THESIS

IDENTIFICATION OF DIRECT TARGETS OF SERINE/ARGININE-RICH 45 PROTEIN ISOFORMS BY TRIBE (TARGETS OF RNA-BINDING PROTEINS IDENTIFIED BY EDITING) IN *ARABIDOPSIS THALIANA*

Submitted by

Nikki Huynh

Graduate Degree Program in Cell and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2021

Master's Committee:

Advisor: A.S.N. Reddy

Deborah Garrity
Jeffrey Wilusz

ABSTRACT

IDENTIFICATION OF DIRECT TARGETS OF SERINE/ARGININE-RICH45 PROTEIN ISOFORMS BY TRIBE (TARGETS OF RNA-BINDING PROTEINS IDENTIFIED BY EDITING) IN *ARABIDOPSIS THALIANA*

In plants, as in other eukaryotes, alternative splicing (AS) is an essential post-transcriptional step. This RNA processing converts pre-mRNAs from a gene to several mature mRNA transcripts, which can then produce distinct proteins to increase proteome complexity or regulate gene expression through multiple mechanisms. Small nuclear ribonucleoproteins (snRNPs) that form a spliceosomal complex are recruited onto the splicing sites of the pre-mRNA sequence to generate the final mature mRNAs. Additional aid from SR (serine/arginine-rich) and SR-like proteins is crucial in executing these splicing events.

SR45, an SR-like protein, is a pre-mRNA splicing regulator that assists the spliceosomal complex by recruiting major spliceosome components and identifying the splice sites of the pre-mRNA. In general, SR45 plays significant tasks in spliceosomal assembly as well as other post-transcriptional events, plant development, and stress responses. SR45 is likely to be an ortholog of the human RNPS1 protein, a component of the exon-exon junction complex, which has a major influence in localization, export, surveillance, and translation of spliced mRNA. However, the mechanisms by which SR45 participates in these diverse roles are still largely unknown. SR45 pre-mRNA is also alternatively spliced into two distinct variants, a long isoform (SR45.1) and a short isoform (SR45.2), differing in eight amino acids. Previous work has found that these two isoforms have distinct functions, but also share some common roles. A long-spliced *SR45*

isoform complemented the flower phenotype and seed production in the mutant, whereas a short-spliced *SR45* isoform did not complement either feature. However, *SR45.2* recovered the root phenotype of the mutant, but SR45.1 displayed no change. The molecular understanding of how these isoforms regulate phenotypes remains unknown. Research on SR45 protein's various regulatory functions and its molecular approaches is relatively nascent, and much work is still needed to understand the precise roles of splice variants at the molecular level.

In this study, I aimed to determine the common and distinctive RNA targets of SR45.1 and SR45.2 globally in Arabidopsis to gain insights into how they regulate different biological processes. Identifying the shared and unique targets for each isoform will allow us to gain insights into how these isoforms perform biologically distinct functions and how they share common roles. I used recently developed RNA editing tools, TRIBE and HyperTRIBE, to identify the specific RNA targets of these two SR45 isoforms. It is a simple RNA binding target method that does not require large amounts of starting material or the tedious work of immunoprecipitation assays. This method consists of creating a chimeric protein that fuses the RNA binding protein (RBP) of interest (mediates target specificity) with a deaminase enzyme, ADARcd (*adenosine deaminase acting on RNA* catalytic domain). The editing specificity is determined by the RNA recognition of the protein of interest and the deaminase converts adenosine into inosine that is in the proximity of the target sites.

We have prepared constructs with each *SR45* isoform fused to the *Drosophila* wild type (TRIBE) and mutated (HyperTRIBE) ADAR catalytic domain (dADARcd). To create the HyperTRIBE enzyme domain, a mutation was incorporated to change one amino acid at position 488 (E488Q) of the original *SR45* TRIBE construct using Q5 site-directed mutagenesis. We generated transgenic lines with six different constructs, confirmed the expression of the

introduced constructs in all lines, and the functionality of SR45-dADAR fusions was verified by the complementation of *sr45* mutant phenotypes. Predicted phenotypes were observed for each construct. Homozygous transgenic lines harboring the fusion protein construct (HyperTRIBE-*SR45* short; HyperTRIBE-*SR45* long; TRIBE-*SR45* short; and TRIBE-*SR45* long) and lines expressing TRIBE and HyperTRIBE ADAR catalytic domain alone in the *sr45-1* mutant background along with wild-type and *sr45* mutant were used for RNA-sequence analysis.

For the bioinformatics analyses, I first ran the pipeline on *Drosophila* RNA-seq data to reproduce their results and to familiarize myself with all steps and parameters of the pipeline. I then tailored the pipeline for the Arabidopsis dataset by adjusting the parameters. In the HyperTRIBE-expressing lines, both HyperTRIBE-*SR45* lines (long and short) yielded consistent and reproducible results as well as identified more editing sites compared to the TRIBE-*SR45* lines. Nevertheless, the majority of the edit sites had a low editing efficiency of 1-5% found in both TRIBE and HyperTRIBE lines. We did extensive analyses to determine if these potential transcripts are spliced variants and whether it coincides with a previously published dataset of SR45-associated RNAs (SAR). We also discuss where these target genes are spatially located by comparing our list of target genes with genes that are expressed in specific domains in shoot apical meristem and developing leaves [Tian et al., 2019]. We found that a substantial amount of target genes from the TRIBE/HyperTRIBE-SR45 lines were derived from spliced transcripts and are from the *CLV3 (CLAVATA3)* domain, which is responsible for controlling the size of the shoot apical meristem (SAM). This evidence suggests that SR45 has a role in regulating alternative splicing in the meristem tissues. A gene ontology (GO) analysis on the target transcripts was performed to further reveal biological processes that are affected by the targets of each isoform. Based on the GO terms, we found that SR45.1 potentially has a major regulatory

role in external-response stimulus compared to SR45.2. Each SR45 isoforms also had distinct GO terms implicated in different aspects of the cell cycle. However, we did observe both SR45 isoforms sharing common targets pertaining to plant development as well as intracellular trafficking. Overall, this study led to the identification of unique and common targets of each isoform and provided some insights into how these isoforms may function in distinct and shared functions in regulating developmental and stress responses.

ACKNOWLEDGEMENTS

I would like to thank Dr. Anireddy Reddy, Dr. Prasad Kasavajhala, and the rest of the members of our lab for supporting me throughout the years and always believing me. I appreciated all of the time and effort that they had given to me in the aim to guide me and strengthen my skillsets as a scientist. I would also like to thank my family, peers, and my partner, Erik, for supporting me and encouraging me to pursue and finish this degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Regulation of gene expression

In most eukaryotic cells, gene expression is continuously modulated by various regulatory pathways throughout the central dogma of molecular biology including at the co-/post-transcriptional level. Major post-transcriptional modifications include the 5' capping, 3' polyadenylation, and precursor mRNA (pre-mRNA) splicing. Co-/post-transcriptional splicing is categorized into two types: constitutive and alternative splicing. Both participate in regulating the abundance of gene expression at the RNA level. With constitutive splicing, a multi-exon gene is processed and spliced in the same exact manner by using one set of consensus splicing sites and consequently, generating only one type of mRNA transcript. However, alternative splicing allows regulated production of two or more distinct mRNAs and protein variants from a single gene via differential usage of splice sites [Keleman et al., 2013]. Alternative splicing can rapidly enhance transcriptome complexity thus increasing the diversity of proteins produced from eukaryotic genomes and reprogram gene expression in response to diverse internal and external cues [Reddy et al., 2013].

Co-/post-transcriptional processing

Recent studies in animals and plants have shown that pre-mRNA splicing mainly occurs during transcription [Glono and Kornbliht, 2020; Reddy et al., 2020]. In animals, major spliceosomal components and splicing regulators initiate the formation of the spliceosome by interacting with the carboxy-terminal domain of the RNA polymerase II co-transcriptionally, while the completion of the splicing catalysis happens co-transcriptionally or post-

transcriptionally [Pandya-Jones and Black, 2009; Luco et al., 2011]. Spliceosome assembly is spatially linked with RNA polymerase II (RNAP II), implicating temporal coordination between splicing and transcriptional elongation [Martins et al, 2011]. Spliceosomal components including splicing factors are enriched in subnuclear membraneless compartments that are formed by liquid-liquid phase separation (LLPS) [Herzel et al., 2017]. Recent studies have shown that a large percentage of the alternative splicing events in plants occurs co-transcriptionally and its splicing efficiency is dependent on certain features, such as total intron number, intron position, splicing, and gene expression level [Zhu et al., 2020; Li et al., 2020]. In recent evidence, the majority of the introns in protein-coding genes that undergo co-transcriptional splicing had a higher splicing efficiency compared to introns in non-coding RNAs. [Li et al., 2020]

This proximity and interaction between the proteins involved in transcription and splicing suggested that achieving efficient and productive co-transcriptional processing is done by transcriptional and spliceosomal machinery working concurrently in these membraneless organelles. For brevity, I will be using the term, “post-transcription”, instead of “co-transcription” throughout this thesis.

Membraneless organelles

Membraneless organelles are multicomponent structures containing many proteins and RNA (or DNA) in the nucleoplasm or cytoplasm generated by liquid-liquid phase separation [Gomez and Shorter; 2019]. Because of their morphology and dynamics, they were initially recognized as organized storage bodies of proteins and RNA. Initially, these cellular bodies were described as compartments of the cell to partition specific biochemical processes from one another and have been distinguished in various forms as stress granules, nuclear speckles,

nucleolus, and photobodies. Aforementioned, nuclear speckles are recognized as a nuclear compartment for RNA-binding proteins, especially a high fraction consisting of splicing factors [Fang et al., 2004; Spector and Lamond, 2011]. Nuclear speckles are shown nearby pre-mRNAs and active transcription sites, where splicing factors and other mRNA-associated proteins would continuously diffuse out of speckles to bind onto the mRNA and execute steps of splicing. The morphology and size of these nuclear bodies are dependent on the splicing proteins composition which varies depending on the cell state and transcriptional activity [Ali and Reddy, 2006; Lorkovic´ and Barta, 2004; Fang et al., 2004; Tillemans et al., 2005].

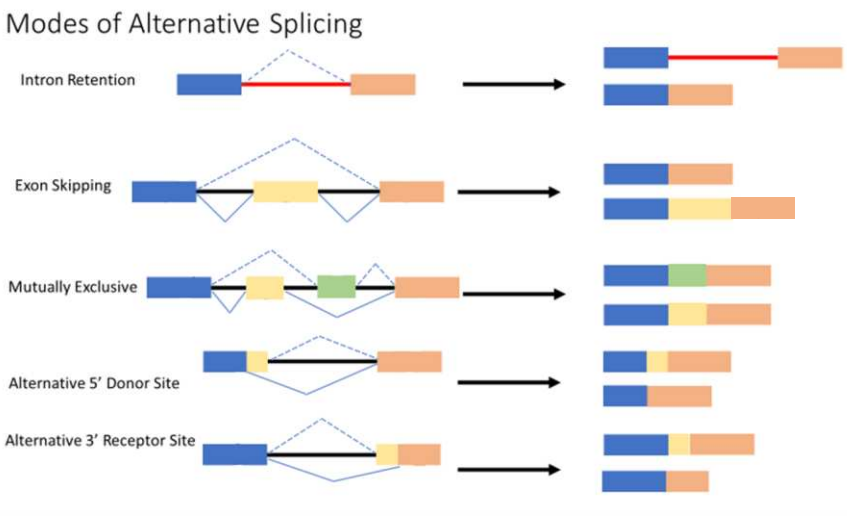
Alternative splicing

Alternative splicing generates a variety of mature RNA transcripts from pre-mRNAs by ligating exons dictated by an assorted combination of splicing sites. There are five primary modes of alternative splicing that are traditionally depicted in basic splicing mechanisms: intron retention, exon skipping, mutually exclusive exons, alternative 5' donor site, and 3'acceptor site (Figure 1). The splicing machinery ligates these exons, leaving a pool of the final, processed mRNA spliced products. These alternatively spliced mRNA isoforms are distinctive from each other in their sequence and could then be translated to fully functional proteins with altered biological roles. The splice variants not only increase proteome complexity but also fine-tune gene expression in diverse ways by affecting the transport, stability, localization, and/or translatability of mRNAs under normal and stress conditions, which will be further elaborated below.

Comparative analysis of splicing between plants and animals

Alternative splicing occurs frequently in plants and animals; it occurs in about 95% of multi-exonic genes in animals and over 60% of genes in plants [Filichkin et al., 2010; Kayna et al., 2012, Wang et al., 2008] (Figure 1). A majority of alternative splicing events in plants are intron retention (IR) while the most prevalent mode of alternative splicing witnessed in mammals is exon skipping [Reddy et al., 2012]. However, a study in 2014 has shown that IR is prevalent in mammals as well [Braunschweig et al, 2014]. Comparative analysis on splicing pattern differences between plants and animals is reflected in the variation of their intron composition, and length [Reddy, 2001].

Analysis of splicing of plant pre-mRNAs in HeLa cell splicing extracts has shown inaccurate splicing of introns [Brown et al., 1986; Hartmuth and Barta, 1986]. Similar imprecise splicing results were reported when expressing animal pre-mRNAs in plants [van Santen and Spritz, 1987]. The incorrect splicing of plant pre-mRNAs in animal splicing extract and animal pre-mRNAs in plants suggest some inherent differences in splice site recognition between plants and animals. Yet the core components of the splicing machinery and splicing process are the same; the spliceosomal assembly and its major components are conserved in both systems [Simpson et al., 2002; Reddy, 2007, Reddy et al., 2013]. Moreover, the conserved sequences that define the 5' splice site, 3' splice site, and branch point are analogous between plants and metazoans. Another shared concept between these species, the recognition of the 5' site is identified by the U1snRNP, U1-70K, and non-snRNP splicing factors, while the recognition of the 3' site is targeted by a set of U2 auxiliary factor (U2AF) proteins, which are responsible for recruiting U2snRNP on the pre-mRNA [Kohtz et al., 1994; Golovkin and Reddy, 1996; Day et al., 2012; Wu et al., 1999; Wahl et al., 2009; Mackereth et al., 2011].



Observed Frequency	
Humans	Plants
~5%	~60%
~40%	~8%
~10.7%	N/A
~7.9%	~7.5%
~18.4%	~15.5%

Figure 1. Different modes of alternative splicing. the most prevalent types of alternative splicing and the frequency of each event in humans and plants. [Adapted from Reddy, A. S. N., Marquez, Y., Kalyna, M., and Barta, A., Plant Cell, 2013]

Epigenetic modifications in chromatin landscape affect splicing

Another layer in this tightly coordinated network to regulate RNA splicing is epigenetic modifications to the chromatin landscape. Transcriptional elongation rate is dependent on the accessibility of DNA sequence for RNAP II binding, which is also affected by nucleosome positioning, DNA methylation status, histone modification, and chromatin architecture [Luco et al., 2011]. Influencing the transcription dynamics can orchestrate the selection of strong and weak splice sites, the timing of the spliceosomal assembly, and overall alternative splicing patterns [Luco et al., 2011; Giono and Kornbliht, 2020]. Studies in animals and plants elucidated that the nucleosomes are enriched in exons, especially around the exon-intron and intron-exon boundaries [Schwartz et al., 2009; Tilgner et al., 2010; Braunschweig et al., 2013; Chodavarapu et al., 2010; Jabre et al., 2021]. Thus, nucleosome positioning crucially distinguishes exons and consequently influences splicing.

