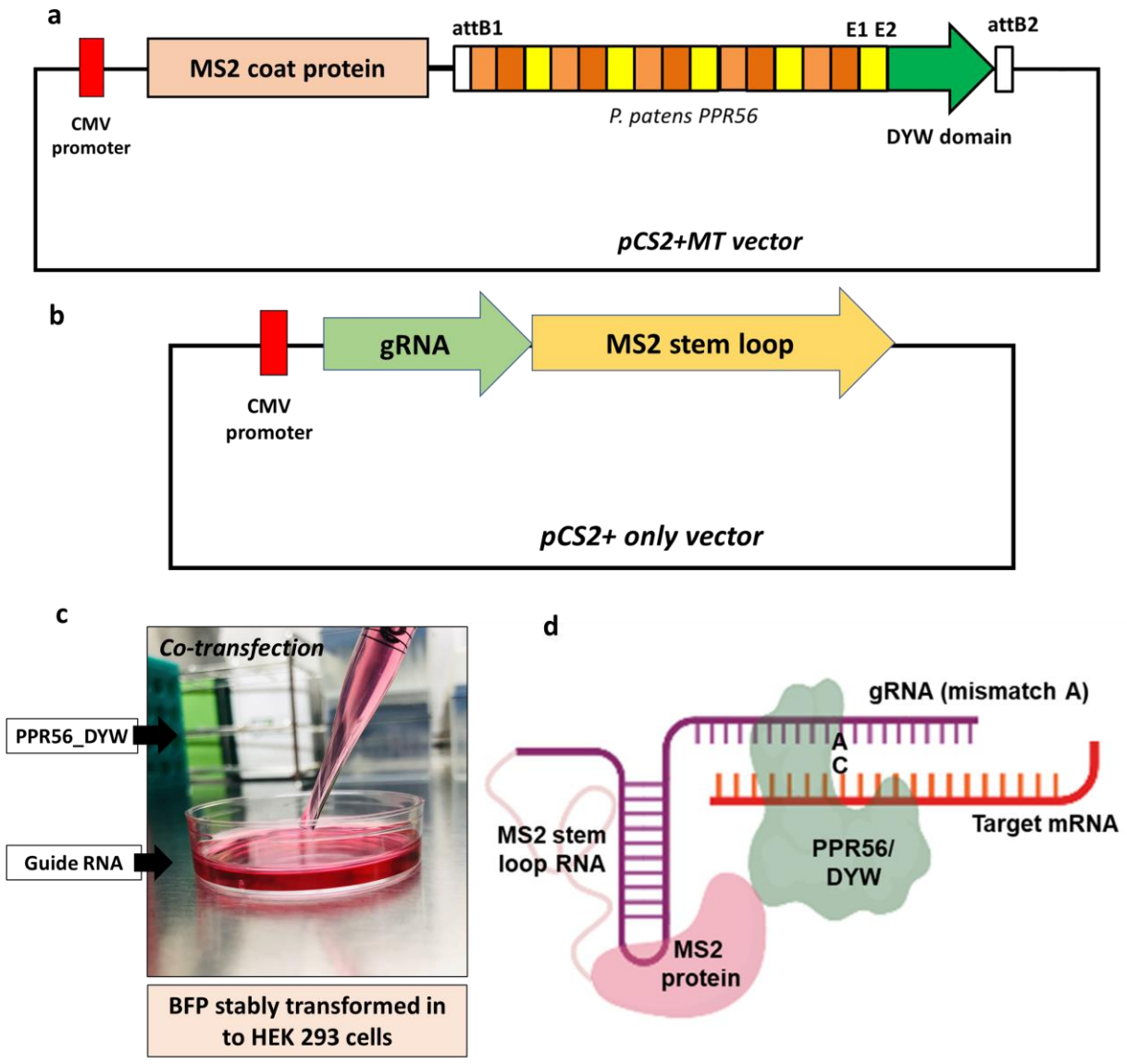


Figure 5.3. schematic for the preparation of MS2_PPR56DYW construct **(a)** PPR56_DYW was cloned downstream of MS2 in pCS2+MT vector under the control of the pol II CMV IE-94 promoter. The resultant plasmid was designated as pCS2+MT-MS2HB-PPR56_DYW, **(b)** preparation of the gRNA to direct the deaminase to target. **(c)** transfection of both factors into BFP stably transformed HEK cells. **(d)** Diagrammatic representation of C to U RNA editing by MS2-PPR56DYW system.

5.3.2. Analyzing the C-to-U RNA editing with DYW domains derived from *Arabidopisis* and *Anthoceros*

In bacterial system, the expression of the two PPR proteins called PPR65 and PPR56 from the moss *Physcomitrella patens* leads to the editing of co-expressed respective targets from *ccmFC* and *nad3/nad4* transcripts[34] . The E. coli RNA editing system demonstrated that single DYW-type PPR proteins can be enough for efficient C-to-U RNA editing without any additional plant organelle-specific factors[36]. Therefore, we used the PPR56 as the RNA editing factor, incorporated with MS2 coat proteins, and the MS2 stem loop with the gRNA, which guided the catalytic deaminase enzyme to the target sites. We successfully achieved the 85-100% efficiency for C to U RNA editing for the first time using MS2-PPR56DYW combination. The chromatogram for sanger sequencing showing the conversion of CCC (BFP) into CCU (GFP) is shown in Figure 5.4a. We further, modified the C terminal DYW domains of PPR56. We cloned the DYW domains from three different sources, such as *P. patens*, *Arabidopsis*, and *Anthoceros*. The raw sequencing data were analyzed using the DNADynamo and Sequence Scanner, version 2 (Applied Biosystems). The efficiency for the C-to-U RNA editing was calculated by measuring the area. At the target site, the edited and unedited products were presented together, a dual peak (C [unedited] and T [edited]) was observed. While the single peak for the T [edited] at the targeted site for C, confirmed the 100% conversion.

5.3.3. RFLP digestion

In the context of genetic code restoration, validation of GFP restoration by PCR-Restriction fragment length polymorphism (RFLP) is very important. Therefore, to confirm the specificity at

the sequence level, the RT-PCR products of BFP and the restored GFP genes were subjected to PCR-RFLP analysis using *BtgI* restriction enzyme. *BtgI* enzymes can cut CCACGG site in BFP genes, but not the restored GFP genes. Therefore, the total length of 324 bp in BFP was cleaved into two fragments of 201 bp and the 123 bp. While it remained uncleaved (324 bp) in the restored GFP genes. The 6% PAGE, polyacrylamide gel electrophoresis for the BFP (control) and the restored samples is shown in figure 5.4 b.

The BFP stabled HEK293 cells were transfected with MS2-PPR56DYW construct and the MS2-gRNA. After 48 h of treatment, wildtype GFP proteins gives the green color fluorescence under fluorescent microscopy. The MS2-DYW transfected cells were observed for fluorescence under the Juli light fluorescent microscope. All three MS2-PPR56DYW construct derived for *P.patens*, *Arabidopsis* (OTP86DYW) and, *hornworts* expressed the green fluorescent after 48 h of transfection, indicating the C to U editing from BFP to GFP genes (figure 5.4 c). The Percentage of C to U RNA editing was calculated for all the three designed system based on peak area. The chromatograms representing the C to U RNA editing using DYW domains from three different plant sources, such as moss, Arabidopsis and the hornworts are shown in figure 5.5 a. We observed the percentage of RNA editing efficiency in hornworts DYW was significantly lower as compared to the others.

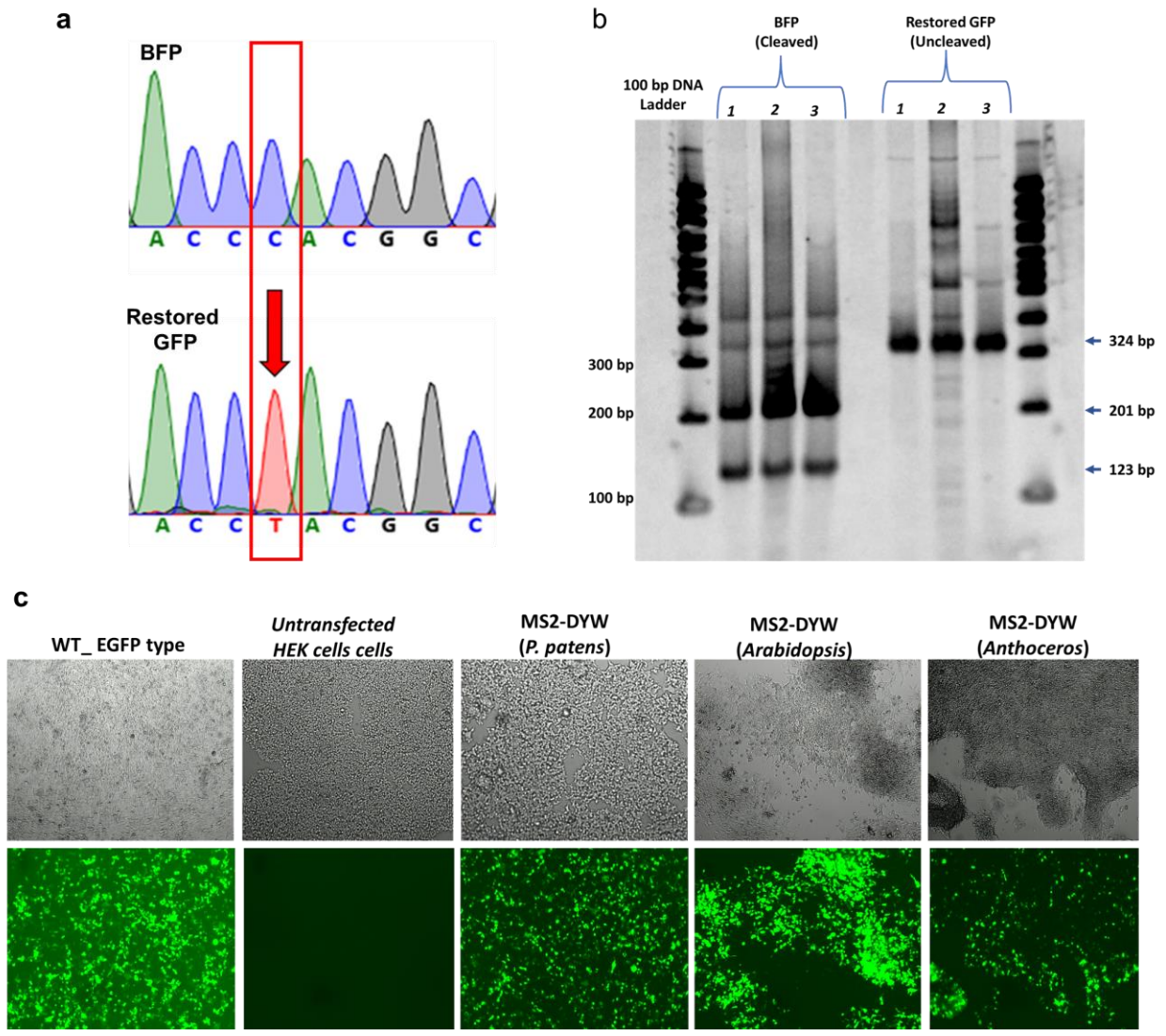


Figure 5.4. a. Sanger sequencing results showing the conversion of CCC into CCT. **b.** PCR–RFLP of cDNA extracted from transfected cells (HEK 293 stably expressing BFP), restriction-digested with *Btg*I. BFP (1; *P.patens*, 2; *Arabidopsis*, 3; *Anthoceros*) was cleaved into fragments of 201 and 123 bp, whereas restored GFP (1; *P.patens*, 2; *Arabidopsis*, 3; *Anthoceros*) was not cleaved and remained at 324 bp. **c.** Stably BFP expressing HEK 293 cells were transfected with wild type of GFP, non-transfected (control) and transfected (PPR56/DYW and gRNA). Green fluorescence expression was observed only in transfected cells, implying that catalytic domains were necessary for C-to-U editing. Imaging was performed by Juli light fluorescent microscopy.

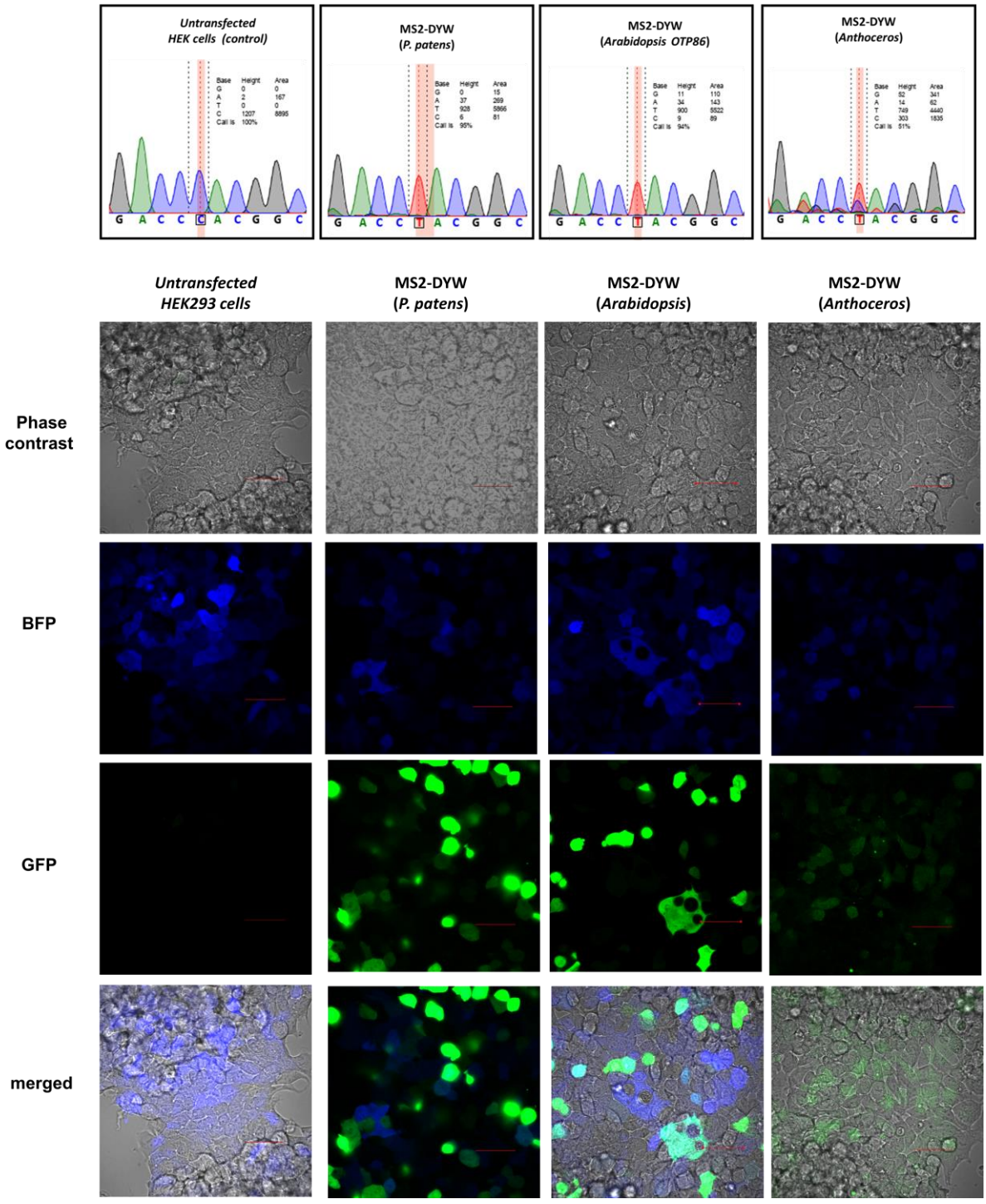


Figure 5.5. Stably BFP expressing HEK 293 cells were transfected with plasmids (DYW + guide RNA). At 48h of transfection, cells were observed for fluorescence. Cells were observed at phase contrast, Blue fluorescence protein (BFP) expression, Green fluorescence protein (GFP)

expression, and the merged. GFP expression was observed only when two factors were present, implying that both factors were necessary for C-to-U editing. Imaging was performed by LSM confocal microscopy.

5.3.4. Comparison of MS2_DYW with MS2_APOBEC1

While comparing the editing efficiencies for the two systems, we identified that the GFP can was restored up to 100% using MS2_DYW. However, about 21% restoration was achieved by MS2_APOBEC1(figure 5.6 a). In the restored samples, about 4.3 % of all SNVs show a C-to-U alterations in MS2_DYW system, which was earlier reported as 6.7% (figure 5.6 b). In the box plots, the median value of the restored C-to-U alterations is approximately 1 compared to the efficiency of editing-negative (figure 5.6 c). These results indicate a sufficiently low off-target effect. However, at more than 10 coverage of comparable C-to-U, the jitter plots show that there are hundreds of specific off-target sites (figure 5.6 d). Of these, about 4 % of C-to-U change occurred at 479 sites. Overall, we can predict from the above-mentioned analysis of the RNA-seq result is that the MS2_DYW system along with sgRNA is quite specific, comparatively providing the maximum editing efficiency and the lesser off-target effects.

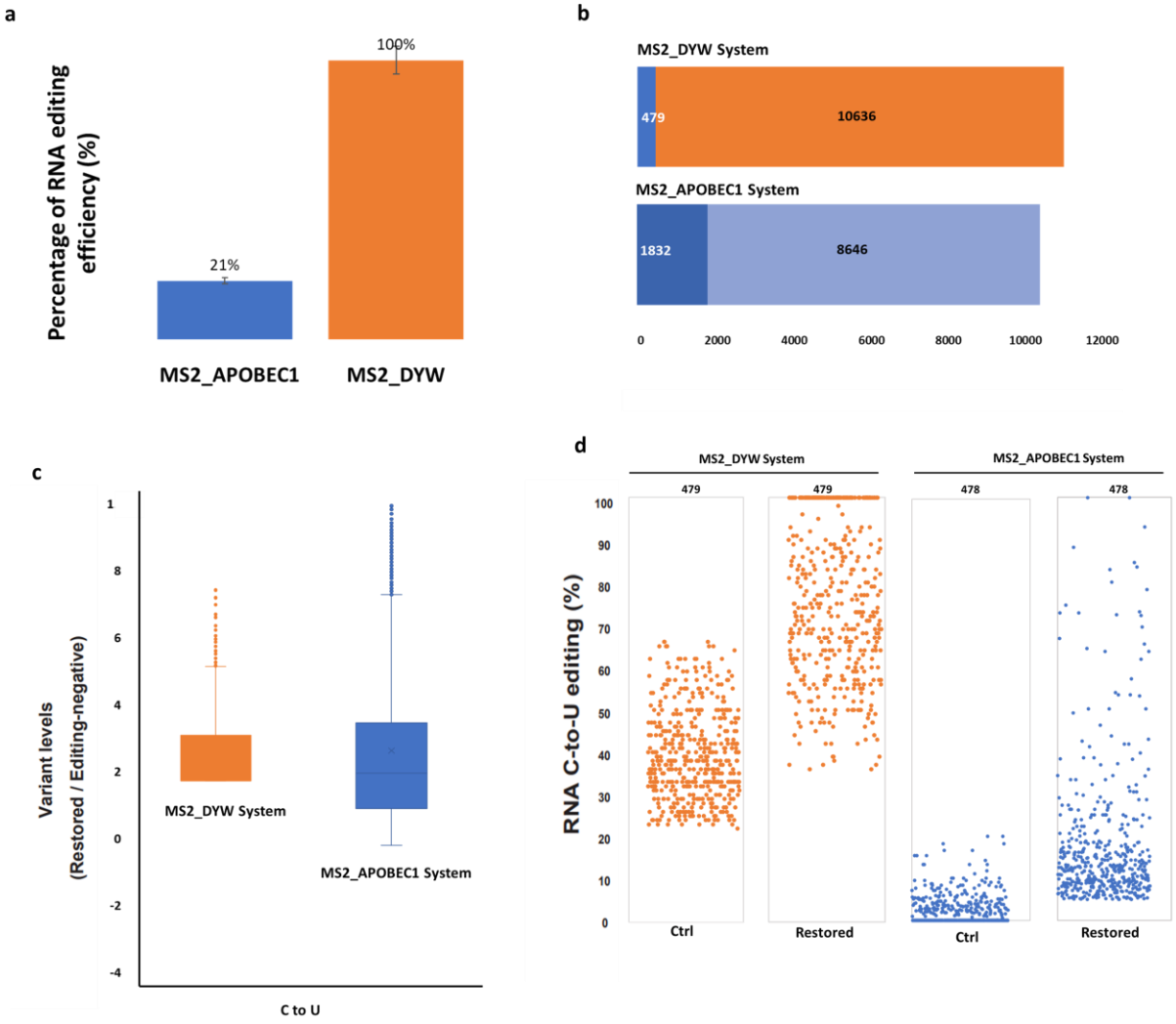


Figure 5.6. MS2-DYW system induces some off-target C-to-U RNA editing in HEK293 cells. **a.** comparison of percentage of RNA editing using MS2_DYW system and MS2_APOBEC1 system. **b.** Percentages of expressed genes with at least one edited cytosine (C-to-U) in total SNVs. **c.** Box plots showing rate of cytosines edited by MS2-DYW compared to editing-negative control. X mark is median. **d.** Jitter plots showing efficiencies of C-to-U edits (y-axis) identified from RNA-seq experiments in HEK293 cells modified by MS2-DYW or editing-negative control. n, total number of modified cytosines identified.