Moreover, epigenetic marks and the chromatin state play a crucial role in cultivating a splicing memory for plants to prime themselves from stresses [Listerman et al., 2006; Thiebaut, 2019; Ling et al., 2018]. Stress-dependent chromatin modulation mediates the generation of spliced transcripts temporally and spatially to allow plants to adapt in the short and long term. However, it is unclear how environmental signals are capable of triggering these epigenetic modifications to provide plants plasticity from exposure to multiple stresses.

Roles of alternative splicing in mRNA stability

Alternative splicing does not only just diversify the cell with proteome complexity but rather regulates the type and amount of mRNA transcripts being generated. First, alternative splicing contributes significantly as a post-transcriptional mRNA quality control. Even though

multiple transcripts are being spliced, some of these spliced variants are found in undetectable quantities at the protein level, suggesting that expressed genes normally have one major isoform [Ezkurdia et al., 2015; Tress et al., 2017a,b; Abascal et al., 2015]. Especially in plants, spliced transcripts with intron retention are most likely to harbor premature terminal codons (PTC), creating nonsense mRNAs that are directed to three fates: be sequestered to be processed at a designated time, be degraded by nonsense-mediated decay (NMD) pathway, or escaped this degradation mechanism to be translated as either truncated, non-functional or truncated, different functional proteins [Filichkin and Mockler, 2012; Kalyna et al., 2012] (Figure 2). Emerging evidence indicates that under stress or different developmental stages, most PTC-containing transcripts are sequestered in the nucleus either for further mRNA processing or degradation in a time-dependent manner, reaffirming that splicing is coupled with NMD and sequestration [Boothby et al., 2013; Filichkin S.A. et al., 2015; Gohring et al., 2014; Hartmann et al, 2018]. The NMD machinery in Arabidopsis is a consequential mRNA surveillance mechanism in regulating the function, timing, and abundance levels of productive spliced transcripts.

Roles of alternative splicing in translation and translation efficiency

MicroRNAs (miRNAs) are also coupled with alternative splicing to attenuate gene expression and ultimately influence the stability and translation of target mRNAs [Bartel, 2009; Voinnet, 2009]. Splicing controls miRNA-mediated regulation of gene expression in multiple ways. Target transcripts of miRNAs, especially regulatory proteins involved in plant growth, development, and stress responses, are alternatively spliced where candidate miRNA sites of these transcripts are either present or lost (Figure 3) [Rogers and Chen, 2013b]. Splicing also indirectly affects miRNA regulation by altering miRNA biogenesis by modulating splicing

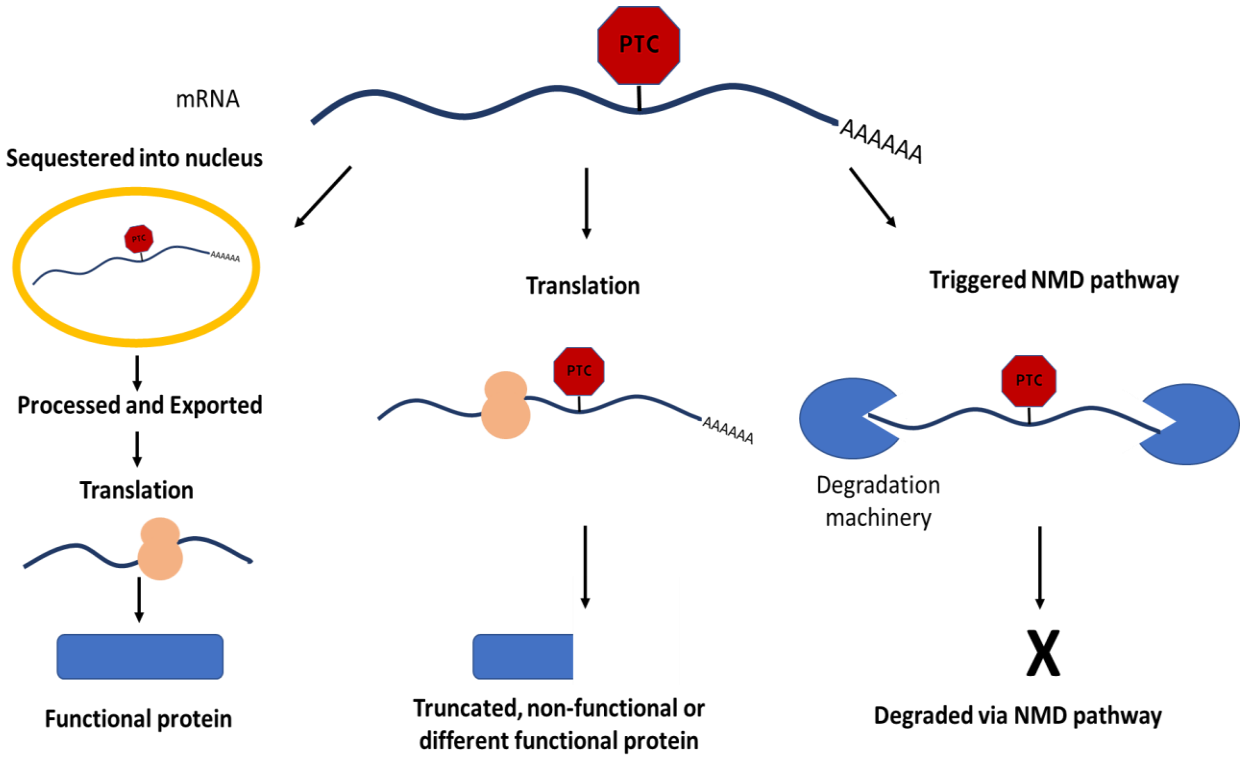


Figure 2. Multiple fates of PTC+ transcripts. The different fates of premature terminal codon (PTC+) transcripts: 1) PTC+ transcript is sequestered in the nucleus to be processed at a later time and exported out to the cytoplasm to be translated into a functional protein, 2) PTC+ transcript escapes the degradation pathway and is translated either into a truncated, non-functional protein or a different functional protein, 3) PTC+ transcript triggers the nonsense-mediated mRNA decay (NMD) pathway and is degraded.

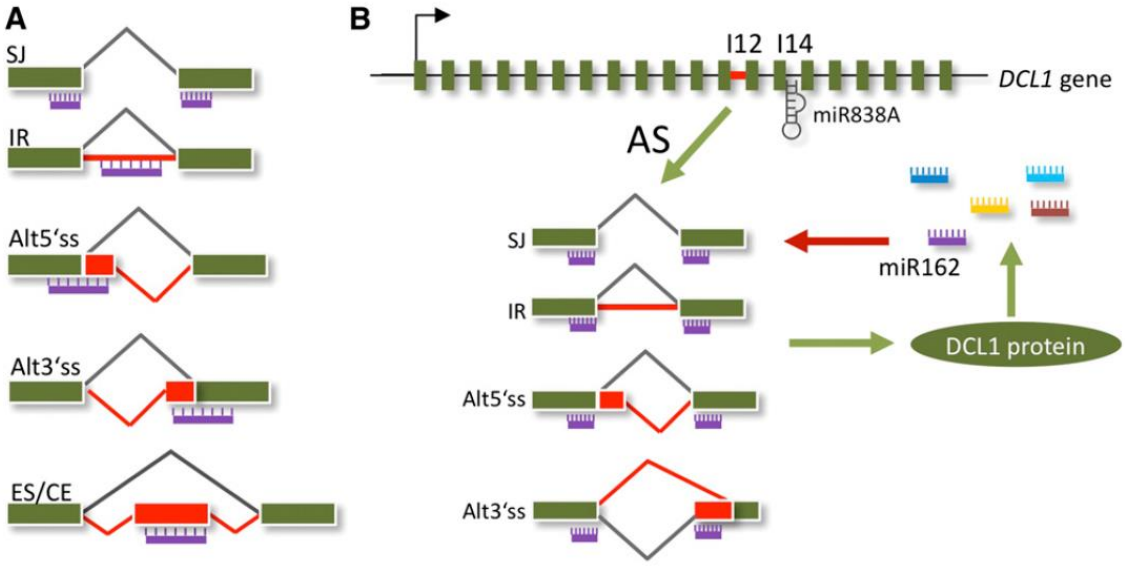


Figure 3. Alternative splicing regulates miRNA production. A) The various type of splice variants that contain or lack target sites of miRNA. SJ, splice junction; ES/CE, exon skipping/cassette exon ; Alt5'ss, alternative 5' splice site ; Alt3's ; alternative 3' splice site ; IR, intron retention. The red lines indicate alternative splicing events. B) Alternative splicing patterns of the dicer-like 1 (DCL1) pre-mRNA transcript, a crucial endonuclease involved in miRNA production, alters the expression level of miRNAs. miRNA binding site for miR162 (violet) is split between E12 and E13 of DCL1 transcript; miR838A is encoded in the intron 14 of DCL [Reproduced from Reddy et al., Plant Cell, 2013].

patterns of pri-mRNAs encoding enzymes, such as dicer-like 1 (DCL1) protein, a crucial endonuclease involved in miRNA production.

Translation efficiency is also shown to be regulated by alternative splicing by a process called intron-mediated enhancement (IME) [Rose, 2018]. In plants, IR transcripts are retained frequently, suggesting that the RNAP II processivity may be dependent on specific intronic regions. For instance, in Arabidopsis, studies have shown that the MHX (magnesium/proton exchanger) gene with an extended 5' UTR intron was responsible for enhancing gene expression and that localization of the intronic element is dependent on its ability to induce expression levels [Akua and Shaul, 2013; Meng et al, 2021]. However, the underlying mechanism of IME in plants has not been fully illuminated.

Roles of alternative splicing in plant stress tolerance

Plants have developed defense mechanisms to acclimate to multiple, recurring, and chronic stress responses [Vinocur and Altman, 2005]. "Molecular stress memory" was coined to demonstrate how plants achieve stress tolerance by priming themselves in response to the initial episodes of stress exposures [Sani et al., 2013; Hilker et al., 2016]. Especially in temperature stress, several altered transcripts of SR proteins were generated, some appeared even as new spliced variants [Palusa et al., 2007; Filichken et al., 2010; Lopato et al., 1999]. This supporting evidence suggested that abiotic stress modulates the splicing patterns of the pre-mRNA of splicing factors, leading to changes in splicing patterns of pre-mRNAs and thereby an altered transcriptome for plant survival. Accumulated IR transcripts that escaped the NMD pathway were shown to be associated with stress-inducible isoforms. In a heat stress study, a substantial increase of splice variants was reflected in non-primed plants and the most prevalent altered transcripts contained IR [Ling et al., 2018]. Whereas primed plants had comparable splicing

patterns as the wild-type plant, that have not been subjected to the stress. Primed plants have an improved adaptation to different environmental cues by fine-tuning splicing patterns and differentially expressing abiotic responsive genes associated with stress memory [Ling et al., 2018; Sanyal et al., 2018].

Major components of the spliceosome

In both constitutive and alternative splicing mechanisms, the initial splicing step follows the transcription of the pre-mRNA from DNA. This transcript is composed of regions of non-coding sequences, introns, and regions of protein-coding sites, exons. Introns are excised out from the sequence and the remaining exons are ligated together as the final processed mRNA transcript. These splicing patterns occur during the trans-esterification reactions done by a highly dynamic complex called the spliceosome. The spliceosome is composed of five major small nuclear ribonucleoproteins (U1, U2, U4, U5, U6 snRNPs), and these spliceosomal subunits constantly reassemble and rearrange throughout each step of the splicing process to execute transesterification reactions. However, these subunits cannot associate with each other as a complex or perform these splicing events independently. These subunits are recruited with the assistance of other mRNA-binding proteins, where the majority of them belong to the conserved serine/arginine (SR) protein family.

***Trans*-acting factors and *cis*-regulatory elements**

Splicing is mediated by a complex, interconnected network of mRNA-binding proteins that conduct synergistic effects by binding one to another on the "splicing code" of the pre-mRNA. The synthesis of these spliced variants is dictated by the binding of *trans*-acting proteins to *cis*-acting RNA sequences on the precursor mRNA. These *trans*-acting RNA binding proteins

can either be splicing activators, proteins that optimize the chance of defining a particular spliced site, or repressors, proteins that serve to inhibit the selection of a spliced junction. *Cis*-acting regulatory sites correspond to a certain *trans*-acting protein based on its role. These regulatory sequences are grouped into two majority types: splicing enhancers, sequences that splicing activators bind to, and silencers, RNA sites that splicing repressors bind to. These RNA sequence elements can either be found in the intron or nearby exons targeted for splicing. They differ from each other in nucleotide sequences and facilitate different protein-protein interactions. SR proteins are defined as these *trans*-acting regulatory elements and play a crucial role in pre-mRNA splicing.

Heterogeneous ribonucleoprotein particles (hnRNPs)

Another essential class of RNA binding protein found in plants and animals is heterogeneous ribonucleoprotein particles (hnRNPs), which are tightly linked with mRNA processing like splicing and are temporally associated with nuclear RNA and mRNA in the cytoplasm [Wachter et al., 2012]. Their splicing roles resemble intronic/exonic splicing silencers. For instance, in a mammalian study, hnRNP-like protein in the neuro-oncological ventral antigen (NOVA) family was found to suppress splicing when bound to an intron that is either before or after the alternatively spliced exon [Park et al., 2011]. They inhibit splicing by blocking access of the spliceosome to the polypyrimidine tract which is regulated by the phosphorylation of some components of the hnRNP complex, the protein K, and the polypyrimidine tract binding protein. These splicing regulators, similar to SR proteins, govern many plant developmental processes, post-transcriptional regulation, telomere regulation, and in general, gene expression level [Beyer et al., 1977; Martinez-Contreras et al., 2007; Eversol and Maizels, 2000].

Serine/Arginine-Rich Proteins (SR Proteins)

The SR protein family is diverse; in Arabidopsis, 19 genes encode SR proteins, and the primary transcripts of 15 of these genes are also alternatively spliced into 95 different isoforms [Kalyna and Barta, 2004; Palusa et al., 2007; Reddy, 2004]. SR proteins are spatially and temporally distributed throughout the plant, suggesting their involvement in various regulatory processes and interaction networks with other proteins [Lopato et al., 1999, Reddy, 2004; Reddy, 2007; Lazar and Goodman, 2000]. Plants have twice as many SR proteins compared to metazoan organisms where 11 out of 19 Arabidopsis SR proteins had no direct counterparts; further lending support to the hypothesis that these unique SR proteins have evolved to perform plant-specific functions. Plant SR proteins are grouped into 6 different subfamilies and 3 of them are conserved in animals, SRSF1 (SR subfamily), SRSF2 (SC subfamily), and SRSF7 (RSZ family) [Barta et al., 2010; Manly et al., 2010] (Figure 4). There is some occurrence of functional redundancy between the plant and animal SR proteins, but in plants, the SR protein family is more diverse and also has distinctive plant-specific functions which is the result of genome amplification, specifically interchromosomal duplication events [Duque, 2011].