5.3.5. Analyzing the RNA editing efficiency by deletion primers; P2, L2, S2, E1, E2

Further, we analyzed the RNA editing by the deletion mutation for PPR56 protein. We designed the primers for different length of nucleotide fragments at the positions such as P2, L2, S2, E1 and E2 domains. MS2 construct for only DYW (Figure 5.7) was identified with no C-to-U RNA editing for EGFP. However, MS2 constructs incorporated with PPR56 showed the successful conversion of cytidine into uridine at the targeted site. Deletion mutations for the primers designed at the position P2, L2, S2, E and E2 were identified with the C-to-U editing. while only DYW domain was unable to caused editing in BFP producing HEK293 cells (Figure 5.7). It shows that the extended DYW domains up to E1E2 motif are having the deamination enzymes that make it sufficient for C-to-U RNA editing.

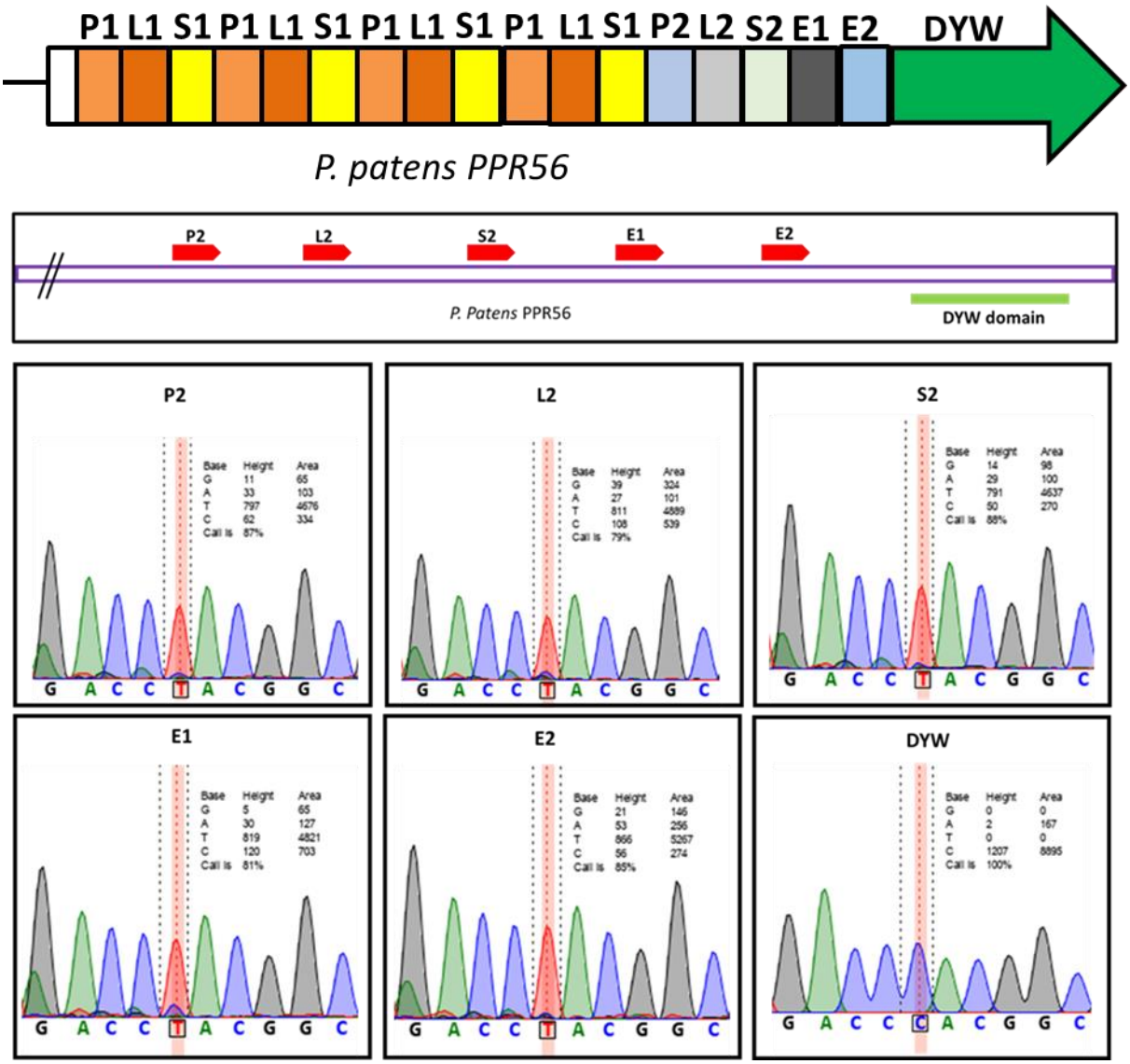


Figure 5.7. Analysis of the RNA editing efficiency by deletion primers.

5.4. Conclusion

We have developed an artificial RNA editing mechanism by combining the deaminase domain of plants DYW with a guide RNA (gRNA) which is complementary to target mRNA. In this artificial enzyme system, gRNA is bound to MS2 stem-loop, and plants specific deaminase domain of DYW, which can convert mutated target nucleotide C-to-U, is fused to MS2 coat protein. As a target RNA, we used RNA encoding blue fluorescent protein (BFP) which was derived from the gene encoding GFP by T>C mutation. Earlier MS2 system has been used with APOBEC1, showed 21% of GFP restoration. Upon transient expression of both components (DYW and gRNA), we confirmed the restoration of original sequence of mutated GFP revealing an editing efficiency of up to 85-100%. In addition, the off-target editing was also identified as least to make this system suitable for medical application. We successfully developed a bio-engineered RNA editing system using deaminase (PPR56DYW) in combination with MS2 system for C-to-U RNA editing in HEK293 cells.

5.5. References

1. Heidenreich M, Zhang F. Applications of CRISPR–Cas systems in neuroscience. *Nature Reviews Neuroscience*. 2016;17(1). doi:10.1038/nrn.2015.2
2. Kim YG, Cha J, Chandrasegaran S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences*. 1996;93(3). doi:10.1073/pnas.93.3.1156
3. Porteus MH, Carroll D. Gene targeting using zinc finger nucleases. *Nature Biotechnology*. 2005;23(8). doi:10.1038/nbt1125
4. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature Biotechnology*. 2011;29(2). doi:10.1038/nbt.1775
5. Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*. 2014;157(6). doi:10.1016/j.cell.2014.05.010
6. Gaj T, Sirk SJ, Shui S lan, Liu J. Genome-Editing Technologies: Principles and Applications. *Cold Spring Harbor Perspectives in Biology*. 2016;8(12). doi:10.1101/cshperspect.a023754
7. Marx V. Base editing a CRISPR way. *Nature Methods*. 2018;15(10). doi:10.1038/s41592-018-0146-4
8. Zhang XH, Tee LY, Wang XG, Huang QS, Yang SH. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy - Nucleic Acids*. 2015;4. doi:10.1038/mtna.2015.37
9. Gott JM, Emeson RB. FUNCTIONS AND MECHANISMS OF RNA EDITING. *Annual Review of Genetics*. 2000;34(1). doi:10.1146/annurev.genet.34.1.499

10. Keegan LP, Gallo A, O'Connell MA. The many roles of an RNA editor. *Nature Reviews Genetics*. 2001;2(11). doi:10.1038/35098584

11. Lane DA, Kunz G, Olds RJ, Thein SL. Molecular genetics of antithrombin deficiency. *Blood Reviews*. 1996;10(2). doi:10.1016/S0268-960X(96)90034-X

12. Olds RJ, Lane DA, Ireland H, et al. A common point mutation producing type 1a antithrombin iii deficiency: AT129 CGA to TGA (ARG to STOP). *Thrombosis Research*. 1991;64(5). doi:10.1016/S0049-3848(05)80011-8

13. Maas S, Rich A, Nishikura K. A-to-I RNA Editing: Recent News and Residual Mysteries. *Journal of Biological Chemistry*. 2003;278(3). doi:10.1074/jbc.R200025200

14. Katrekar D, Chen G, Meluzzi D, et al. In vivo RNA editing of point mutations via RNA-guided adenosine deaminases. *Nature Methods*. 2019;16(3). doi:10.1038/s41592-019-0323-0

15. Stafforst T, Schneider MF. An RNA-Deaminase Conjugate Selectively Repairs Point Mutations. *Angewandte Chemie International Edition*. 2012;51(44). doi:10.1002/anie.201206489

16. Vogel P, Hanswillemenke A, Stafforst T. Switching Protein Localization by Site-Directed RNA Editing under Control of Light. *ACS Synthetic Biology*. 2017;6(9). doi:10.1021/acssynbio.7b00113

17. Vogel P, Stafforst T. Site-Directed RNA Editing with Antagomir Deaminases - A Tool to Study Protein and RNA Function. *ChemMedChem*. 2014;9(9). doi:10.1002/cmdc.201402139

18. Azad MTA, Bhakta S, Tsukahara T. Site-directed RNA editing by adenosine deaminase acting on RNA for correction of the genetic code in gene therapy. *Gene Therapy*. 2017;24(12). doi:10.1038/gt.2017.90

19. Bhakta S, Azad MTA, Tsukahara T. Genetic code restoration by artificial RNA editing of Ochre stop codon with ADAR1 deaminase. *Protein Engineering, Design and Selection*. 2018;31(12). doi:10.1093/protein/gzz005

20. Bhakta S, Sakari M, Tsukahara T. RNA editing of BFP, a point mutant of GFP, using artificial APOBEC1 deaminase to restore the genetic code. *Scientific Reports*. 2020;10(1). doi:10.1038/s41598-020-74374-5

21. Lei L, Chen H, Xue W, et al. APOBEC3 induces mutations during repair of CRISPR–Cas9-generated DNA breaks. *Nature Structural & Molecular Biology*. 2018;25(1). doi:10.1038/s41594-017-0004-6

22. Blanc V, Xie Y, Kennedy S, et al. Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. *RNA*. 2019;25(1). doi:10.1261/rna.068395.118

23. Grünewald J, Zhou R, Garcia SP, et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature*. 2019;569(7756). doi:10.1038/s41586-019-1161-z

24. CHESTER A. Optimization of apolipoprotein B mRNA editing by APOBEC1 apoenzyme and the role of its auxiliary factor, ACF. *RNA*. 2004;10(9). doi:10.1261/rna.7490704

25. Hanswillemenke A, Kuzdere T, Vogel P, Jékely G, Stafforst T. Site-Directed RNA Editing in Vivo Can Be Triggered by the Light-Driven Assembly of an Artificial Riboprotein. *Journal of the American Chemical Society*. 2015;137(50). doi:10.1021/jacs.5b10216

26. Johansson HE, Liljas L, Uhlenbeck OC. RNA Recognition by the MS2 Phage Coat Protein. *Seminars in Virology*. 1997;8(3). doi:10.1006/smvy.1997.0120

27. Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM. Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell*. 1998;2(4). doi:10.1016/S1097-2765(00)80143-4

28. Bhakta S, Sakari M, Tsukahara T. RNA editing of BFP, a point mutant of GFP, using artificial APOBEC1 deaminase to restore the genetic code. *Scientific Reports*. 2020;10(1). doi:10.1038/s41598-020-74374-5

29. Doniwa Y, Ueda M, Ueta M, Wada A, Kadowaki K ichi, Tsutsumi N. The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. *Gene*. 2010;454(1-2). doi:10.1016/j.gene.2010.01.008

30. Schallenberg-Rüdinger M, Lenz H, Polsakiewicz M, Gott JM, Knoop V. A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. *RNA Biology*. 2013;10(9). doi:10.4161/rna.25755

31. Andrés-Colás N, Zhu Q, Takenaka M, de Rybel B, Weijers D, van der Straeten D. Multiple PPR protein interactions are involved in the RNA editing system in Arabidopsis mitochondria and plastids. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(33):8883-8888. doi:10.1073/pnas.1705815114

32. Rovira AG, Smith AG. PPR proteins – orchestrators of organelle RNA metabolism. *Physiologia Plantarum*. 2019;166(1):451-459. doi:10.1111/ppl.12950

33. Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A. RNA Editing in Plants and Its Evolution. *Annual Review of Genetics*. 2013;47(1). doi:10.1146/annurev-genet-111212-133519

34. Ohtani S, Ichinose M, Tasaki E, Aoki Y, Komura Y, Sugita M. Targeted Gene Disruption Identifies Three PPR-DYW Proteins Involved in RNA Editing for Five Editing Sites of the Moss Mitochondrial Transcripts. *Plant and Cell Physiology*. 2010;51(11). doi:10.1093/pcp/pcq142

35. Barkan A, Small I. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology*. 2014;65(1). doi:10.1146/annurev-arplant-050213-040159

36. Oldenkott B, Yang Y, Lesch E, Knoop V, Schallenberg-Rüdinger M. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in Escherichia coli. *Communications Biology*. 2019;2(1). doi:10.1038/s42003-019-0328-3

Chapter 6

A non-sequencing approach for the rapid detection of RNA editing

6.1. Introduction

Identification of single-nucleotide variants (SNVs) in the genomic RNA, including A-to-I, C-to-U or U-to-C, could be used to find RNA-editing events. Detection of SNV in the RNA remains a challenging task and it needs to be addressed technically. Conventionally, the ratio of edited-to-non-edited RNA is determined by direct sequencing, allele-specific real-time PCR, or denaturing HPLC approaches [1–6]. However, the low accuracies due to a higher level of noise and less time/cost effectiveness of these approaches pose technological bottlenecks for RNA-based SNV detection [7,8]. In this stream, we engaged to develop a new protocol using the principles of temperature gradient gel electrophoresis (TGGE) for identifying single nucleotide polymorphisms (SNPs) as an alternative to eliminate the need for direct RNA sequencing approach in RNA-editing assay.

Electrophoresis is recognized as one of the most preferred method for separation and analysis of biomolecules in life science laboratories. In TGGE, DNA fragments of the same size but with different sequences can be separated which is based on decreased electrophoretic mobility of a partially melted double-stranded DNA molecule in porous gels containing a linear temperature gradient [9,10]. The melting of DNA fragments proceeds in discrete melting profile. Once a domain with the lowest melting temperature reaches it melting temperature at a particular position in the gel, a transition of a helical to a partially melted molecule occurs, and migration of the molecule will practically halt. Therefore, TGGE which utilizes both mobility (size information) and temperature-induced structural transition of DNA fragments (sequence-dependent information) makes this approach highly resolvable and powerful. The featuring points in a melting pattern which correspond to those where structural transitions of DNA occurs from double-stranded to single-stranded DNA are assigned and categorized in 3 types including strand initial-dissociation point, strand mid-dissociation point, and strand end-dissociation point (Figure 6.1). Sequence

variation, even a single base difference, within such domains causes the melting temperatures to differ, and molecules with different sequences will show discrete melting (denaturation) pattern during the electrophoresis. Therefore, TGGE can be used for analyzing single-nucleotide variants in the genomic RNA and thus can be invaluable as a high-throughput RNA-editing detection method. However, this throughput gains can be lost when traditional gel electrophoresis based TGGE is used. Using a miniaturized version of TGGE, micro-TGGE (μTGGE), the gel electrophoretic time was shortened, and the analysis was accelerated with 100-fold productivity [11]. The simplicity and compactness was further improved by introducing a viable solution for field-applicable, hand-held and affordable gel electrophoresis system from BioSeeds Corporation, named PalmPAGE [12]. In this work, we introduced a new protocol using μTGGE (hardware) and uMelt (software) [13] to examine 4 types of RNA editing sites in mammalian HEK293 cells and Arabidopsis tissues and rapidly detect the RNA-editing types including A-to-I, C-to-U or U-to-C (Figure 6.1A). With an average run-time of 15-30 min, this protocol enables rapid, reliable and easy identification of RNA-editing without the need of direct RNA sequencing.

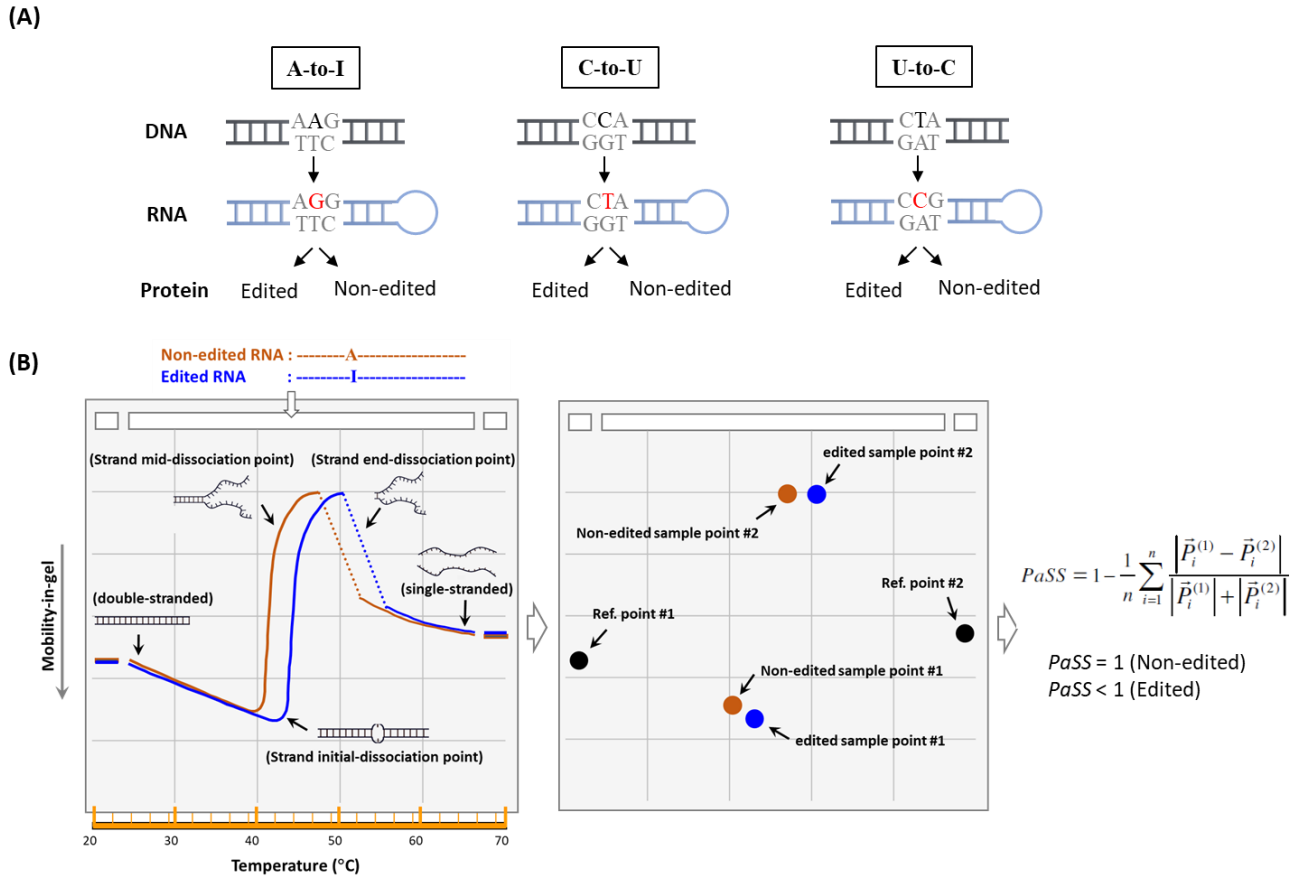


Figure 6.1: The procedure used to identify RNA editing by TGGE. (A) Types of RNA editing events. **(B)** The schematic representation of TGGE pattern for edited and non-edited gene. In TGGE, sample is layered on the top of a slab gel migrate downward with drawing a characteristic curvature caused by the temperature gradient. The featuring points in melting pattern are assigned and processed to calculate PaSS (pattern similarity score) value. The PaSS calculation is performed as described in Equation 1.

6.2. Methods and Materials

1. Optimization of the target fragment

In order to predict the gene fragment regions with significant difference in melting profiles between edited or non-edited regions, the melting curve of gene fragments was predicted by the uMelt HETS web-based tool[13]. uMelt HETS predicts the shape of melting curves for heteroduplex and homoduplex products. Three gene fragments of near 300 bp in length were designed by keeping the location of editing or non-editing sites either at 5'-terminal end or middle position or 3'-terminal end of gene fragment. The fragment which shows the maximum shift of the melting region on the helicity axis between edited and non-edited sites was selected for further analysis. For uTGGE analysis, the selected gene fragment was synthesized by PCR amplification. Both the forward and reverse primers were designed using software DNADyanamo https://www.bluetractorsoftware.com and verified using the NCBI Primer-BLAST tool. Primers were purchased from Eurofins (Japan) diluted in TE buffer at the concentration of 100 pmol/uL in a salt-free condition. Each primer set was diluted to the concentration of 10 pmol/uL using distilled water.

2. RNA extraction and PCR amplification of target fragment

We extracted the total RNA from the source of edited and non-edited genes. The RNA was isolated from *Arabidopsis* tissues using the Plant Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions given in the manual and, from HEK293 cells using TRIZOL extraction method [14]. The synthesis of the corresponding cDNA was performed using ReverTra Ace enzyme (200U/ µl) at 55°C for 60 minutes for RT-PCR. We confirmed the concentration by NanoDrop-1000 spectrophotometer and keep at -25°C. PCR

Amplification assays include 2min initial denaturation at 95°C followed by 35 cycles each with 2min denaturation at 95°C, 30s the annealing temperature depending upon the set of primers, 2min synthesis at 72°C, and a final step of synthesis for 7min at 72°C. PCR products were purified by adding 2U ExoI (TAKARA) and 0.5U Shrimp Alkaline Phosphatase (TAKARA) and incubate in PCR machine at 37°C for 1h followed by 15min at 80°C and, then hold at 4°C. In a new PCR tube, add 11.5μL H20, 2.5 μL of forward primer, and 1 μL of PCR products (with ExoSAP). Sequenced by eurofins genomics. https://eurofinsgenomics.jp/jp/home/. Each sequencing results were validated with both the forward and the reverse primers. The lists for pair of primers for the studied genes are given in Table1. Diluted the purified PCR products up to 200ng with de-ionized water. Mixed with 6X gel loading dye into a 250 μL tube and were subjected to electrophoresis on a 6% of 8M polyacrylamide gel.