Circular RNA and backsplicing

Circular RNAs (circRNAs) biogenesis results in primarily non-coding single-stranded RNAs that are covalently linked to form a closed-loop structure where the downstream 3′ donor splice sites are linked to the upstream 5′ acceptor splice sites (Figure 5) [Jeck et al., 2013; Memczak et al., 2013; Jeck and Sharpless, 2014]. This event is also known as “back splicing”. Found in all eukaryotes, circRNAs play a role in modulating alternative splicing and regulating plant growth, developmental, and stress responses [Ren et al., 2018; Sun et al., 2016]. CircRNAs

Subfamily name	Aliases	New protein /gene symbol	Accession
SR subfamily	At-SRp30	At-SR30	At1g09140
	At-SRp34, SR1	At-SR34	At1g02840
	At-SRp34a	At-SR34a	At3g49430
	At-SRp34b	At-SR34b	At4g02430
	Os-SRp32	Os-SR32	Os03g22380
	Os-SRp33a	Os-SR33a	Os05g30140
	Os-SRp33b	Os-SR33	Os07g47630
	Os-SRp20*	Os-SR40	Os01g21420**
RSZ subfamily	At-RSZp21, SRZ21	At-RSZ21	At1g23860
	At-RSZp22, SRZ22	At-RSZ22	At4g31580
	At-RSZp22a	At-RSZ22a	At2g24590
	Os-RSZp21a	Os-RSZ21a	Os06g08840
	Os-RSZp21b	Os-RSZ21	Os02g54770
	Os-RSZp23	Os-RSZ23	Os02g39720
SC subfamily	At-SC35	At-SC35	At5g64200
	Os-SC35a	Os-SC34	Os08g37960
	Os-SC35b	Os-SC32	Os07g43050
	Os-SC35c	Os-SC25	Os03g27030
SCL subfamily	At-SCL28	At-SCL28	At5g18810
	At-SCL30	At-SCL30	At3g55460
	At-SCL30a	At-SCL30a	At3g13570
	At-SCL33, SR33	At-SCL33	At1g55310
	Os-SCL25	Os-SCL25	Os07g43950
	Os-SCL26	Os-SCL26	Os03g25770
	Os-SCL30a	Os-SCL30a	Os02g15310
	Os-SCL30b	Os-SCL30	Os12g38430
	-	Os-SCL28	Os03g24890
	-	Os-SCL57	Os11g47830
RS2Z subfamily	At-RSZ32	At-RS2Z32	At3g53500
	At-RSZ33	At-RS2Z33	At2g37340
	Os-RSZ36	Os-RS2Z36	Os05g02880
	Os-RSZ37a	Os-RS2Z37	Os01g06290
	Os-RSZ37b	Os-RS2Z38	Os03g17710
	Os-RSZ39	Os-RS2Z39	Os05g07000
RS subfamily	At-RSp31a	At-RS31a	At2g46610
	At-RSp31	At-RS31	At3g61860
	At-RSp40, At-RSp35	At-RS40	At4g25500
	At-RSp41	At-RS41	At5g52040
	Os-RSp29	Os-RS29	Os04g02870
	Os-RSp33	Os-RS33	Os02g03040

Figure 4. SR proteins subfamilies. There are six different subfamilies of SR proteins. The SR subfamily has a mammalian ortholog, SRSF1, the SC subfamily has a mammalian ortholog, SRSF2, the RCZ subfamily has a mammalian ortholog, SRSF7. The plant-specific SCL subfamily has an N-terminus charged extension and the plant-specific RS2Z subfamily has two zinc knuckles (2 ZnK) in between an RNA recognition motif (RRM) and two SP-rich regions. The plant-specific RS subfamily proteins contain two RRMs on the N-terminal and an arginine-serine rich (RS) domain rich on the C- terminal [Barta et al., Plant Cell, 2010].

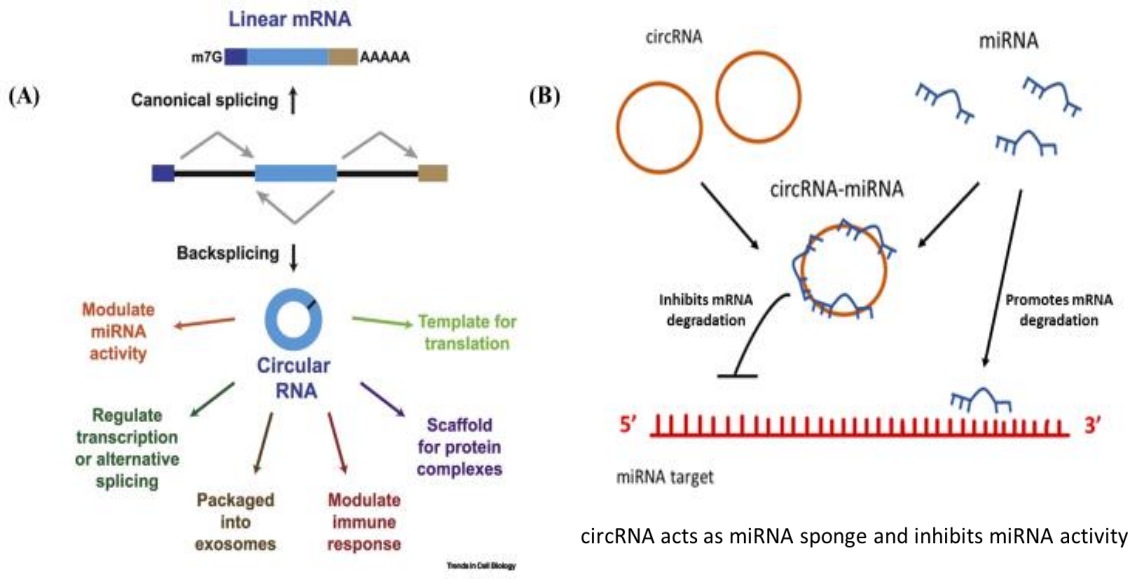


Figure 5. Regulatory roles of CircRNAs. A) In canonical splicing, splice events occur in a linear manner by using the consensus splicing sites; however, in backsplicing, the mRNA is covalently linked into a loop at the 5' and 3' ends. CircRNAs have been shown to modulate different regulatory mechanisms [Reproduced from Xiao, M.-S., Ai, Y., and Wilusz, J.E., Trends Cell Biol, 2020]. B) CircRNAs can sequester miRNA and prevent them from binding to their miRNA binding site and thus inhibiting mRNA degradation.

also serve as a “miRNA sponge”, meaning that these noncoding family members contain miRNA-binding sites which when bound, serve as a suppressor on the miRNA activity [Hansen et al., 2013]. Differential expression levels of miRNAs, circRNAs, and mRNAs were observed in Arabidopsis leaves suggesting that circRNA has a role in leaf senescence, induced expression of circRNA have been identified in Arabidopsis leaves during a pathogen attack implying a regulatory role in plant immunity [Meng et al., 2018; Sun et al., 2016].

There are a few putative Arabidopsis circRNAs experimentally verified but overall a small fraction, about 5% of them constituted as miRNA sponges [Ye et al., 2015]. These circRNAs also contain much fewer miRNA-binding sites compared to animals. Though circRNAs participate in the miRNA pathway, this raises the question of whether circRNA corroborates with SR proteins in a synergistic relationship or circRNA has an antagonistic association via SR protein-mediated regulation of circRNA biogenesis. Further studies are needed to address these questions.

Domains in SR proteins

The diversified functions of SR proteins may also be dependent on their domain structures. Generally, SR proteins are characterized to have either one or two RNA recognition motif (RRM) domains at the N-terminus and an arginine/serine-rich (RS) domain residing at the C-terminus region. The RRM domain mediates the RNA binding specificity by attaching itself to certain regulatory sequences of the pre-mRNA transcript. The RS domain is responsible for mediating interactions with other mRNA-associated proteins as well as major spliceosomal subunits during the splicing process [Golovkin and Reddy, 1999; Day et al., 2012; Ali et al., 2007] (Figure 6). These proteins are not just limited to being splicing regulators but have been

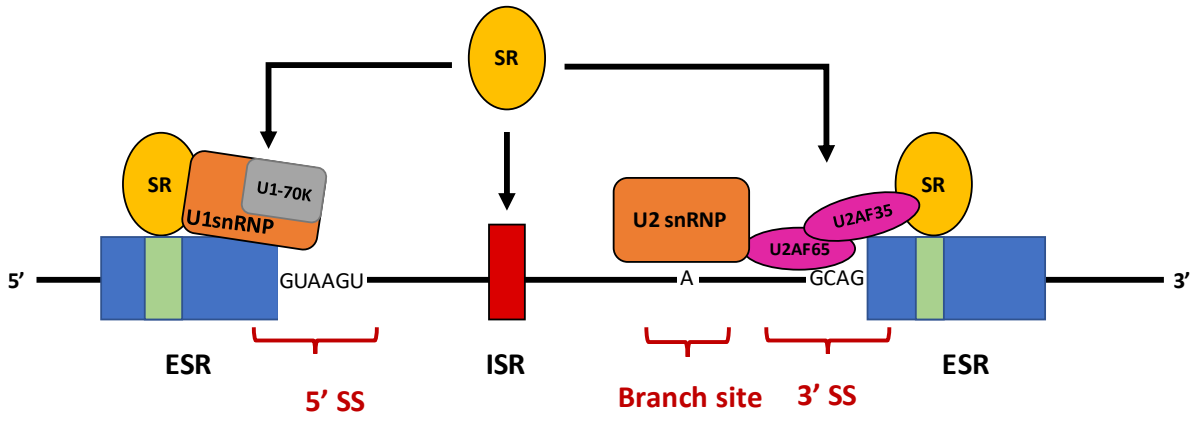


Figure 6. Roles of SR proteins in pre-mRNA splicing. SR proteins assist in the spliceosome assembly by interacting with *cis*-regulatory elements and other *trans*-acting elements. An SR protein binds to the ESR by the 5' splice site which recruits the U1-70K protein, a subunit of U1snRNP, onto the pre-mRNA. Another SR protein binds to ESR by the 3' splice site and stabilizes and associates with U2AF65 and U2AF35, a heterodimer that binds to the polypyrimidine tract upstream of the 3′ splice site. This set of splicing proteins then recruits the U2snRNP, one of the major small ribonuclear proteins that constitute the spliceosome, onto the branch site. Lastly, an SR protein binds to the ISR to mediate the bridge between the 5'splice site and 3' splice site to initiate the beginning steps of the spliceosomal assembly. SR, serine/arginine-rich protein; SS, splice site; ESR, exonic splicing regulatory sequence; ISR, intronic splicing regulatory sequence; U1snRNP and U2snRNP, U1 and U2 small ribonucleoproteins; U1-70K; U2AF65 and U2AF35, U2 auxiliary factor 65 and U2 auxiliary factor.

shown to modulate nuclear localization and export, mRNA stability, splicing events, and translation [Ali and Reddy, 2006; Ali et al., 2003, 2008]. Prior findings have observed that SR proteins are essential pre-mRNA splicing regulators yet further investigations on their other functions in plants are needed.

Localization and dynamics of SR proteins

RNA-binding proteins involved in splicing are common targets for phosphorylation which significantly impacts the timing and coordination of splicing. The localization and dynamics of plant splicing regulators are heavily regulated by phosphorylation, ATP, and transcription [Ali and Reddy, 2006]. The domains of SR proteins are hypo-phosphorylated initially as they reside in the nuclear bodies (Figure 7). Soon the SR proteins are phosphorylated by nearby SR protein kinases, causing these splicing regulators to diffuse out from the speckle and facilitate protein-protein interactions or catalyze a splicing reaction [Reddy, 2008; Barta et al., 2008]. Besides, in response to ATP depletion, AtSR34 and AtSR45 displayed slower mobility in the nucleoplasm and different localization patterns, demonstrating how the mobility and kinetic dynamics of SR proteins are ATP dependent [Ali and Reddy, 2006]. Transcription also plays an essential factor in the mobility and assembly of nuclear speckles. Instead of numerous small speckles, inhibition of transcription redistributed SR proteins to be accumulated in a larger speckle as well as an abrupt cease to their movements [Ali et al., 2003, Tillemans et al., 2005, Ali and Reddy, 2008b]. Influenced by these key factors, many of these splicing proteins, transcription factors, and snRNP proteins are accumulated in these nuclear storage bodies to later execute their biological functions in a temporally and spatially controlled manner.

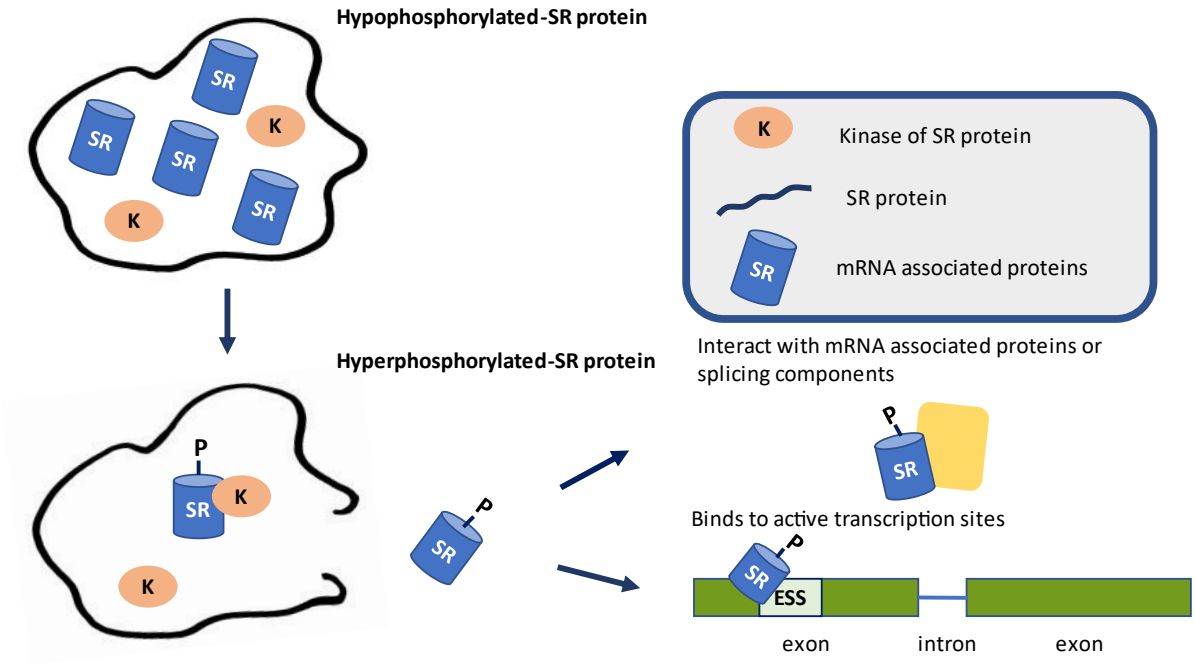


Figure 7. Phosphorylation of SR proteins affects their localization and dynamics. The activity/function and localization of the SR proteins are dependent on their phosphorylation/dephosphorylation status. SR proteins initially reside in nuclear bodies, but once SR proteins are phosphorylated by an SR protein kinase, it is diffused out to the speckles where it facilitates protein-protein interactions or binds to active transcription sites.

Arabidopsis Splicing Factor, SR45

SR45 is an SR-like splicing regulator that assists in the assembly of the spliceosomal complex by recruiting the major spliceosomal components together as a whole structure and facilitating splice site recognition in pre-mRNA transcripts [Ali et al., 2007]. This protein has a distinctive arrangement of its domain structure where the RS domains flank on both sides of the RRM domain (Figure 8). Despite its unique domain structure compared to other SR proteins, SR45 is considered a bona fide splicing factor. In a splicing-deficient S100 cell extract, the SR45 protein showed splicing activity on the beta-globin pre-mRNA substrate in a concentration-dependent manner [Reddy, 2004] (Figure 9).

SR45 role in splicing regulation

As mentioned above, SR45 greatly impacts alternative splicing by differentially selecting the 5' and 3' splice sites. In Arabidopsis, SR45 was originally identified as an interacting partner of U1-70K, a major subunit of the U1snRNP that is involved in 5’ splice site recognition, in a yeast two-hybrid (Y2H) experiment [Golovkin and Reddy, 1999]. Later, *in vivo* studies have shown that U1-70K associates with the RS domains of SR45 and colocalizes with SR45 and SR1 proteins in nuclear speckles [Ali et al., 2003]. In another Y2H screen with SR45, it was found to have another interacting protein partner: U2AF^{35}b, one of the U2AF paralog that constitutes the U2AF complex and is known to function in the 3' splice site recognition (Day et al., 2009). Even though three other SR proteins (SCL33, RSZ22, and RSZ21) are known to interact with U1-70K, none of them were reported to associate with U2AF other than SR45, indicating that SR45 has a role in splice site recognition (Golovkin and Reddy, 1998, 1999). Based on this evidence, SR45 presumably binds to the pre-mRNA and recruits U1-70K to the 5'

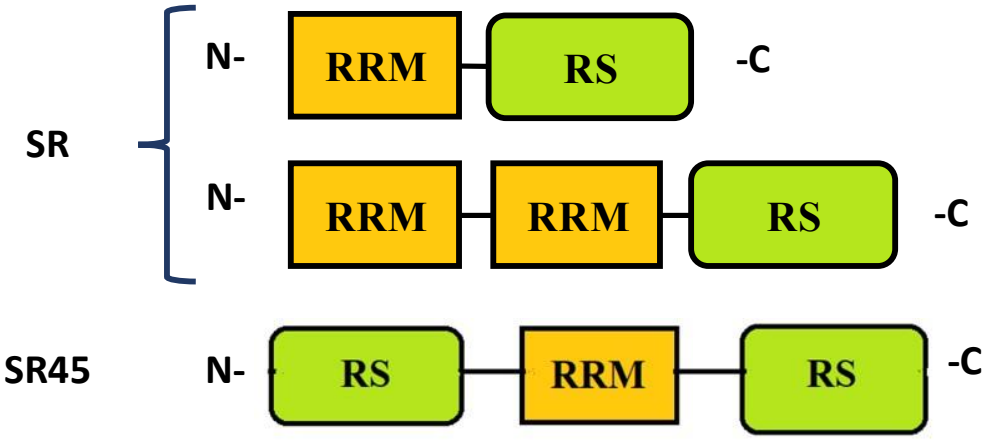


Figure 8. Modular organization of SR and SR45 proteins. SR protein possesses either one or two N-terminal RRM domains, which mediates the RNA binding, and a C-terminal RS domain that facilitates interaction with other partnering proteins. SR45 is an SR-like protein due to its distinct domain arrangement where it has two RS domains flanking on both ends of the RRM domain. RS, serine/arginine-rich domains; RRM, RNA recognition motifs domains.

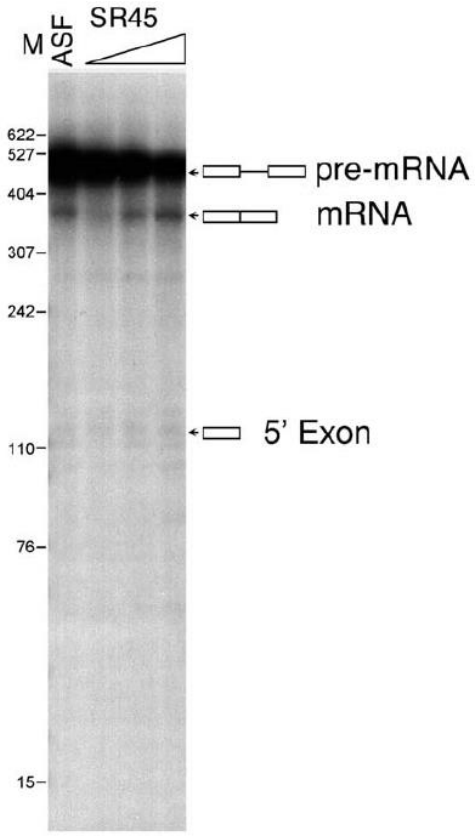


Figure 9. SR45 is a bona fide splicing factor. Reproduced from Ali et al.'s paper, in a splicing-deficient S100 cell extract, different concentrations of purified SR45 protein expressed in insect cells was added to analyze its splicing activity on the β-globin pre-mRNA substrate. These results verified that SR45 was capable of splicing the substrate in a concentration-dependent manner and at a level like the recombinant human ASF (alternative splicing factor). The arrows indicate where the pre-mRNA, spliced mRNA, and 5' exon is positioned on the blot. Boxes indicate the exons, and the line is the intron [Ali et al., PLoSONE, 2007].

splice site and recruits U2AF to the 3' splice site. Then, the U2AF complex guides the binding of the U2snRNP, another major component of the spliceosomal complex, to the branch point of the pre-mRNA sequence. Utilizing bimolecular functional complementation (BiFC) and Y2H assays, these assessments further strengthened the reasoning that U1-70K interacts with both U2AFa and U2AFb proteins, deducing a connection between these sets of proteins [Ali et al., 2008]. Thus, the association between SR45 and these crucial splicing components, U1-70K and the U2AF complex demonstrates a likelihood that SR45 forms a bridge between the 5' and 3' splice sites [Day et al., 2012].

SR45 regulation of plant developmental processes

Not only does SR45 play significant roles in the spliceosomal assembly, but it is also involved in other post-transcriptional processes, plant development, and stress responses. Multiple pleiotropic traits, especially growth and developmental abnormalities and sensitivity to stress cues, were exhibited in the *sr45-1* mutant of Arabidopsis thus presumes the importance of this splicing factor in plant reproduction and defense [Ali et al., 2003]. These affected traits of the *sr45-1* line included delay in flowering, stunted roots, altered number of petals and flower reproductive organs defects, and reduction in seed yield [Ali et al., 2007; Xing et. al, 2015] (Figure 10). SR45 delayed flowering time by increasing the expression of flowering time C (FLC), a flowering repressor [Ali et al., 2007]. In global gene expression analysis with *sr45-1*, high levels of *FLC* and suppressed expression of FLC target genes were discovered. Therefore, SR45 negatively regulates this repressor gene to regulate the flowering time in plants.

RNA-seq analysis with wild-type and *sr45-1* inflorescence tissues identified a total of 542 genes whose pre-mRNA splicing is altered in an SR45-dependent manner. Many of these genes were found to encode mRNA binding proteins [Zhang et al., 2017]. In a gene ontology

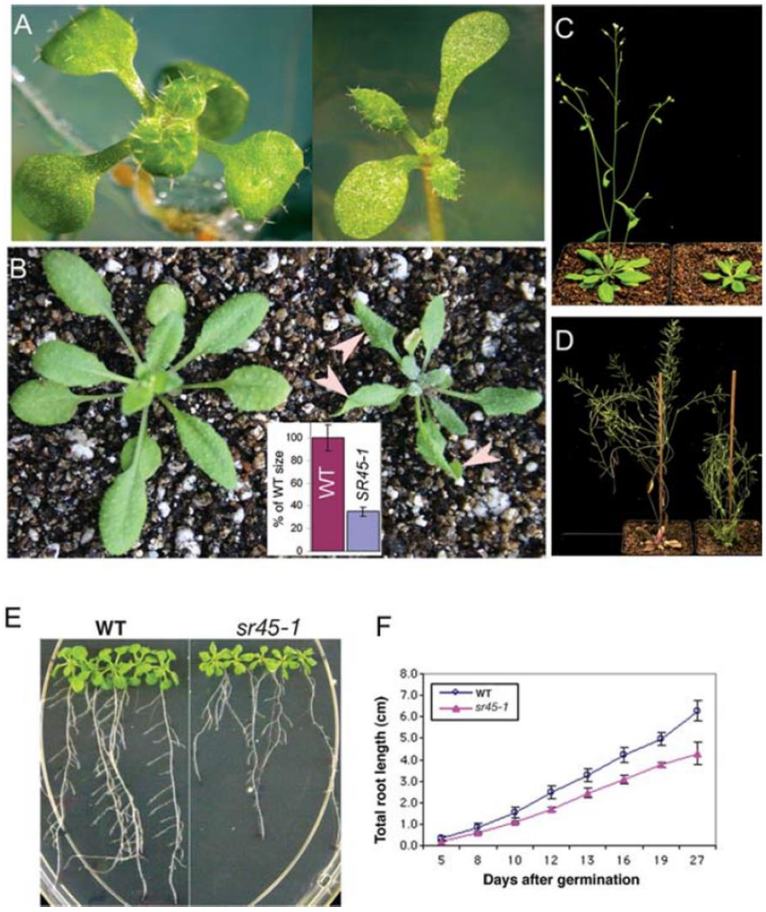


Figure 10. Phenotypic traits of *sr45-1*. A-D) Visual display of various growth stage of *sr45-1* plants under 16/4 hours (day/day) photoperiod. The wildtype line is shown on the left panel and the *sr45-1* line is shown on the right panel. A) The eight-day-old seedlings are grown on MS medium; emerging narrow leaves start to appear on the *sr45-1* plant. B) The twenty-day-old, *sr45-1* plants have pointed leaves and stunted size compared to the wild-type plant. C) Thirty-five-day-old plants are shown. D) Fifty-four-day-old plants are grown. E) Seedlings are grown for twenty-seven days on an MS plate to display the root growth of the wildtype and *sr45-1* lines. *SR45-1* plant has shortened roots compared to wildtype indicating that SR45 may contribute to plant root development. F) A graph showing the total root length after twenty-seven days after germination [Reproduced from Ali et al., PLoSONE, 2007].

analysis, these mRNA-associated proteins function in various post-transcriptional processes such as RNA degradation, stress granule formation, and polyadenylation processing. This further validates that SR45 action in alternative splicing is substantially influential in the regulation of a network of downstream genes modulating different mRNA metabolism processes [Xing et al., 2015]. However, the molecular mechanism of how SR45 operates in these mRNA regulatory processes is still largely unknown.

SR45 modulation of RNA-mediated DNA methylation and other SR proteins expression

The SR45 protein cooperates with other mRNA-associated proteins within many different regulatory pathways such as in RNA-mediated DNA methylation (RdDM). In Arabidopsis, the *sr45-1* mutation in a DICER-LIKE3 (DCL3), ribonuclease III enzyme, mutant background exhibited a major reduction in methylation levels in RdDM-dependent targets, resulting in a late-flowering phenotype [Ausin et al., 2012]. Their double-mutant relationship suggests an additive impairment to DNA methylation, but no further evidence on whether these proteins play a direct role in this pathway was provided. However, reduced small RNA (siRNA) and ARGONAUTE 4 (AGO4) abundance levels were a direct effect from *sr45-1* mutant, revealing that SR45 may be involved in RdDM before the siRNA production.

A few splicing targets of SR45 include other SR genes. SR45 sustains a specific balance in the abundance levels of alternative *SR* transcripts by facilitating the 3' splicing site of the longest intron of *SR* transcripts [Ali et al., 2007]. Normally, *SR* transcripts are spliced at the distal 3' site to amplify more of the shorter transcripts, but *sr45-1* mutant generated a larger pool of longer transcripts from the usage of the proximal 3' splice site (Figure 11). Hence, a balance of *SR* genes is crucial for plant growth and development. For instance, ectopic expression of *SR30* and *RSZ33* genes resulted in abnormalities in the flower and root

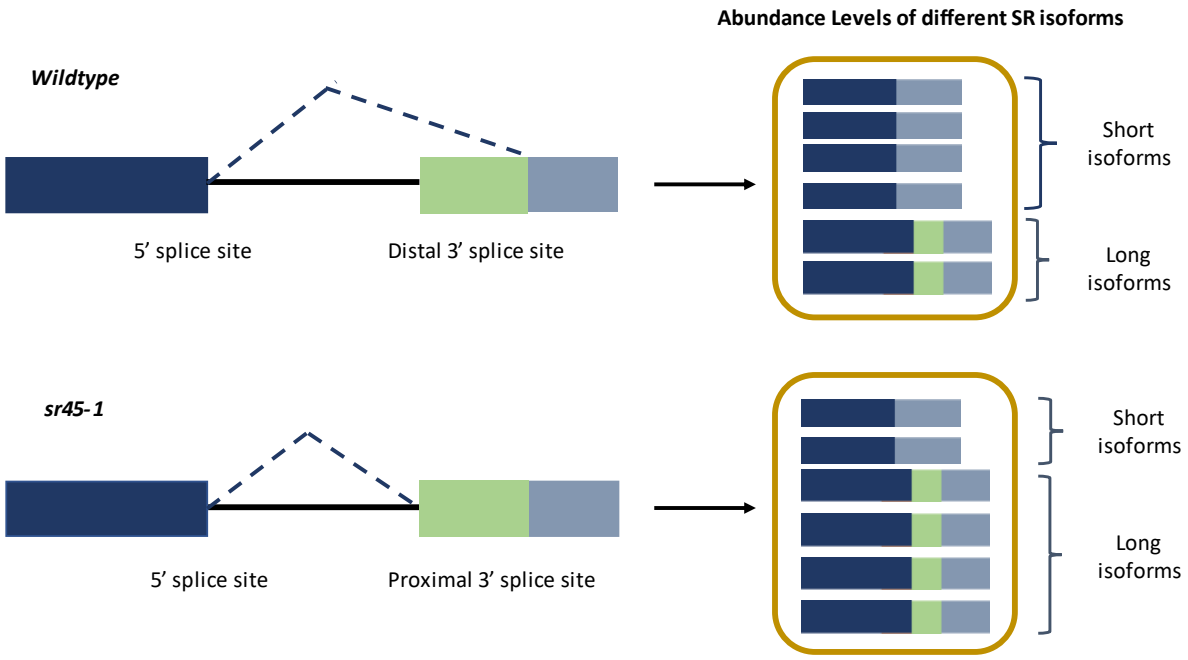


Figure 11. SR45 affects the splicing of other SR protein pre-mRNAs. SR45 regulates the specific abundance level of transcripts of SR protein-encoding genes. The typical alternative spliced pattern of the wildtype prefers the usage of a distal 3' splice site on its longest intron, thus increasing the ratio of its shorter isoforms. In sr45-1, the pre-mRNAs of other SR proteins have an increase in the usage of the proximal 3' splice site, and the abundance of the longest isoforms are significantly induced.

development [Kaylna et al., 2003] (Figure 12). Thus, SR45 modulates the splicing patterns of other *SR* genes, which then regulate the splicing of other downstream genes that are involved in different developmental processes. This cascade effect may be the rationale of how *sr45-1* plants exhibit a pleiotropic phenotype (Figure 13).

The human ortholog of SR45: RNPS1

It has been shown that SR45 is likely to be an ortholog of the human RNA-binding protein with serine-rich domain 1 (RNPS1), a component of the exon-exon junction complex, which has a major influence in splicing events, nonsense-mediated decay (NMD), localization, export, surveillance, and translation of spliced mRNA [Tange et al., 2004; Le Hir et al., 2001; Lykke-Andersen et al., 2001]. RNPS1 and SR45 both have demonstrated binding specificity to selective mRNA transcripts. In a previous study, RNPS1 preferentially associates to an isoform of p34cdc2-related protein kinases, PITSLRE kinases, which provided insights into the role of this kinase role in nuclear speckle and spliceosomal regulation [Loyer et al., 1998]. Different known protein-protein associations of RNPS1 have shown that this protein positively, when associating with p54, and negatively, when interacting with SART3, regulates alternative splicing [Harada et al., 2001; Sakashita et al., 2004]. Thus, the identification of SR45-associated transcripts will elucidate SR45 specificity and the genes that are significantly affected by this splicing factor.