3. μTGGE analysis

3.1. Assembly of the gel cassettes

A new design of gel cassette was used for μTGGE analysis. This consists of a set of three 1-inch sized gel plates: bottom gel plate, top gel plate and lane-former plate (see Figure 6.2). The top gel plate are sandwiched between other two plates and assembled in the gel casting holder for gel polymerization.

3.2. **Polyacrylamide Gel preparation**

To prepare 6% of 8 M polyacrylamide, weigh 7.2 g of urea and dissolve into 10 mL tube. Heat up in microwave for 20-30s and bring to the room temperature. Added 3 mL of 5X TBE buffer, 2.25 mL of 40(w/v)%-acrylamide/bis (19:1), 75 μl of 10X ammonium persulfate (APS) and 15 μl of

tetramethylethylenediamine (TEMED) into the solution. Immediately, poured the gel solution slowly into the gel casting holder, avoiding the air bubbles. Soaked the buffer pads (0.5 × 2.5 cm) in 2 mL of 1× TBE buffer and use as a source of running buffer.

3.3. Generation of melting profiles of cDNA using micro temperature gradient gel electrophoresis (μTGGE) unit.

Place the gel cassette in the horizontal electrophoresis chamber unit followed by placing the upper and lower buffer pads as shown in Figure 2d. Now, load 10 μl of PCR product in the middle longer well and 1 μl each in both the side wells. After a one min wait, connected the power unit and supply 100 V for 12 min duration at a linear temperature gradient from 25 to 65°C. After the run has been complete, take out the cassette and remove the upper glass cover. Poured a 300 μL volume of 10× SYBR gold nucleic acid gel stain on top of the gel, and the melting profiles were visualized using the blue LED flashlight installed in the palm-sized electrophoretic device. A soft file of image was saved for PaSS calculation. Repeat every electrophoretic experiment thrice to confirm the reproducibility of data.

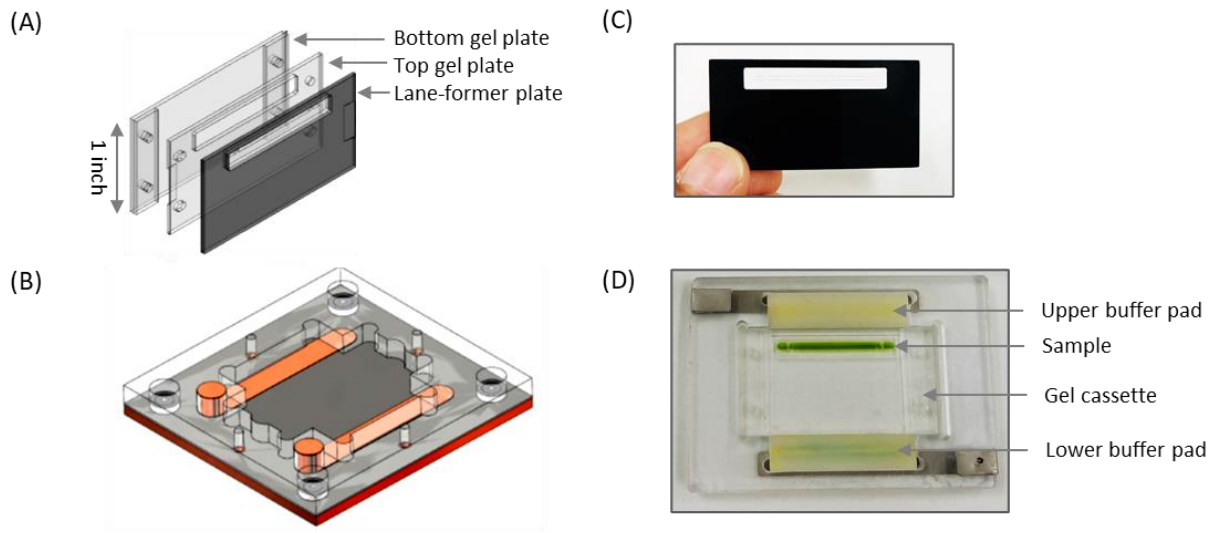


Figure 6.2: Steps involved in Temperature gradient gel Electrophoresis. **(A)** Assembly of 3 tiny sized glass gel plates, **(B)** glass cassette for the TGGE Run, **(C)** 1-inch glass plates hold into figures**, (D)** cassette with loaded samples,

4. *PaSS* calculations

Download the TGGE analyzer software and follow the below steps:

4.1.Open the "micro TGGE"

4.2.Open the JPEG image file. If you don't have JPEG image file, convert the file to a JPEG

4.3.Push the "フレーム" (Frame) and fix the area (frame) of the image

4.4.Push the " 座標補正" (Coordinate correction) and add two references points.

4.5.Push the "特徴点追加" (addition of feature points) and add sample point and save the data

4.6.Push the "サンプル" (Sample) and then push the "類似サンプルを検索" (search for simple

points) to compare two or more image.

6.3. Results and discussion

6.3.1. TGGE can be used to identify a single nucleotide base change in RNA editing events

TGGE-based melting profiles were analyzed for different types of RNA editing samples including the C-to-U, A-to-I(G) and the U-to-C types. We selected four edited genes: C-to-U RNA editing in blue fluorescent protein gene produced in HEK293 cells and edited by deaminase enzymes of APOBEC1 (Bhakta et al., 2020); A-to-I RNA editing in EGFP genes containing ochre stop codon (TAA) produced in HEK293 cells and edited by adenosine deaminase Acting on RNA (ADAR1)[16], and U-to-C RNA editing in two nuclear genes, AT2G16586 and AT5G02670, recently identified in *Arabidopsis thaliana* [17,18]. Single base nucleotide differences among the RNA edited sample and corresponding wild type non-edited sample were confirmed by sequencing method followed by TGGE analysis to identify the difference in melting curves (Figure 6.3). For C-to-U RNA editing type, the non-edited sample with the original C base showed a longer melting pattern at strand end-melting point than the edited sample with the modified U base, which exhibits a shorter melting pattern (Figure 6.3A). Similar to this phenomenon, the edited sample with modified I(G) base displayed a longer melting pattern at strand end-melting point than the non-edited sample with original A base in A-to-I(G) RNA editing type (Figure 6.3B). For the "reverse" U-to-C RNA editing type, the edited gene with modified C base showed a longer melting pattern between strand initial-melting and strand end-melting points. In this type, the edited C base exhibits an additional diffused melting pattern near the strand initial-melting point in gene AT2G16586 (Figure 6.3C), while it can be clearly visualized at end-melting point in gene AT5G02670 (Figure 6.3D). The lists of genes selected for uTGGE profiles with the RNA editing sites are given in Table 6.1.

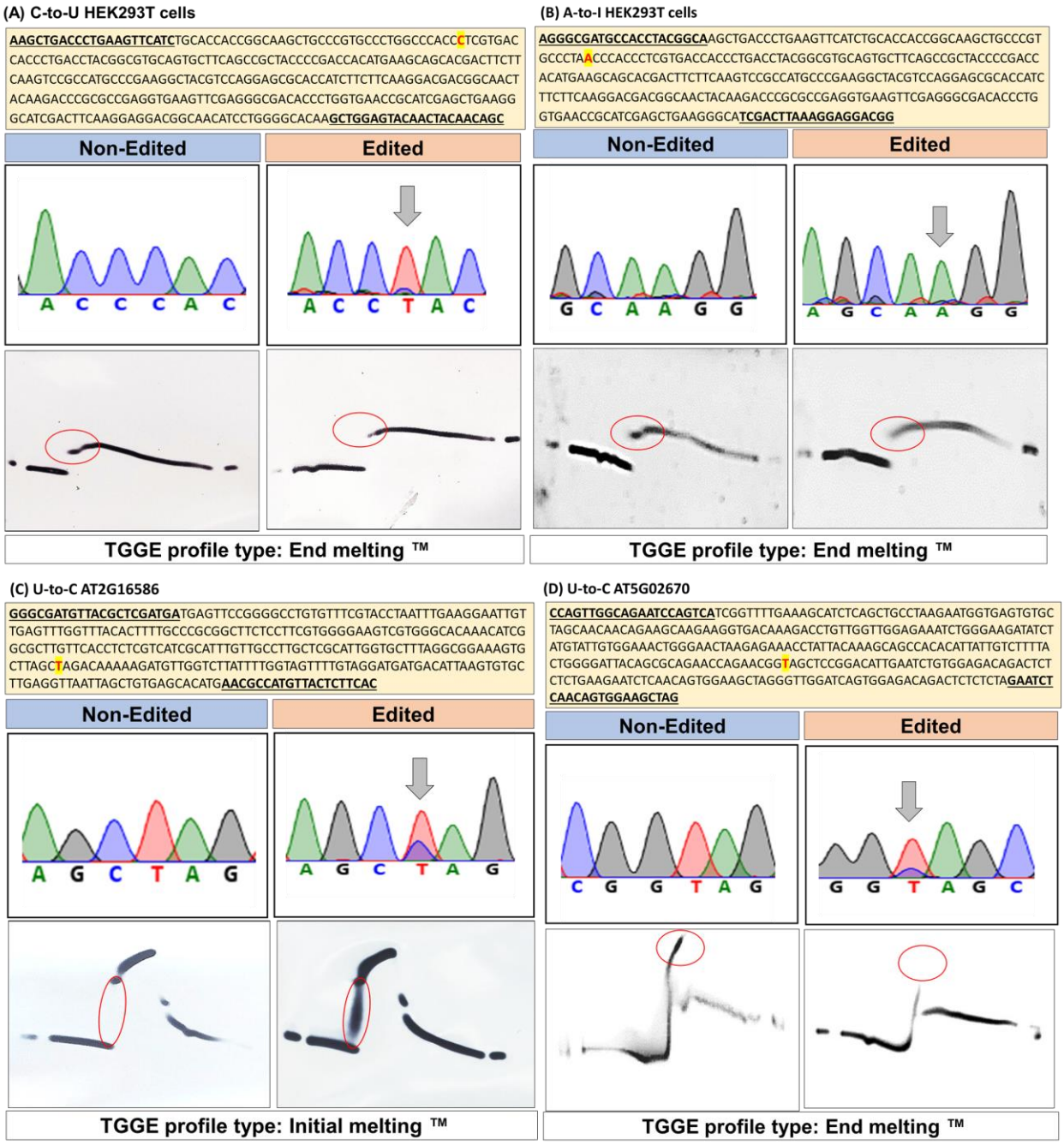


Figure 3: Temperature-Gradient Gel Electrophoresis analysis for single base nucleotide change in various RNA editing. Melting profiles for C-to-U, A-to-I(G) and U-to-C RNA editing was compared for wild type and edited genes. The differences in the band patterns are indicated by red arow.