SR45 negatively regulates glucose and ABA signaling pathways

Glucose and hormone pathways are interconnected with plant growth and metabolism since it is reported that glucose amplifies abscisic acid (ABA) content, which modulates growth arrest during unfavorable environmental conditions [Lopez-Molina et al., 2001; Carvalho et al.,

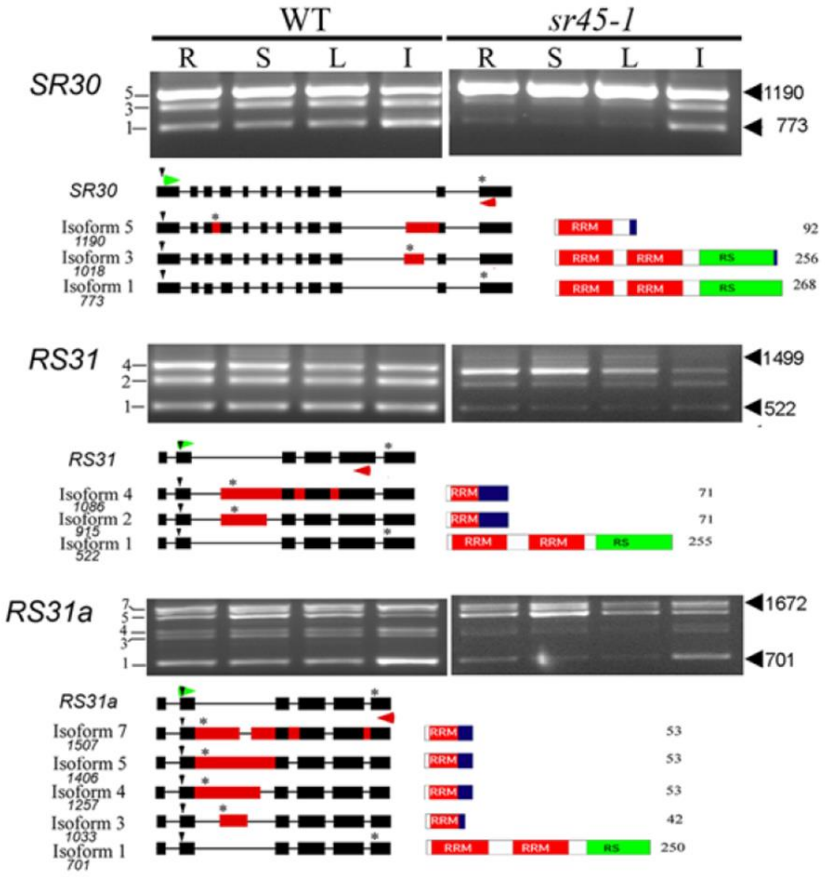


Figure 12. The loss of SR45 resulted in altered alternative splicing of pre-mRNAs of other *SR* genes. The AS pattern of pre-mRNAs of *SR* genes *(SR30, RS31*, and *RS31a*) is changed in *sr45-1*. The expression and AS pattern of *SR* genes were analyzed using RT-PCR with primers specific to each gene. DNA sizes are labeled on the right and the name of each *SR* gene is shown on the left. R, root; S, stem; L, leaf and I, inflorescence. The bottom panel shows the gene structure and alternatively spliced isoforms, and the right of the bottom panel displays the predicted protein domains. The numbers by the predicted proteins are the number of amino acids in the protein. Black boxes represent exons and lines represent introns. Black rectangles suggest constitutively spliced exons and the red rectangles represent the included regions in splice variants. Vertical arrowhead and '*' indicate the start and stop codons; Horizontal green and red arrowheads above and below gene structures serve as the forward and reverse primers [Reproduced from Ali et al., PLoSONE, 2007].

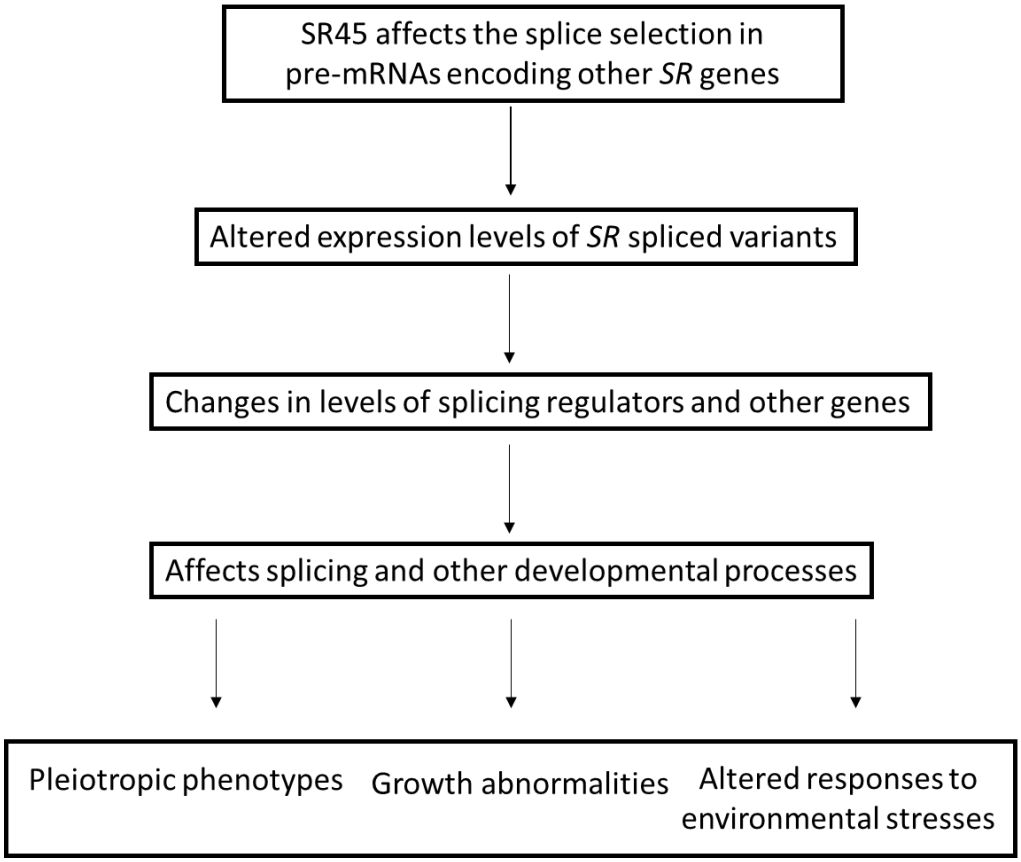


Figure 13. The cascade effect of SR45 splicing regulation of other *SR* genes. SR45 differentially modulates splicing patterns of alternative transcripts expressing other *SR* genes, which then modulate splicing and other RNA processing activities that ultimately control many aspects of the plant growth and development.

2010]. Glucose and ABA are key components for sugar signaling during early growth stages and perhaps even SR45 plays a major role coinciding with these elements. It was shown that *sr45-1* mutant is hypersensitive to glucose and ABA, suggesting that SR45 negatively modulates glucose signaling [Carvalho et al., 2010]. Regulation of degradation of the energy sensor SNF1-related kinase 1 (SnRNK1) by SR45 is important in negatively regulating glucose signaling [Carvalho et al., 2016]. Overall, this study highlighted that SR45 negatively regulates Glc-specific genes involved in the sugar signaling pathway.

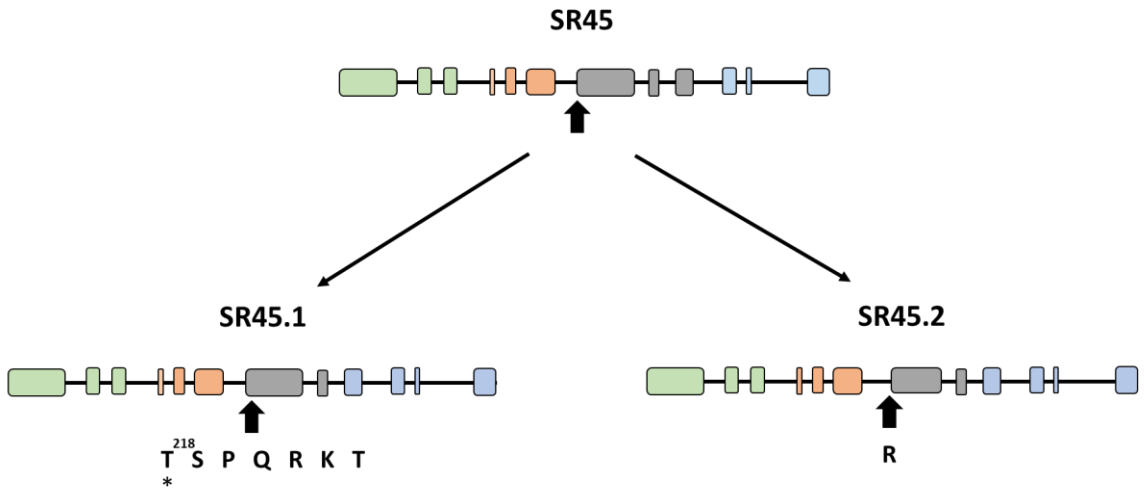
SR45 functions in stress responses

Splicing is highly variable and dependent on developmental cues and abiotic stress responses. Additionally, different splicing patterns in plants have implicated a strong impact on agricultural yield, plant growth and development, and plant responses to environmental conditions. The mechanistic insights of splicing affected by external signals in plants are yet to be fully elucidated. Nonetheless, sufficient evidence has indicated a linkage with the SR45 splicing factor and plant adaptation to multiple stresses. Most of the gene ontology (GO) terms have fallen in the category of stress and hormonal responses, indicating SR45 major functionality is associated with plant adjustment to environmental stress responses [Xing et al., 2015]. About 30% of transcripts of ABA signaling genes were involved with SR45 during the post splicing level and most of these genes encode phosphatases and kinases, suggesting the SR45 role in regulating the expression of crucial ABA signaling genes. This evidence is consistent with RNPS1 role in splicing and post-splicing. When exposed to virulent pathogens, the *sr45-1* mutant displayed a stronger resistance to these pathogens in their initial defense and had a significant enrichment in plant defense genes and salicylic acid signaling genes [Zhang et al., 2017]. This evidence strongly suggests that SR45 may play a role in suppressing the expression

of genes associated with innate immunity, especially in salicylic acid biosynthesis. The loss of SR45 has shown misregulation and different splicing patterns of other SR proteins, especially during abiotic and biotic stress.

SR45 isoforms: SR45.1 and SR45.2

SR45 pre-mRNA itself undergoes alternative splicing, generating two distinct splice variants that encode proteins differing in eight amino acids, a long isoform (SR45.1) and a short isoform (SR45.2). Because of a different 3' splice site selection at the beginning of the seventh exon, SR45.1 has 21 more nucleotides with additional amino acids (TSPQRKTG) as compared to SR45.2 [Palusa et al., 2007] (Figure 14). Two of these residues exclusively found in the SR45.1 alternative spliced region are shown to be major phosphorylation sites. In a study, it was found specifically that T218 requires phosphorylation for SR45.1 to constitutively splice *SR30* transcript and thus promoting flower petal development [Zhang et al., 2014]. Previous work has discovered that these two isoforms are functionally different with some shared common roles. Hypersensitivity to glucose was fully complemented by both spliced isoforms when shortened cotyledon growth and reduced hypocotyl elongation were reverted to wildtype in lines of SR45.1 and SR45.2 expressed in *sr45-1* mutant background [Carvalho et al., 2010]. These results suggest that SR45 isoforms seem to share a biological role in early seedling growth. However, a long-spliced SR45 isoform complemented the flower phenotype and seed production in the mutant, whereas a short-spliced SR45 isoform did not complement either feature [Zhang and Mount, 2009]. Conversely, SR45.2 recovered the root phenotype of the mutant, while SR45.1 displayed no significant effect on its root phenotype. In response to salt stress, it was demonstrated that SR45.1 positively regulates the expression and splicing patterns of stress-responsive genes whereas SR45.2 impaired these expression levels and spawned irregular



* Phosphorylation at Threonine 218 is required for the function in flower petal development

Figure 14. Two major SR45 isoforms. The *SR45* transcript generates two major isoforms due to an alternative acceptor splice site at the beginning of the seventh exon: SR45.1 (long isoform) and SR45.2 (short isoform). SR45.1 sequence differs by eight amino acids (TSPQRKT) compared to SR45.2 (R). Studies have shown that the SR45.1-Threonine218 of the eight distinctive amino acids is a major phosphorylation site that promotes flower petal development. The RS1 domain (green); the RRM domain (orange); the RS2 domain (blue).

splicing events [Albaquami et al., 2019] (Figure 15). This indicated that SR45.1 is required for salt tolerance and restores ion homeostasis because of its ability to successfully rescue the salt stress-sensitive phenotype of the mutant. Whereas SR45.2 had similar phenotypes with stunted roots and fewer emergence of new leaves, indicating that it does not complement salt stress sensitivity phenotype. The molecular strategies for obtaining these isoform phenotypes need further research. Research on SR45 protein's various regulatory functions and its molecular approaches is relatively nascent so much work is needed to understand the role of these splice variants.

Importance of elucidating SR45 isoform functionalities

Alternative splicing contributes to proteome complexity and diversity in multicellular eukaryotes and plays important role in plant development and stress responses. Thus, understanding the role of splicing factors in regulating alternative splicing is not only important to our understanding of plant developmental and stress responses at the post-transcriptional levels but will also open new avenues to develop crops with desired traits and enhance agricultural production. Agricultural issues have been linked to aberrant splicing events [Reddy et al., Plant Cell, 2013; Brown and Staiger Plant Cell 2013] so functional analysis of major splicing factors and their splice isoforms can help us improve seed quality, plant growth, and plant stress responses. Not only that, shedding insights on the composition of the "splicing code" and the key players of plant-specific splicing may allow researchers to predict the type of alternatively spliced products that are generated. Discovering the *in vivo* RNA targets of SR45 protein isoforms will provide mechanistic insights into isoform functions and paves the way for modulating specific isoform expression for crop improvement. With substantial work on deciphering the importance of each isoform, researchers can selectively express more of one

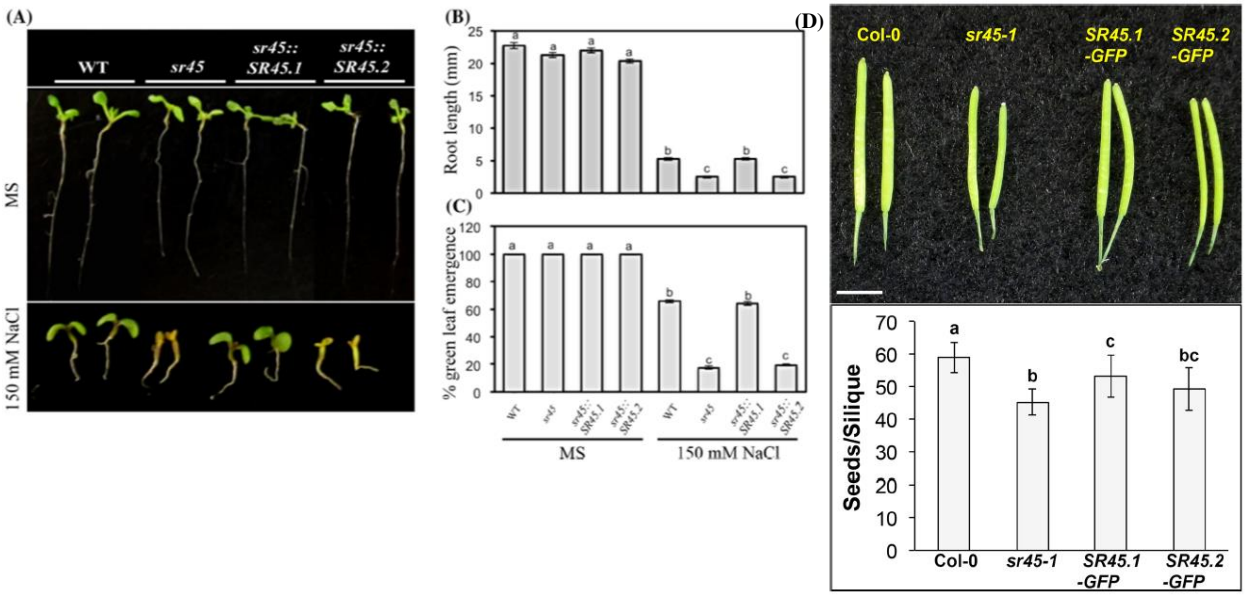


Figure 15. SR45 isoforms have distinct biological functions. A) Images of seedlings that were grown in either MS or MS supplemented with 150 mM NaCl to phenotypically show the responses of SR45 and its isoforms to salt stress at the germination stage. *Sr45* and *sr45.2* complemented lines had similar stunted growth and withered leaves, while *SR45.1* line recovered its phenotype as the wildtype [Reproduced from Albaqami et al., Plant Molecular Biology, 2019]. B and C) The root length and number of new green leaf emerging from each sample were quantified. In both cases, the seeds grown in MS medium had similar results. In medium supplemented with NaCl, *sr45* and *SR45.2* lines had significant inhibition of root growth and reduction in green leaves emerging. *SR45.1* is significantly induced in response to NaCl treatment comparable to wild-type levels D) A comparison of seed number per silique among Col-0 and SR45-GFP complemented lines were shown in the image visually and seeds per silique were quantified from Zhang et al.'s paper. The reduced number of seeds in *sr45* was drastically improved in *SR45.1* (90% of Col-0), but not in *SR45.2* (83%). This study proved that SR45 long isoform promoted seed and flower development. [Zhang et al., BMC genomics, 2017].

isoform that is beneficial to the plant's growth and development. Therefore, researchers can exploit this genetic control to attain desired proteins, as well as reprogram certain aspects in post-transcriptional regulatory pathways for plants so that they can cope with various adverse environmental conditions. Overall, SR45 protein is within this dynamic, interconnected regulatory network of post-transcriptional proteins that are reliant on one another to maintain multiple cellular processes. The purpose of this work is to identify RNA transcripts associated with each SR45 isoform, which can enlighten our knowledge on various mRNA regulations affected and ultimately, elucidates the capabilities of SR45 protein in plants.