Table 1: Lists of genes studied for melting Profiles.

S.No	RNA editing	Source	Position	Gene Id	Forward Primer	Reverse primer	Sequence length
1.	C-to-U	HEK293T cells	48th	EGFP	AAGCTGACCCT GAAGTTCATC	GCTGTTGTAGT TGTACTCCAGC	324
2.	A-to-I	HEK293T cells	59th	EGFP	AGGGCGATGCC ACCTACGGCA	CCGTCCTCCTT TAAGTCGA	300
3.	U-to-C	*Arabidopsis*	152th	AT2G16586	GGGCGATGTTA CGCTCGATGA	GTGAAGAGTAA CATGGCGTT	301
4.	U-to-C	*Arabidopsis*	169th	AT5G02670	CCAGTTGGCAG AATCCAGTCA	CTAGCTTCCAC TGTTGAGATTC	300

6.3.2. Quantitative analysis of TGGE melting pattern in RNA editing events

To evaluate the reproducibility in TGGE-based melting profiles of RNA editing events, the information obtained in melting patterns of TGGE was quantitatively processed by calculating Pattern Similarity Scores (*PaSS*)[11]. As described in Figure 6.1B, the featuring points in melting pattern that correspond to structural transitions between double-stranded to single-stranded DNA were used. In order to eliminate experimental variables, computer-aided normalization was performed using two internal reference points, i.e., reference point #1 for band position of sample in double-stranded form (the most left lane) and reference point #2 for band position of sample in single-stranded form (the most right lane). The featuring points of melting pattern are obtained by normalizing the coordinates of the featuring points with both the internal reference points and used

to calculate the *PaSS*. The *PaSS* value provides a measure how two set of melting pattern can be closely superposed, generating a higher value (maximum: 1) for highly similar melting patterns. Thus, the *PaSS* value between melting patterns of non-edited and edited is assumed to be less than 1. As shown in Figure 6.4, the differences between the *PaSS* value of non-edited and edited samples were measured higher in case of C-to-U and A-to-I types of RNA editing, in comparison with U-to-C type of RNA editing. Each experiment was repeated thrice, and the average value was used for the analysis. We observed that the higher of *PaSS* value differences in C-to-U type RNA editing or A-to-I type RNA editing can be represented to the location of editing sites near the 5'-terminal ends (i.e., at 48th and 59th position of about 300 bp long fragment) than the lower of *PaSS* value differences in U-to-C type of RNA editing which is rather located in the middle of 5' and 3'-terminal ends (i.e., at 152th and 169th of about 300 bp long fragment). It indicates that the edited sites located at the terminal positions were more intend to represent the greater difference between the edited and non-edited sites.

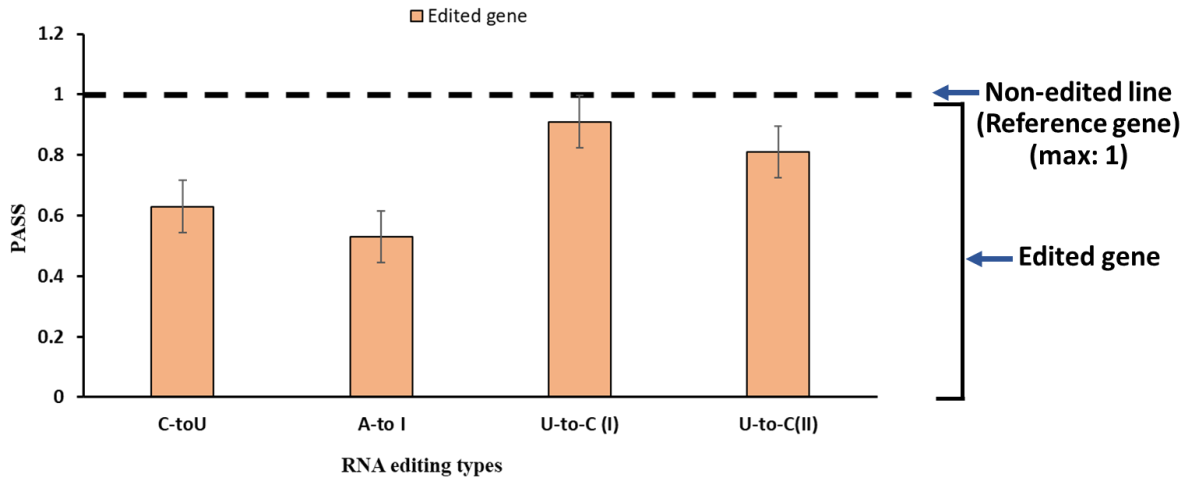


Figure 6.4: Bar graph showing the average PASS values for the edited genes.

6.3.3. Optimization of TGGE melting pattern to differentiate RNA editing events

During *PaSS* value analysis, we observed that the difference in the values were varied with the difference in the positions of RNA editing sites. Therefore, we further investigated the melting pattern for a defined editing site in C-to-U type RNA editing gene by shifting the position of editing site in three ways, i.e., at 5'-terminal end, 3'-terminal end and center of 5' and 3'-terminal ends. The fragment length was kept constant as 300 bp (Figure 6.5A). Prior to TGGE analysis, the melting patterns of non-edited fragment and all three set of edited fragments were theoretically predicted using uMelt HETS web-based tool. As shown in Figure 6.5B, the modification of C base to U base shifts the melting curve earlier along the temperature axis. After changing the position of editing sites, the order of earlier melting curve shift is: 5'-terminal end > 3'-terminal end > center of 5' and 3'-terminal ends. Very interestingly, *PaSS* value obtained by TGGE melting patterns was consistent with the predicted results (Figure 6.5C). The nucleotide base differences located at the 5' or 3'-terminal ends displayed a higher variation in *PaSS* value between edited and non-edited genes. These results suggested that a prior knowledge of difference in melting curves between edited and non-edited gene can guide to optimize and determine the analyzing gene fragment for RNA editing.

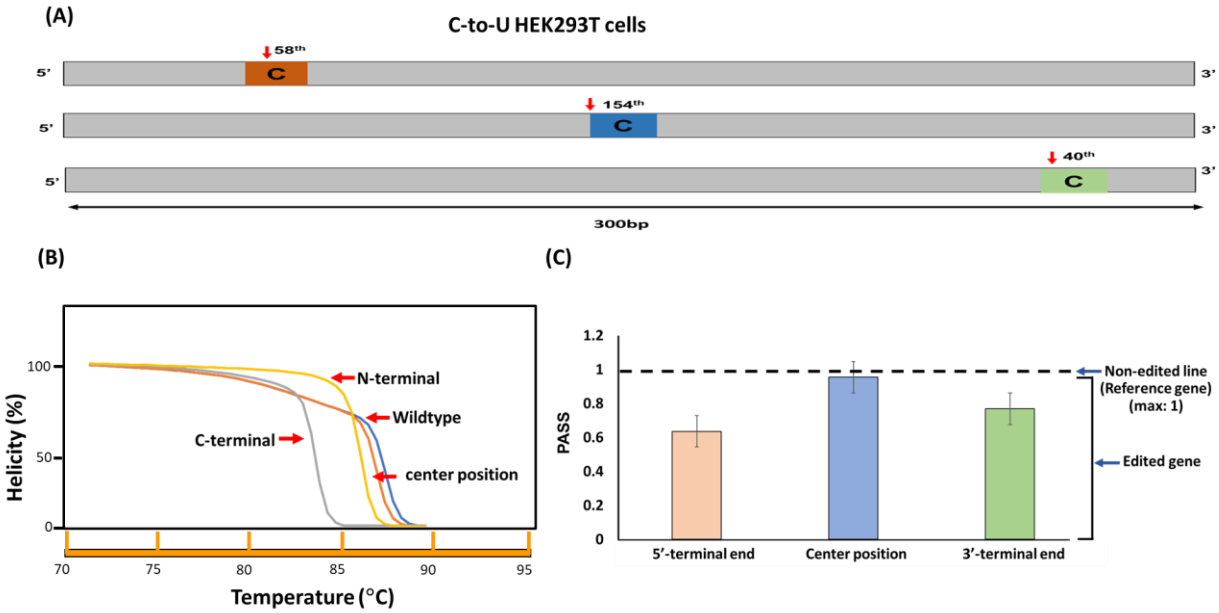


Figure 6.5: Position-specific PASS analysis

6.4. Conclusion

We have developed an RNA editing detecting tool using a portable micro-Temperature gradient gel electrophoresis technique. This tool can be used to confirm the edited genes prior to the Sanger sequencing. Being just 1-inch in size, it required very less amount of samples and gives a fast detection for the differences between the sequences even with a single nucleotide change. It provides a robust, cost-effective, and less time-consuming method for RNA editing analysis. Protocols are very easy to follow for any experts or the new in the field. Here, we showed that the edited and non-edited genes have the difference in their band patterns for single base nucleotide modification due to change in melting points. We also validated the similarity for the uTGGE curves to the uMELT curves for the same gene fragment to the specific RNA editing. we further confirmed that the nucleotide base differences located at the 5’ and 3’ terminals are having the higher difference with their reference genes, as compared the center position. Concerning the limitations, this method cannot be used to confirm the type of nucleotide modifications. Sanger sequencing is required further to confirm the specific nucleotide changes.

6.5. REFERENCES:

1. Chateigner-Boutin AL, Small I. A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. *Nucleic Acids Research*. 2007;35(17). doi:10.1093/nar/gkm640
2. Chen YC, Kao SC, Chou HC, Lin WH, Wong FH, Chow WY. A real-time PCR method for the quantitative analysis of RNA editing at specific sites. *Analytical Biochemistry*. 2008;375(1). doi:10.1016/j.ab.2007.12.037
3. Barbon A. Glutamate receptor RNA editing: a molecular analysis of GluR2, GluR5 and GluR6 in human brain tissues and in NT2 cells following in vitro neural differentiation. *Molecular Brain Research*. 2003;117(2). doi:10.1016/S0169-328X(03)00317-6
4. Gallo A. Micro-processing events in mRNAs identified by DHPLC analysis. *Nucleic Acids Research*. 2002;30(18). doi:10.1093/nar/gkf516
5. Schiffer HH, Heinemann SF. A Quantitative Method to Detect RNA Editing Events. *Analytical Biochemistry*. 1999;276(2). doi:10.1006/abio.1999.4369
6. Burns CM, Chu H, Rueter SM, et al. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature*. 1997;387(6630). doi:10.1038/387303a0
7. Marian AJ. The Bottleneck in Genetic Testing. *Circulation Research*. 2015;117(7). doi:10.1161/CIRCRESAHA.115.307344
8. Bird A. Genome Biology: Not Drowning but Waving. *Cell*. 2013;154(5). doi:10.1016/j.cell.2013.08.010
9. Riesner D. SG, ZR, ORA, WM, HW, VS& HK. Temperature-gradient gel electrophoresis of nucleic acids: analysis of conformational transitions, sequence variations, and protein-nucleic acid interactions, Electrophoresis 1989: 10, 377-389. *Electrophoresis*. Published online 1989.

10. Jones BM, Knapp LA. Temporal Temperature Gradient Electrophoresis for Detection of Single Nucleotide Polymorphisms. In: ; 2009. doi:10.1007/978-1-60327-411-1_9

11. Biyani M& NK. Hundred-fold productivity of genome analysis by introduction of micro-temperature gradient gel electrophoresis. *Electrophoresis* . Published online 2001.

12. Rathore H, Biyani R, Kato H, Takamura Y, Biyani M. Palm-size and one-inch gel electrophoretic device for reliable and field-applicable analysis of recombinase polymerase amplification. *Analytical Methods*. 2019;11(39). doi:10.1039/C9AY01476D

13. Dwight Z, Palais R, Wittwer CT. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics*. 2011;27(7). doi:10.1093/bioinformatics/btr065

14. Rio DC, Ares M, Hannon GJ, Nilsen TW. Purification of RNA Using TRIzol (TRI Reagent). *Cold Spring Harbor Protocols*. 2010;2010(6). doi:10.1101/pdb.prot5439

15. Bhakta S, Sakari M, Tsukahara T. RNA editing of BFP, a point mutant of GFP, using artificial APOBEC1 deaminase to restore the genetic code. *Scientific Reports*. 2020;10(1). doi:10.1038/s41598-020-74374-5

16. Azad MdTA, Qulsum U, Tsukahara T. Comparative Activity of Adenosine Deaminase Acting on RNA (ADARs) Isoforms for Correction of Genetic Code in Gene Therapy. *Current Gene Therapy*. 2019;19(1). doi:10.2174/1566523218666181114122116

17. Ruchika, Okudaira C, Sakari M, Tsukahara T. Genome-Wide Identification of U-To-C RNA Editing Events for Nuclear Genes in Arabidopsis thaliana. *Cells*. 2021;10(3). doi:10.3390/cells10030635

18. Ruchika, Tsukahara T. The U-to-C RNA editing affects the mRNA stability of nuclear genes in Arabidopsis thaliana. *Biochemical and Biophysical Research Communications*. 2021;571.

doi:10.1016/j.bbrc.2021.06.098

Chapter 7

Final conclusion and Future prospective

1.1.Final conclusion

In **Chapter 2**, I was the first one to report the rare occurrence of U-to-C RNA editing events in the nuclear genes of *Arabidopsis thaliana.* Cytosine-to-Uridine (C-to-U) RNA editing involves the deamination phenomenon, which is observed in animal nucleus and plant organelles; however, it has been considered the U-to-C is confined to the organelles of limited non-angiosperm plant species. Although previous RNA-seq based analysis implied U-to-C RNA editing events in plant nuclear genes, it has not been broadly accepted due to inadequate confirmatory analyses. Here I examined the U-to-C RNA editing in Arabidopsis tissues at different developmental stages of growth. In this study, the high-throughput RNA sequencing (RNA-seq) of 12-day-old and 20-day-old Arabidopsis seedlings was performed, which enables transcriptome-wide identification of RNA editing sites to analyze differentially expressed genes (DEGs) and nucleotide base conversions. The results showed that DEGs were expressed to higher levels in 12-day-old seedlings than in 20-day-old seedlings. Additionally, pentatricopeptide repeat (PPR) genes were also expressed at higher levels as indicated by the log2FC values. RNA-seq analysis of 12-day and 20-day-old Arabidopsis seedlings revealed candidates of U-to-C RNA editing events. Sanger sequencing of both DNA and cDNA for all candidate nucleotide conversions confirmed the seven U-to-C RNA editing sites. This work clearly demonstrated presence of U-to-C RNA editing for nuclear genes in Arabidopsis, which provides the basis to study the mechanism as well as the functions of the unique post-transcriptional modification.

Chapter 3 In my further study, I investigated the U-to-C RNA editing-related genes in Arabidopsis tissues and the effects on mRNA stability, with a special focus on PPR proteins. I previously shown the extensive occurrence of U-to-C RNA editing in 12-day and 20-dayold Arabidopsis seedlings. Here, I have demonstrated the effects of this “reverse” RNA editing on the mRNA stability for all

seven edited genes. I also identified U-to-C RNA editing in the nuclear PPR gene (AT2G19280) in 12-day-old seedlings of Arabidopsis thaliana. The U-to-C RNA editing sites were found in the untranslated region (3' UTR) of the mature mRNA and may affect its secondary structure. I also examined the correlation between U-to-C RNA editing-related genes and their mRNA abundance. Furthermore, I investigated the effects of U-to-C RNA editing in Arabidopsis using the transcription inhibitor actinomycin D (Act D). The addition of Act D to the cell suspension culture of transgenic Arabidopsis generated by Agrobacterium-mediated transformation showed that single nucleotide base conversion adversely affected the mRNA secondary structure and stability.

Chapter 4 Pentatricopeptide repeats (PPR) proteins are exclusively act as sequence-specific RNA-binding proteins within mitochondria and chloroplasts in almost all land plants. Genome-wide analysis of the hornworts, *Anthoceros agrestis*, revealed the PPR proteins in this species contain unique C-terminal DYW-like domains with specific signatures. These domains are the strongest candidates for the U to C RNA editing enzyme since such domains were not observed in other model plant species having only C to U RNA editing. In present work, I have explored the study on three different variants of C-terminal PPR proteins of hornworts, GRP-type, DRH- type and DYW-types. I investigated the RNA editing events by cloning the Hornworts PPR genes. An expression system was developed in which the Hornworts specific PPR protein variants were cloned with PPR56 (truncated DYW), *Physcomitrella patens* (moss) editing factor. The assay system allowed to study RNA editing by hornworts PPR genes with its potential target sequences in bacterial and mammalian cells. We measured the gene expression levels of DYW variants. In addition, we demonstrated the functional homology of DYW domains with APOBEC1in mammalian cells.

Chapter 5 A genetic disorder is a disease caused by a change in the nucleotide base sequences. Many genetic diseases are raised due to T-to-C point mutations. Hence, editing of mutated genes represents a promising strategy for treating these disorders. The MS2 system were initially developed with ADAR1 and APOBEC1 for A-to-I(G) and C to-U RNA editing, respectively. In this study, an expression system was designed with PPR56, *Physcomitrella patens* (moss) editing factor. The assay system allowed to study RNA editing by plant-derived PPR genes with its potential target sequences in animal cells. We engineered an artificial RNA editing mechanism by combining the deaminase domain of plant derived PPR56 with a guideRNA (gRNA) which is complementary to target mRNA. As a target RNA, we used RNA encoding blue fluorescent protein (BFP) which was derived from the gene encoding GFP by a single T-to-C substitution. Upon transient expression of both components (PPR56 and gRNA), we confirmed the restoration of sequence of GFP revealing an editing efficiency of up to 85-100%, while previous developed system with APOBEC1 only showed about 20 % editing efficiency. Furthermore, we identified that the C-terminal E2-DYW domain of PPR56 is sufficient for C-to-U conversion in the MS2 system. This high efficiency bio-engineered RNA editing system using plant C-to-U RNA editing enzyme in combination with MS2 will be a powerful tool to rescue T-to-C point mutations without altering genomic DNA.

Chapter 6 Lastly, I believe analysis of RNA editing is essentially an important process, however, largely relies on Sanger sequencing and RNA-seq technology for the detection and quantification of edited sites. Base nucleotide change validation through Sanger sequencing and the whole RNA sequencing can be both expensive and time-consuming. In this study, I introduced a non-sequencing approach for the rapid detection of RNA editing using a portable micro-Temperature gradient Gel Electrophoresis (μTGGE). This is based on the principle of

electrophoresis, which use temperature to denature the samples as it moves across the polyacrylamide gel. In this method, a fragment of double-stranded DNA when heated, forming a gradient of double-stranded DNA to partially separated strands to completely separated single-stranded DNA. A sample of RNA editing with different nucleotide bases show the different melting profiles because of their different melting profile. Here, we demonstrated the difference between the melting profiles for edited and non-edited (wild type) RNAs. Reproducibility was evaluated from measuring the pattern similarity scores (PaSS) between the band patterns obtained with the edited and non-edited RNAs. This tool is providing a simple, and cost-effective method for detecting even a single base modification in RNAs. We expect that our rapid analyzing tool will foster further discoveries in this rapidly expanding field of molecular biology.

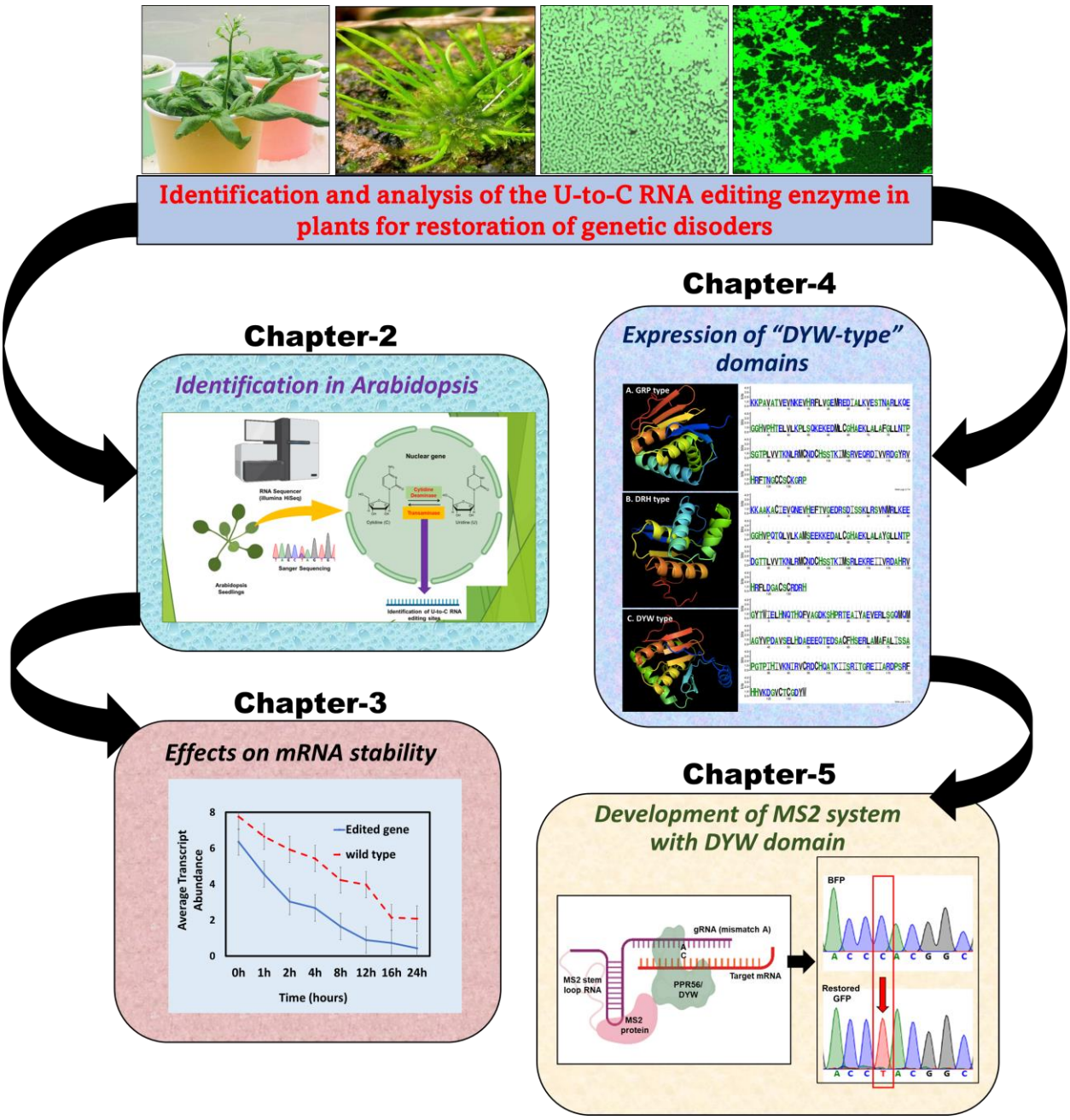


Figure 7.1. schematic representation of thesis summary.