Objectives of this study

In this study, I intend to determine the global and distinctive RNA targets of SR45.1 and SR45.2 isoforms in *Arabidopsis thaliana (A. thaliana)* to gain insights into how they regulate different biological processes. Detecting the shared and unique binding regions from each isoform will allow us to investigate how these isoforms perform biologically distinct functions. I will be using a recently developed RNA editing tool, HyperTRIBE [McMahon et al., 2016], to identify the specific RNA targets of these two SR45 isoforms. It is a simple method that is based on the editing of RNA that an RNA binding protein (RBP) binds and it does not necessitate large amounts of starting material, or the tedious work associated with immunoprecipitation assays. Utilizing TRIBE/HyperTRIBE for the detection of *in vivo* RBP targets has not been demonstrated in plants yet. A few papers have been published; most of these used this technique in *Drosophila* and mammalian systems. I will be using both TRIBE and HyperTRIBE in plants in the pursuit of identifying RNA targets of long and short isoforms of SR45. This discovery of the *in vivo* targets of each SR45 isoform can be further analyzed by performing the gene ontology (GO) analysis to gain insights into distinct biological functions of each isoform.

Approaches to identify RNA targets of an RNA binding protein

CLIP (Crosslinking and immunoprecipitation) is considered the gold standard assay to identify RNA-protein interactions. Other derivatives of this RBP target identification tool, such as PAR-CLIP, HITS-CLIP, etc., have been developed using the same concept: protein and RNA are irreversibly UV-crosslinked to attach and stabilize them together as an RNA-protein complex [Vesper and Reddy, 2021]. RNases are used to digest the unbounded RNA, and the bounded complexes are immunoprecipitated. RNA is extracted from the cells and high-throughput sequencing is further performed to analyze the exact RNA-binding protein binding sites [Ule et al., 2005; Huppertz et al., 2014].

Currently, the preferred method for investigating protein-RNA interactions is CLIP as it allows precise identification of RBP binding site on RNA targets. Nevertheless, this method still has disadvantages that could potentially obfuscate its results. CLIP demands a large amount of starting material; hence it is difficult to obtain RNA targets in specific cell types in a tissue. It is almost impossible to purify single cell types from tissues, so CLIP tends to use mixed cell types as starting material [Darnell, 2010; Moore et al., 2014]. Consequently, the use of heterogeneous cells/tissues would make it difficult to assign identified RNA targets to distinctive cell types. Also, random binding sites are likely to create nonspecific complexes and thus present false-positive results. Performing the CLIP procedure is also tedious and has a higher likelihood of errors when executing the cross-linking and immunoprecipitation steps. Likewise, crosslinking efficiency varies depending on the system being used; in plants, typically the efficiency is reduced compared to mammals [Darnell, 2010; Vesper and Reddy, 2021]. This reduced

efficiency indicates the absence of inefficient crosslinking between the RNA and protein, thus losing true RNA targets. Another problem with CLIP is its requirement of a specific antibody for the RNA-binding protein (RBP), which can be costly or difficult to obtain. If a high-affinity antibody has not been utilized in CLIP, non-specific complexes can be precipitated, resulting in a high percentage of false positives. A new assay to overcome these limitations with CLIP is highly valued, especially when interested in RBP target sites in a cell or tissue-specific manner.

TRIBE: A novel method to identify RNA targets of an RBP

The TRIBE (Targets of RNA-binding Proteins Identified by Editing) method was developed in Michael Rosbash's lab from Brandeis University as a novel alternative to identify RNAs that bind to an RBP of interest [McMahon et al., 2016]. TRIBE uses the catalytic domain of RNA-editing enzyme ADAR's (adenosine deaminase acting on RNA) from *Drosophila* to identify RNAs that bind any RBP. The ADAR is comprised of two domains: a double-stranded RNA-binding domain (dsRBM), which is responsible for binding to double-stranded RNA, and a catalytic domain (ADARcd), that hydrolytically deaminates adenosine to inosine, which is read as guanosine by ribosomes and reverse transcriptase [Keegan et al., 2004, Montiel-Gonzalez et al., 2013; Vogel et al., 2014] (Figure 16). In this method, the catalytic domain that deaminates adenine is fused to an RBP of interest to generate a fusion protein, which edits only those RNAs that the RBP binds to. The editing specificity is determined by the RNA recognition motif of the RBP of interest and the deaminase converts adenosine into inosine that is in the proximity of the target sites. Expression of the fusion protein in the cell of interest followed by high-throughput RNA sequencing will display the edited-RNA transcripts and thus allowing us to identify the RNA targets of the RBP of interest. Another recently developed method to identify any RBP

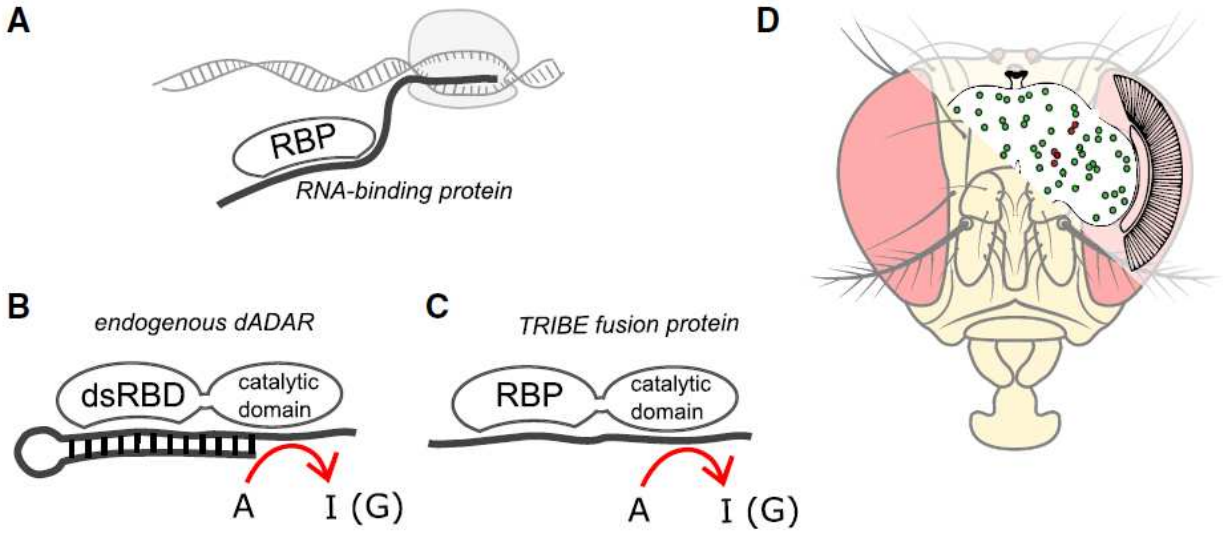


Figure 16. TRIBE method. A) A schematic diagram of the RNA binding protein (RBP) binding to its specific target mRNA transcripts. B) The endogenous *Drosophila* ADAR (dADAR) is composed of two domains: double-stranded RNA binding domain, which will mediate the RNA binding specificity, and a catalytic deaminase enzyme domain, which catalytically deaminases nearby adenine to inosine. The inosine is recognized as guanosine by transcriptase due to its structural similarity. C) Rosbash's group replaced the dsRBD with the protein of interest and fused the domain with the catalytic enzyme to create the TRIBE fusion protein. With this chimeric protein, the RNA recognition is dictated by the RBP followed by the catalytic ADAR domain irreversibly editing the target RNA transcripts. D) The TRIBE method can identify the target sites of the RBP in rare cells *in vivo* by expressing the fusion protein in a cell- or tissue-specific manner using cell/tissue-specific promoters. Rosbash's group was able to use the TRIBE method to find RNA targets of an RBP in *Drosophila* neuronal core circadian pacemaker neuron cells (indicated with red dots] and dopaminergic neurons (indicated with green dots]) [McMahon et al., Cell, 2016].

targets also depends on the attachment of the RBP with an enzyme that tags bound RNA [Lapointe et al., 2015]. In this approach, the chimeric protein contains the RBP of interest that dictates the mRNA binding and C. elegans poly(U) polymerase, PUP-2, that tags the associated RBP-RNA transcript with a tail of uridine at the 3' end. This method also avoids the crosslinking and protein purification steps of CLIP and bioinformatic analyses are needed to identify the binding regions after RNA-sequencing. However, there are drawbacks to RNA tagging (Figure 17). It does not provide the RBP binding sequences as it identifies only the labeled RNA transcript. Because of the extended U- tail, the cell can unintentionally activate the U-tail-dependent degradation pathway and reduce the tagged RNA transcript results. Overall, TRIBE seems to be the superior antibody-independent approach that is simple and accurate in determining the RNA binding regions of any RBP. Previous work of TRIBE has been performed only in *Drosophila* and mammals. Whether this technique is widely applicable to all model systems is yet to be validated [Jin et al., 2020].

Examining *in vivo* RBP target sites by TRIBE is an ideal strategy especially in identifying cell-specific RNA targets of an RBP. Rosbash et al. successfully conducted TRIBE experiments to identify RNA targets with different RBPs, including Hrp48, from specific neuron subtypes (Mcmahon et al., 2016). Not only were the RBP-TRIBE constructs expressed in these specific rare cells, but they were able to extract RNA from a small set of neurons, about 150 fly neurons. The majority of the Hrp48-TRIBE edit sites were enriched in the 3' UTR region, which was expected for their role in translation and further highlighted how the TRIBE method accurately recognized the target mRNA transcripts. Compared to CLIP, TRIBE does not rely on complicated and tedious tools and the availability of a highly-affinity antibody, suggesting this RBP-target method is substantially cost-effective (Figure 18). The TRIBE method involves the

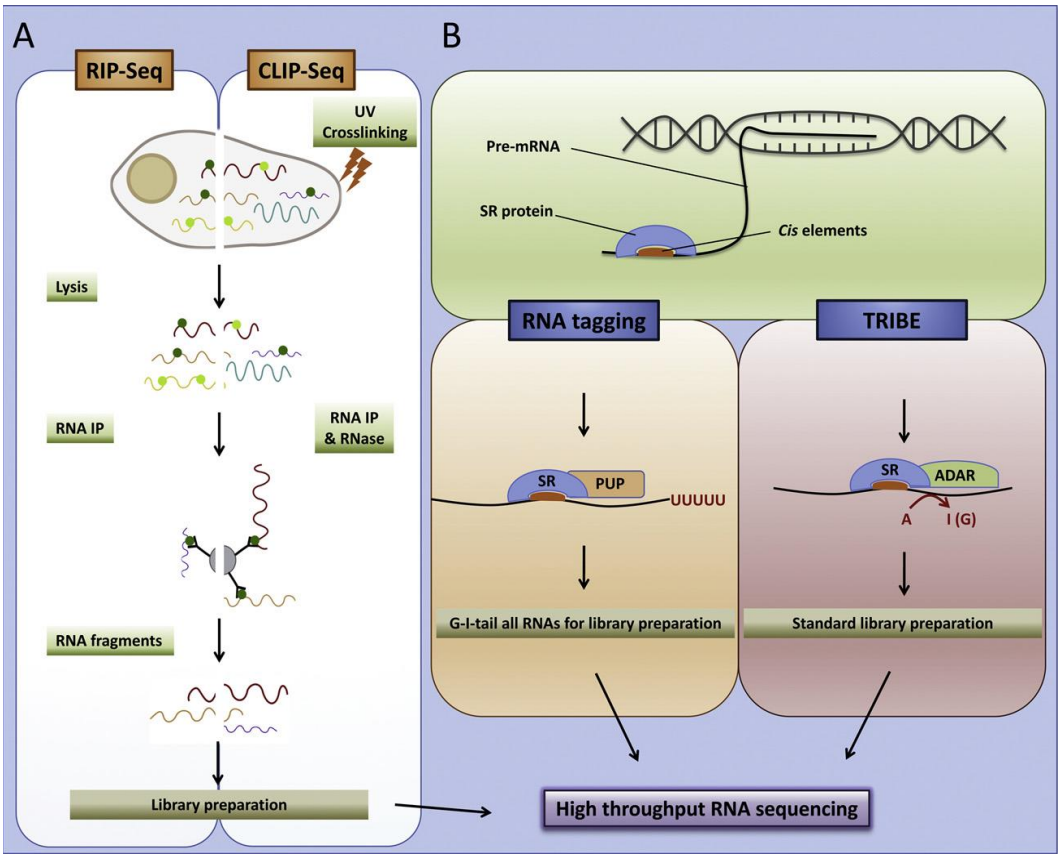


Figure 17. RIP-Seq, CLIP-Seq, RNA-tagging, and TRIBE. A schematic representation of the procedure to identify the *in vivo* RNA targets of the RNA binding protein (RBP). A) In RIP-seq, RNA and the RBP are crosslinked with formaldehyde. The cell is lysed, and the RBP-RNA complex is immunoprecipitated by bead-bound antibodies against the RBP. The captured RNA fragments are utilized for cDNA libraries followed by RNA sequencing. In CLIP-seq, the RNA fragments bound to an RBP are UV crosslinked. After cell lysis, the RNA-protein complexes are immunoprecipitated with a bead-bound antibody and are exposed to Rnases to degrade any unbound RNA fragments. The remaining bound RNA fragments are used to make a cDNA library, which is then subjected to high throughput sequencing. B) RNA tagging and TRIBE are both methods that do not require a specific antibody or immunoprecipitation of the RNA-protein complexes. In RNA tagging, the RBP is fused with a poly(U) polymerase (PUP). Expressing this chimeric protein in the cells of interest will lead to the addition of a uridine tail on the RNA transcript that the RBP binds. The uridine tail is tagged with guanosine (G) and inosine (I) for cDNA library preparation followed by high-throughput sequencing. The TRIBE method relies on the expression of a chimeric protein that consists of the RBP of interest and a deaminase enzyme domain (ADARcd). The RBP is responsible for the binding specificity and the ADARcd catalyzes the adenosine (A)-to-inosine (I) deamination, which is recognized as guanosine by reverse transcriptase. After RNA sequencing, the identification of A-to-G mutations would reveal the potential RBP binding targets [Morton et al., Plant Science, 2019].

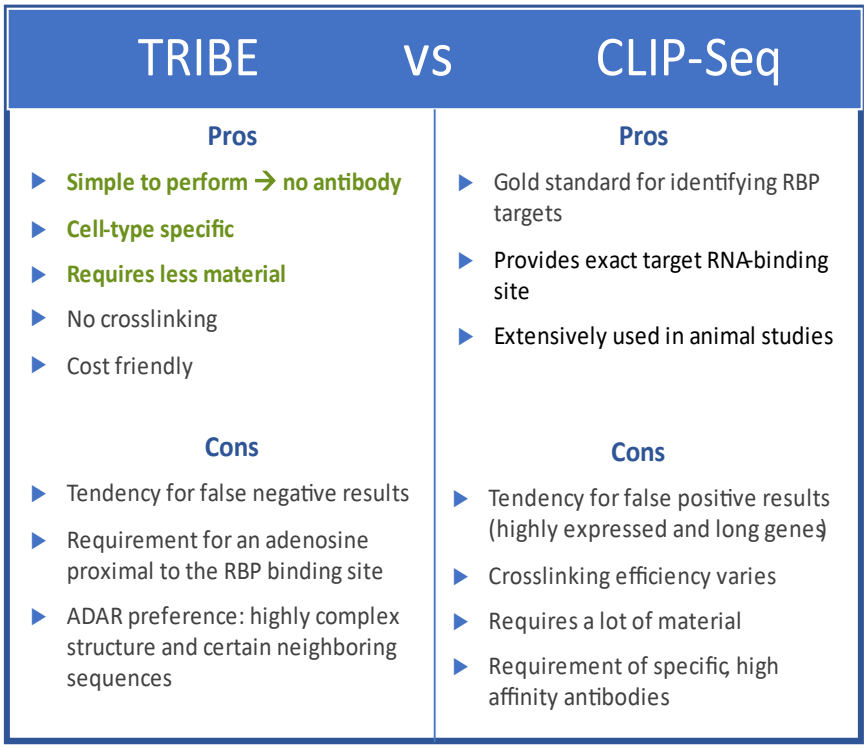


Figure 18. Comparisons between TRIBE and CLIP-Seq. Pro and cons of TRIBE and CLIP-Seq methods for the identification of RNA-Protein interactions.

expression of the TRIBE construct with the fusion protein of the RBP and ADARcd in cells or tissues of interest. Once lines are expressing the TRIBE construct, these lines are prepared for RNA sequencing. The RNA-seq data are then analyzed computationally to identify the edited (A to I (G) mutations) bases in RNAs. When compared to CLIP, the current competitor, TRIBE is a simpler, inexpensive procedure with the capability to reduce false-positives in a cell-type-specific manner. In a recent study, *trans*-editing using TRIBE has been observed from interchromosomal contacts from an RBP, which imitates how transcriptional regulatory proteins associate with nascent transcript and create a transcriptional hub [Biswas et al., 2020]. This finding broadens TRIBE impact from identifying proximal high-confidence RBP targets to elucidating spatial organization of nuclear RNA regulatory proteins. There is also evidence that TRIBE displayed a lower background noise compared to CLIP when they found that the majority of CLIP off-targets are from a poor-quality antibody capturing nonspecific interaction and substantially fewer TRIBE off-target peaks compared to CLIP were from background editing of ADARcd [McMahon et al., 2016; Biswas et al., 2020]

HyperTRIBE enhanced editing efficiency

This unique RBP target identification method, however, has its own limitations and can be further improved. Utilizing a conservative parameter (>10% editing and 10 edits per reads) to filter out the precise TRIBE-edited sites, Hrp48-TRIBE was only able to locate about 25% of the edit sites found by Hrp48-CLIP. This finding indicates that TRIBE produced a high rate of false-negative results [Mcmahon et al., 2016]. Presumably, the low editing efficiency of TRIBE could be due to the ADARcd preference for a specific sequence (typically uridine on the 5' end and guanosine on the 3' end; UAG sequence) and an RNA-double structure surrounding the targeted adenosine. To circumvent these issues, Rosbash's lab used a hyperactive ADARcd with a single

amino acid change (E488Q), which significantly enhanced the ADARcd editing efficiency as well as reduced the sequence bias [Xu et al., 2018]. Comparing edit sites of Hrp48-TRIBE, Hrp48-CLIP, and Hrp48-HyperTRIBE, the results from Hrp48-HyperTRIBE had a larger set of sites overlapping with Hrp48-CLIP exact binding sites than Hrp48-TRIBE marked edits. A substantial amount of edit sites were found in both HyperTRIBE and TRIBE, but some of the sites did not meet TRIBE required parameters of editing. These below thresholds edit sites were counted and found in 30% of Hrp48-HyperTRIBE editing so this verified that HyperTRIBE was able to minimize the false-negative results with TRIBE editing. Even expressing the HyperTRIBE/TRIBE in *Drosophila* neuron cells, comprising of a small quantity, HyperTRIBE had a significant increase in the editing of about 11-fold compared to TRIBE editing events, suggesting that HyperTRIBE editing percentage is much higher than the TRIBE method [Xu et al., 2018] (Figure 19). Moreover, HyperTRIBE had multiple adenines converted within the binding region of their chimeric protein, where results showed HyperTRIBE had an average of three edits per gene and TRIBE had about one edit per gene. The editing preference originally found in TRIBE editing has been less frequent in HyperTRIBE editing events. The bias for UAG neighboring elements near the edited adenine has strikingly minimized in HyperTRIBE target sites and has more leniency on editing with less structural requirements near the adenine of the RBP site. The mapping of RBP sites in specific cell types and conditions from the HyperTRIBE approach has been validated in the mammalian system, where HyperTRIBE fused with Mushashi-2, a protein responsible for cell fate determination, in human hematopoietic stem cells and leukemia stem cells [Nguyen et al., 2020]. This further supports that HyperTRIBE can confidently distinguish cell-specific RNA targets in specific cell types including rare cell types.

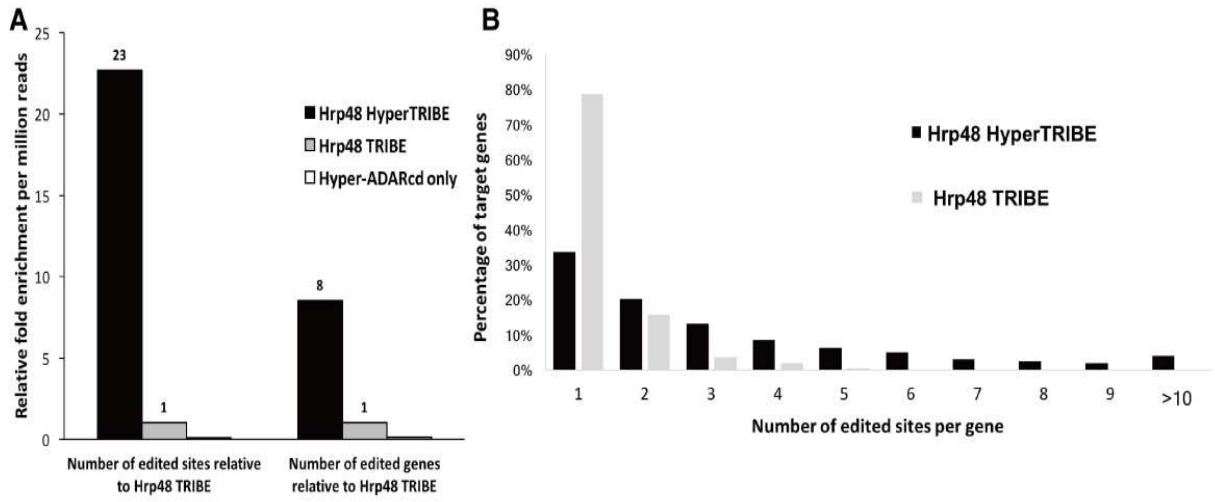


Figure 19. HyperTRIBE enhanced editing efficiency. A) Xu et al. normalized the number of genes and edited sites to the sequencing depth of each sample before the relative fold change of each sample was calculated by comparing Hyper-ADARcd as the control. In HyperTRIBE-expressing cells, a dramatic increase in editing was found in both target genes and edited sites compared to the TRIBE lines. More specifically, HyperTRIBE replicates reported 10,689 common edit coordinates, TRIBE replicates had 291, and Hyper-ADARcd had 11. B) The number of edited sites per gene was observed in Hrp48 HyperTRIBE and Hrp48 TRIBE lines. A larger proportion of target genes contain multiple editing sites (one to more than 10 edited sites) in the HyperTRIBE lines whereas the target genes edited by the TRIBE lines mostly harbored 1 edited site [Xu et al., RNA, 2018].

High-throughput RNA sequencing applications

High throughput sequencing is done to identify the A-to-I conversion of the RBP target transcript and there are many sequencing platforms that can be applied for this step. The widely used short-read sequencing platform is Illumina. This next-generation sequencing has a 99.9% base calling accuracy and is typically utilized for the majority of TRIBE and HyperTRIBE current data and interpretations [Pfeiffer, 2018]. However, Illumina sequencing utilizes a reverse transcriptase when an RNA template is used as the source which can lose information during the process of transcribing RNA into cDNA. And reverse transcriptase recognizes inosine as guanosine which can be problematic when the transcriptase is not able to identify all inosines accurately.

Another sequencing approach is nanopore sequencing, which provides long read lengths and is capable of doing direct RNA sequencing [Zhao et al., 2019; Gao et al., 2021]. However, with nanopore sequencing, currently, there are no base callers that can accurately identify inosine bases. For my project, I will be utilizing Illumina sequencing reads to assess SR45 splice isoforms RNA targets sites, which will provide new mechanistic insights into distinct biological functions of these isoforms in plant growth, development, and stress responses. Incorporating the TRIBE and HyperTRIBE techniques in this project is not only crucial in observing its editing efficiency in plants but could potentially be recognized as the preferable RNA-protein identification assay for plant studies.

MATERIALS AND METHODS

TRIBE constructs preparation

Drosophila ADAR catalytic domain (*dADARcd*) sequence along with linker sequence was amplified from pMT_A_Blasticidin _HRP48_ADAR(1)-V5 [MacMahon et al., 2016] using Hot start Primestar polymerase (Takara) with primers containing *Not*I in forward and *Bam*HI *Sph*I in reverse primer, respectively. The amplicon was cloned in pGEM-T Easy vector (Promega). For cloning of *SR45* isoforms (short and long isoforms), cDNA was prepared from the total RNA isolated from the transgenic lines of Arabidopsis (expressing each isoform separately) using Superscript III first-strand preparation Kit (Invitrogen). Individual full-length *SR45* isoforms were amplified using Hot Start Primestar polymerase (Takara) with oligos containing *Sac* I, *Asc* I and *Not* I restriction sites in forward and reverse primer. Each full-length *SR45* isoform amplicon was digested with *Not* I and *Sac* I and subsequently ligated into the cloning vector pGEM-T Easy as per the manufacturer's instructions. Clones bearing the amplicon were sequenced using T7, SP6 promoter, and gene-specific primers, to rule out any possibility of PCR-induced errors. *dmADARcd* along with linker sequence was cloned downstream of each isoform of *SR45* by ligating predigested (with *Not* I and *Sph* I) *dmADARcd* fragment and pGEM-T Easy *SR45* (isoforms) vector (Figure 20).

HYPER TRIBE constructs preparation

To make a hyperactive version of the TRIBE, pGEM-T Easy - *dmADARcd* was used. Based on a previous study [Xu et al., 2017] a mutation of E488Q was created using Q5 site-directed mutagenesis kit (NEB) as per the manufacturer's instructions. Primers Q5SDM-F 5'-

CGAGTCCGGTCAGGGGACGAT-5'and Q5SDM-R 5'-ATTTTGGTGCGCAGCTGG-3'were used for the creation of the mutation. The primers for mutagenesis were designed using NEB base changer web tools (NEB). The rest of the cloning process is similar to that described above for the TRIBE constructs.

Plant transformation vector construction

The transgenic lines expressing the fusion protein (*SR45* isoform – *dmADARcd*) were generated in *SR45-1* background by transforming them with *Agrobacterium tumefaciens* (GV3101) containing pFGC5941: *SR45 isoform - dmADARcd* vector. The *SR45 isoform - dmADARcd* was cloned into pFGC5941as *Asc*I and *Bam*HI fragments downstream of the CaMV35S promoter. Individual lines expressing the *dmADARcd* alone or *SR45 isoforms – dmADARcd* were selected on ½ MS media containing 7.5 mg/L BASTA (glufosinate ammonium) with 1% sucrose and 0.5g/L of MES. BASTA-resistant seedlings were transferred to soil and grown in a growth chamber at 20^{o} C under day-neutral conditions. Homozygous transformed lines of independent TRIBE and HYPERTRIBE lines generated after repeated selfing followed by selection on 5mg/L BASTA. Homozygous lines of all genotypes were used for all experiments.

Plant Materials and Growth Conditions

Wild-type (WT) lines for this experiment were *Arabidopsis thaliana* ecotype Columbia-0 (Col-0). Homozygous mutant (*sr45-1*) complemented (*SR45-HYPERTRIBE SHORT*; *SR45-HYPERTRIBE LONG; SR45-TRIBE SHORT*; *SR45-TRIBE LONG*), and lines expressing *TRIBE* and *HYPERTRIBE* catalytic domain alone were generated. Seeds from each line were harvested

from plants grown at 22º C, 120 mmol/m2/s white fluorescent light, long-day16 hr/8 hr light/dark photoperiod.

RT-PCR analysis of expression of TRIBE and HyperTRIBE fusions in transgenic lines

Twenty-day-old seedlings of wild-type, *sr45-1* mutant, *sr45-1* complemented with either *HyperTRIBE* and *TRIBE* fusions as well as lines expressing *TRIBE* and *HyperTRIBE* catalytic domain only were utilized for the RT-PCR expression analysis to further validate each genotype. These seedlings were collected, frozen in liquid nitrogen, and stored in the -80 fridge. The frozen tissues were ground into fine powder in 2 mL microcentrifuge tubes with metal ball bearings using the TissueLyser II. Extraction of total RNA was done using a Trizol RNA isolation protocol and was treated with DNase to digest any genomic DNA. Using the Superscript III reverse transcriptase (Invitrogen), total RNA (1-2 μg) was utilized for cDNA synthesis and cDNA (about 0.5 μL) was used as the template for each reaction. Amplification of PCR products was done using specific primers for each *SR45* isoform and dADARcd alone and fused with each SR45 domain. To confirm each genotype and its expression levels, the PCR products for each sample were visualized on a 1% agarose gel. CYCOPHILIN was used for the loading control.