1.2.Future Prospects

1. Investigation on possibility for **G-to-A editing.** (Ongoing project in Tsukahara lab)

- Rare occurrence of U-to-C RNA editing for nuclear genes was reported in *Arabidopsis* (Ruchika et. al. Cells, 2021). In this work, we identified the additional substitution with maximum number of Nucleotide modification for G-to-A, which possibly be a candidate genes G-to-A RNA editing.

2. Searching for **co-factors**: Creating constructs of hornwort GRP or DRH type DYW domain with protein-Tags and transform into hornwort.
3. Structural analysis: Preparation of constructs for structural analysis of GRP/DRH domains through **Protein-NMR** and **crystallography**. Searching for optimal constructs and conditions for protein expression
4. Reconstruction of U-to-C RNA editosome: Cloning candidate co-factors and co-expression of them with GRP or DRH type PPR proteins in E.coli to assay U-to-C aminase activity.
5. Transformation of **RNAi** and **CRISPR** based constructs into hornworts.
6. To development an Artificial transaminase system with hornworts specific “GRP/DRH” type PPR protein

- Upon transient expression of both components (PPR56DYW and gRNA), the restoration of original sequence of mutated GFP was confirmed, revealing an editing efficiency of up to 85-100%. Therefore, I would like to develop an artificial enzyme system for U-to-C RNA editing by the incorporating the MS2 coat proteins with the candidate genes “GRP” or “DRH”.

LISTS OFPUBLICATION

1. **Ruchika**; Okudaira, C.; Sakari, M.; Tsukahara, T. Genome-Wide Identification of U-to-C RNA Editing Events for Nuclear Genes in *Arabidopsis thaliana*. ***Cells*** 2021, *10*, 635. https://doi.org/10.3390/cells10030635
2. **Ruchika**, Tsukahara T. The U-to-C RNA editing affects the mRNA stability of nuclear genes in Arabidopsis thaliana. ***Biochem Biophys Res Commun.*** 2021 Jul 26; 571:110-117. https://doi.org/10.1016/j.bbrc.2021.06.098
3. **Ruchika**, Toshifumi Tsukahara, and Manish Biyani. A non-sequencing approach for the rapid detection of RNA editing. ***Journal of Visualized experiments, JOVE*** (Accepted)
4. **Ruchika**, Mizuki Takenaka, Toshifumi Tsukahara, Expression of Hornworts specific "DYW-type" PPR proteins for RNA editing. ***Plant molecular biology*** (To be submitted)
5. **Ruchika**, Sakari, M, Tsukahara, T, Takenaka, M. Development of MS2 system with plant derived "DYW" type PPR protein (In preparation).

INTERNATION CONFERENCE PRESENTATIONS

1. **Ruchika**, Toshifumi Tsukahara, RNA-Seq based identification of U-to-C RNA editing in *Arabidopsis* thaliana, in 43rd Annual meeting of **Molecular Biology Society of Japan MBSJ**, 02-04 December 2020, webinar ***(Poster presentation)***

2. **Ruchika,** Toshifumi Tsukahara, Effect of U-to-C RNA editing on the mRNA half-life of PPR protein, 5th World **Plant Genomics and Plant Science Congress**, Paris, France, September 27-28, 2021. ***(Featured Speaker)***

   - ***Young scientist Award***

3. **Ruchika**, Chisato Okudaira, Matomo Sakari, Toshifumi Tsukahara, Genome-wide identification and analysis of U-to-C RNA editing events in *Arabidopsis thaliana* by transcriptome sequencing, 3rd Annual congress on **Plant Biology and Biotechnology** (ACPB-2021) Osaka, Japan October 07-09, 2021. ***(Featured Speaker)***

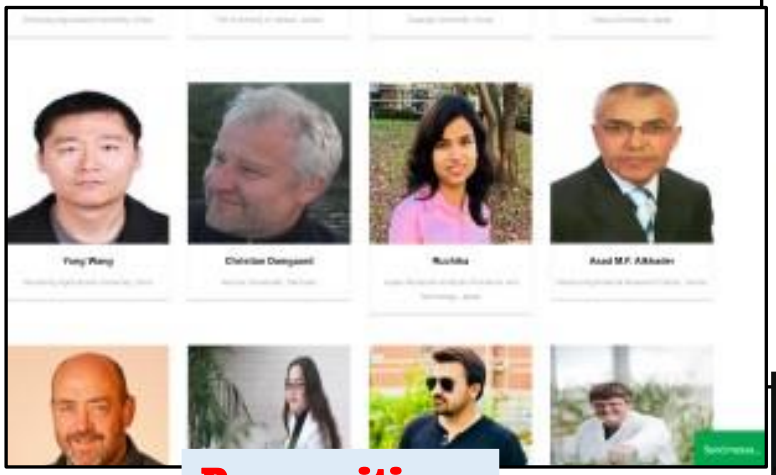
4. **Ruchika**, Toshifumi Tsukahara, Development of RNA editing system using hornwort specific DYW type PPR proteins, **Nature**, Harnessing the plant Microbiome, 22-24th October, 2021. ***(Poster presentation)***

5. **Ruchika**, Toshifumi Tsukahara, Development of bioengineered MS2 system for restoration of genetic code, **Nano today**, Elsevier, 15-18 November 2021. ***(Poster presentation)***

AWARDS AND GRANTS

1. **Young Scientist Award** at plant genomics, 2021.
2. **JAIST research grant (Hoga)** for Doctoral Students, 2021-2022
3. Awarded **JAIST foundation research gran**ts for students to present in international conference.
4. Awarded **JAIST off-campus Research Grant** for conducting research at Kyoto University, japan (December 2019- February 2020).
5. Japanese Government **(Monbukagakusho: MEXT**) University Recommendation Scholarship to work as a PhD scholar in School of Material Science, Japan Advanced Institute of Science and Technology, Japan 2019~

Achievements

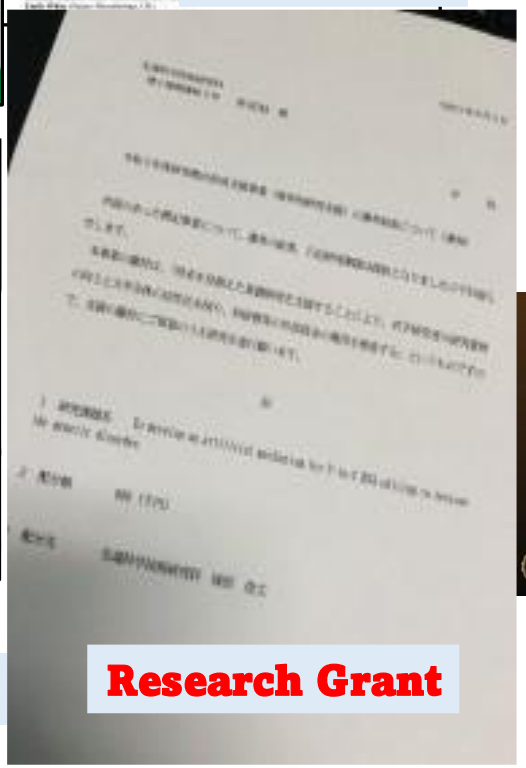


Recognition

Nature conference

Publication

Award



Research Grant

Scholarship

Acknowledgement

Firstly, I wish to express my sincere gratitude to my research supervisor Prof. Toshifumi Tsukahara, Area of Bioscience and Biotechnology, Graduate school of advanced Science at JAIST for providing me with the opportunity to conduct my PhD research in his Laboratory. My dream of studying abroad would have never been possible without his concent. I would like to thank him, who saw the capabilities in me and nurtured my passion for doing a quality research. His kind guidance, valuable suggestions and discussions contributed the necessary impetus in the completion of this work. His utmost faith in me was the prime source of motivation for me to stay focused during the entire time of research and preparation of this thesis.

At the same, I would also like to take this opportunity to thank Prof. Mizuki Takenaka, Plant molecular genetics, Kyoto university, for his constant support and valuable suggestions and constructive critisms throughout my time at a professional level. He was my minor research supervisor (outside JAIST), other than that, he was also a greatest source of inspiration for me to push myself and explore my potential to the fullest. I am very much greatful to him for having virtual meeting for exchanging the acedamic knowledge, onces or twice every month. Also, I am thankful to Professor Sakakibara from Rikyo University, Japan for sharing the hornworts PPR protein databases with us.

I am also grateful to all the Tsukahara lab members, for their timely inputs, cooperation and discussions, for all the time we were working together, and for all the fun we had. I consider myself very lucky to have met so many people along the way in JAIST and gathered memories to be cherished all my life.

I owe a deep sense of gratitude for my seniors and friends for their constant motivation and encouragement. The joyful time I spent on my weekends or holidays was always kept me refreshed and energetic for the upcoming tasks.

I would also like to extend my sincere thanks to JAIST research grants for financial encouragement for attending the International conferences. I was also honoured to receieve JAIST grant (hoga), 2021. I would like to mention Tsukahara Lab (JAIST) for monetary support during off-campus research in Kyoto University. In addition, I am humbled by MEXT Scholarship provided by the Japanese Government throughout my PhD research in Japan.

Lastly, I express my heartfelt gratitude to my parents, brother and sisters for their positivity, relentless support and encouragement in difficult times. They have never questioned my potential. I would like to dedicate this work to my parents, who have always believed in me and my dreams. Their faith makes me stronger every time and fuels my passion to dream big.

Ruchika
Tsukahara Laboratory
JAIST