RNA isolation

We used the RNeasy Plant Mini kit (Qiagen, USA#217004) to prepare RNA for RNA-Seq libraries. About 100 mg of plant tissue of each sample was ground thoroughly by the TissueLyser II. The ground tissues were homogenized in Buffer RLT and vortexed. The lysate was then transferred to a QIA shredder spin column, centrifuged for 2 minutes, and the flow-through supernatant was transferred to a new microcentrifuge tube. A 0.5 volume of ethanol (96-100%) was added to the lysate and then placed into an RNeasy Mini spin column, and

centrifuged for 15 s at 8000 g. This centrifugation step was repeated three more times with different buffers that were provided by the RNeasy Plant Mini kit. At the last centrifugation for 2 minutes, the RNeasy spin column is transferred to a new collection tube, and 30-50 μL of RNase-free water was added to the column membrane. The tube is then centrifuged for 1 minute at 8000 g to collect the RNA. Ribosomal RNA was removed using a Ribozero Plant kit and the sequencing libraries were prepared from rRNA-depleted samples using TruSeq stranded RNA-seq kit (Illumina) as per manufacturer instructions and paired-end sequencing of the library was done at LC Sciences (Houston, USA). All RNA-seq reads will be deposited at NCBI in the GenBank sequence read archive (SRA) under the accession number ***.

Library Construction and Sequencing

Poly (A) RNA sequencing was done at LC Sciences. RNA libraries were made by utilizing Illumina's TruSeq-stranded-mRNA kit. Poly (A) mRNA was captured using oligo-(dT) magnetic beads with two purification washes. After poly(A)-containing mRNAs were purified, the temperature was spiked and was washed with a divalent cation buffer to fragment the mRNAs. The Agilent Technologies 2100 Bioanalyzer High sensitivity DNA Chip was utilized to assess the overall quality and to quantify the sequencing libraries. Paired-end sequencing of all libraries was done on Illumina's NovaSeq 6000 sequencing platform. We obtained 46-52 million reads per sample and the average read length was approximately 140 nucleotides. LC Sciences cleaned the reads by Cutadapt and in-house Perl scripts to remove any adapter contamination or low-quality bases and checked the sequence quality with FastQC. These reads were then mapped to the Arabidopsis reference genome using HISAT2 and were further assembled with StringTie. After assembling, the final format of the output is the BAM file which we used for the part of the

TRIBE and HyperTRIBE analysis protocol. I also did my own read trimming and assembling by utilizing different quality control tools.

Trimming and alignment of sequence libraries

To obtain high-quality reads, Trimmomatics was performed on all 24 samples (TRIBE and HyperTRIBE lines) to remove low-quality reads and to trim off low-quality bases from both ends of the sequencing read due to potential error from random hexamer mispriming (6 nucleotides of the reads were removed). Reads with an average quality score of 25 or more and a minimum length of 19 were retained. To compare the TRIBE and HyperTRIBE sequences with the wtRNA reads, the next step was to align the quality reads onto the reference genome by STAR (Spliced Transcripts Alignment to a Reference). This computational tool mapped each sample to the wildtype Arabidopsis transcriptome (TAIR 10) and created a finished alignment file called SAM (Sequence Alignment/Map) files. To build indices for STAR, we used the Arabidopsis_thaliana.TAIR10.dna.toplevel.fa for the genome FASTA files and TAIR10_GFF3_genes.gff for the genome annotated transcripts.

SAMtools were then executed to remove any low-quality alignments followed by a conversion of the SAM files to a more compressed version called BAM (Binary Alignment/Map). The number of mismatches was limited to at most 7% of the mapped read length and a minimum of 16 bases had to be mapped per read. The maximum number of multiple alignments allowed for a read is one, so if that number is exceeded, it will be considered unmapped. Using the SAMtools to eliminate low-quality alignments, we skipped any alignments with a MAPQ value smaller than 10. The final output is SAM files which we used SAMtools to convert SAM into a BAM file and sort the BAM files before using Picard. PCR duplicates were removed using the Picard tool by identifying reads with the exact mapped location and

preserving the reads with the highest quality while deleting the other reads. Most of the bioinformatic tasks were done on the Summit High-Performance Computing (HPC) system, a collaboration with Colorado State University and UC Boulder.

Loading of alignments to MySQL

The alignment files in SAM format were modified into a matrix layout, in which the file would consist of the counts of each type of nucleotide found at each position on the aligned reads in the genome. Once files are converted in a matrix form, these files are deposited into a MySQL database. This database allows us to do a pairwise comparison between each library with the wtRNA library. This comparison identifies any nucleotide with a conversion of adenosine (from wtRNA library) to guanosine (from sample library) and labels them as unique editing sites into a recorded list. When uploading each sample's data, the script requires the a) SAM file name, b) Mysql table name, c) experiment name, and d) integer for the replicate. The combinations of all these variables should be distinctive to each library as a way to easily retrieve the information from the database when called upon.

Identifying unique RNA editing sites

A nucleotide coordinate is labeled as a unique editing site by using the wild-type DNA nucleotide frequency as the reference against TRIBE and HyperTRIBE RNA nucleotide frequencies. By calling out an RNA library by its Mysql table name and other variables, we can compare every position in that transcriptome and compare that data with the genomic DNA nucleotide frequency to determine whether that site had a conversion from A to G. To lower background noises, the unique editing-site list was filtered with criteria where each editing site has at least a specified percentage of editing and/or has a minimum number of reads. We ran our

RNA libraries through three different thresholds that we found were the most suitable for analyzing our data: a) at least 10% editing, b) at least 5% editing, and c) at least 1% editing. Perl scripts were utilized to adjust these parameters. Executing these Perl scripts generated a bedgraph file (the final output file) that summarizes the transcripts targeted by TRIBE/HyperTRIBE editing and its editing results such as the number of editing sites discovered on that transcript and the editing frequency.

Post-processing and reviewing list of editing sites

We used a python script or https://www.biovenn.nl/ (BioVenn) to optimize the bedgraph files into a high confidence list of TRIBE/HyperTRIBE editing sites by detecting edit sites that overlapped in all triplicates. Edit sites found in TRIBE/HyperTRIBE catalytic domain alone were subtracted from each library by bedtools as an approach to remove any nonspecific editing or single nucleotide polymorphism incorporated in its reads. IGB (Integrated Genome Browser) was used to visualize the editing sites and further validate the presence of the conversion of adenine to guanine. We used gene IDs from the transcript that the edit coordinate was located for the gene ontology (GO) analysis. GO terms were determined for these genes using http://geneontology.org/. To identify targets of SR45 that were positioned on spliced isoforms, Excel and Biovenn were used to process and overlap the results.

To compare expression levels of SR45 throughout each sample, a Deseq2 analysis was done using the aligned SAM files that were generated in the TRIBE/HyperTRIBE analysis protocol as the input. FeatureCounts, a bioinformatic tool to tabulate the number of mapped reads to genomic features, were utilized to output a count matrix on each sample and this is done by assigning the Arabidopsis TAIR10 annotation GTF file as the reference genome. The count matrix is downloaded to RStudio where further downstream analysis is done with the

Bioconductor packages (3.12) which contains tools to analyze differential gene expression based on the negative binomial distribution. Some installed libraries that were performed for this analysis were DeSeq2 (1.20.0), corrplot (0.84), RColorBrewer (1.1.2), and apeglm (3.12). A log fold change shrinkage (0.5) was executed for the visualization of the differential expression levels of SR45.

RESULTS

Data analysis pipeline for the identification of RNA editing sites of *Drosophila* Hrp48 protein

The TRIBE/HyperTRIBE method is a relatively new technique, and there is currently limited information on the utility of this method in different systems. Currently, there is no published data on the TRIBE/HyperTRIBE method in plants, and the published pipeline used for data analysis has only been tested on flies and mammals. Hence, I decided to first reproduce the published *Drosophila* results to familiarized myself with the pipeline. This allowed me to execute each step of the pipeline and understand the parameters used in each step. Once I familiarize myself with the TRIBE/HyperTRIBE pipeline, we could apply this to our plant datasets and test results with different parameters. My study to identify the *in vivo* targets of each SR45 isoform is aimed at testing the utility of this novel TRIBE/HyperTRIBE method and data analysis pipeline in plants.

The bioinformatic pipeline that was used for their HyperTRIBE data from *D. melanogaster* was broken down into four major steps: quality control of reads, mapping of reads to the reference genome, uploading of alignments to Mysql, and identifying the RNA target sites (Figure 20). The RNA-Seq libraries from a HyperTRIBE construct were initially filtered by Trimmomatics to remove any low-quality bases and reads to achieve high-quality reads. Six nucleotides from both ends of the reads were also removed to eliminate potential sequencing errors. Picard was another quality control tool used to discard PCR duplicates as an attempt to prevent any bias when calculating the editing percentage of the HyperTRIBE data (Figure 21).

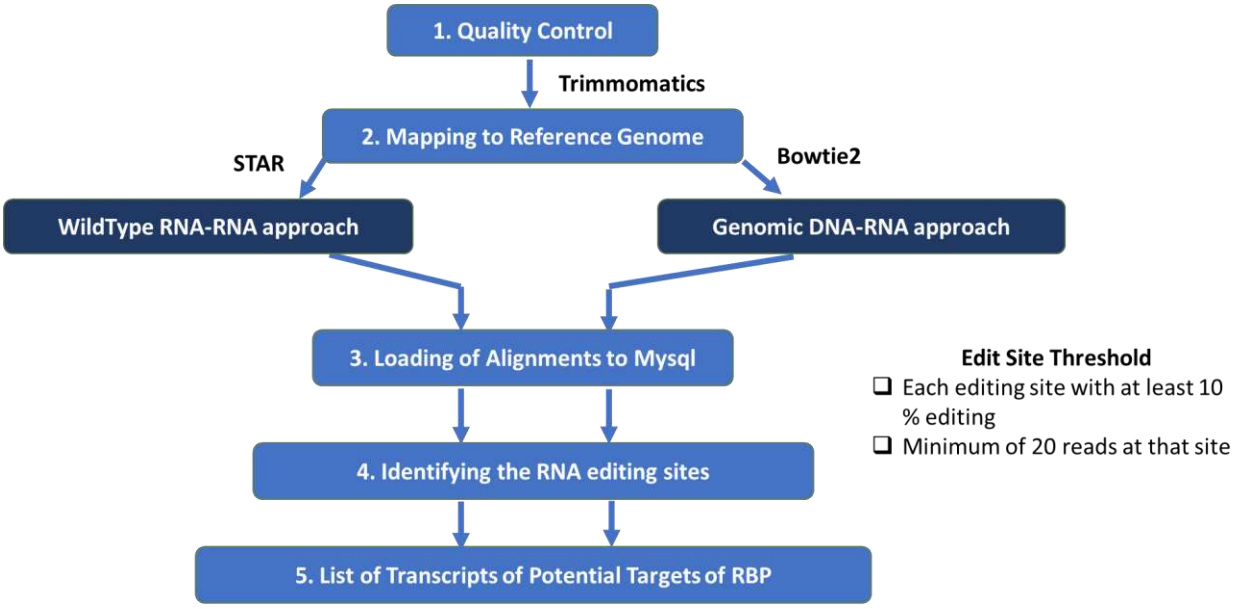


Figure 20. Analysis pipeline for *D. melanogaster* data: Identification of RNA editing sites. The bioinformatic procedures used for the mapped reads of their sample to identify the RBP targets of their protein of interest. Xu et al. [Xu et al., 2018] used two different alignment approaches to compare their HyperTRIBE RNA reads with a control genomic DNA sequence (genomic DNA-RNA approach) or a wild-type transcriptome (wild-type RNA-RNA approach) from the same genetic background. A filtered threshold was implemented to remove any potential single nucleotide polymorphisms and background editing.

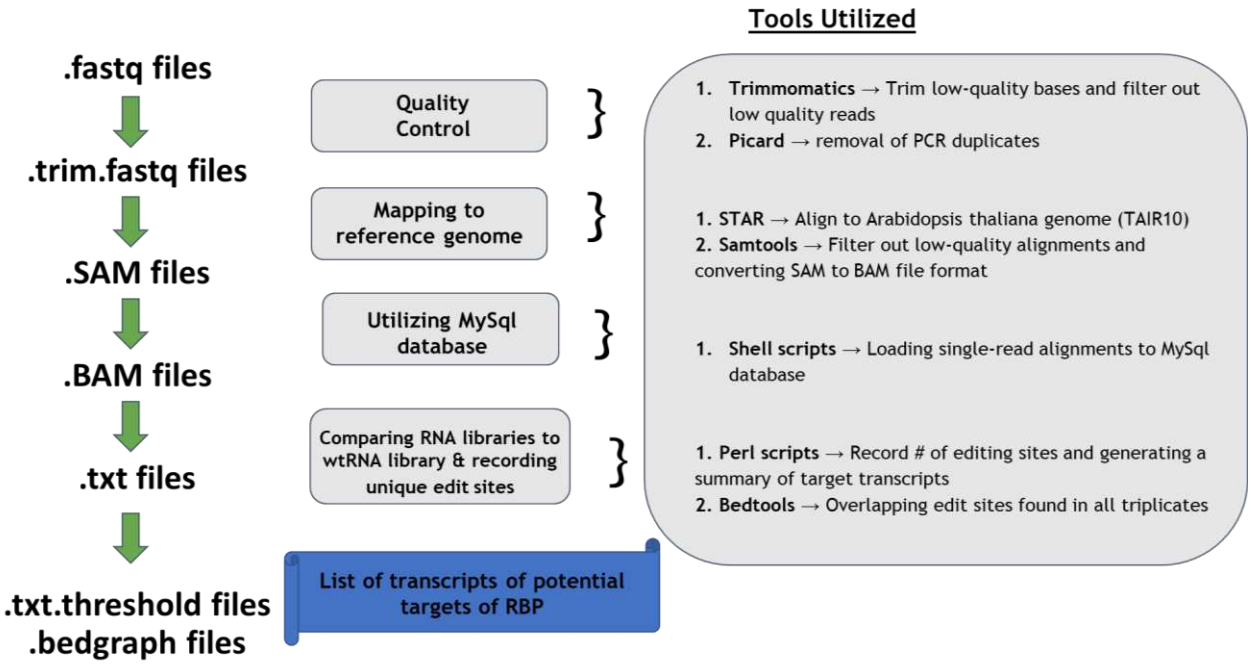


Figure 21. Bioinformatic tools and output files used for the TRIBE/HyperTRIBE analysis pipeline. A detailed description of the type of bioinformatic tools that were utilized and the various output files that were generated for each step of the TRIBE/HyperTRIBE workflow.

The next step is to align reads from each sample to the reference genome; each sample was compared with a control reference genome or a wild-type transcriptome (wtRNA) from the same genetic background [Xu et al., 2018]. For each position of the transcriptome, either the genomic DNA or wtRNA nucleotide frequency was used as a reference for the pairwise comparison against the HyperTRIBE RNA-library nucleotide frequency to detect editing sites. The results from each alignment were cross-referenced with each other to obtain a high-confidence list of editing sites. For the genomic DNA-RNA alignment approach, reads from each sample were aligned to the *Drosophila* transcriptome by STAR or to its genome by Bowtie2.

These alignment files were subjected to final quality control by STAR/Bowtie2 to remove low-quality alignments and minimize any background noise. These alignment files were then converted into a SAM format and uploaded into a Mysql database, which will organize the dataset into a matrix and record how frequently a nucleotide would show up at a certain position in the genome. Then, Perl scripts were used to call out the nucleotide position that has a conversion from A to G, which were then used to calculate the average editing percentage that occurred on that nucleotide coordinate. These edit coordinates were then filtered and compiled to create a list of transcripts that are potential targets of the RBP. Any background editing or single nucleotide morphisms that may have occurred in the control libraries were removed. A threshold of at least 10% editing and a minimum of 20 reads at a site was implemented to fully consider a nucleotide as an edit coordinate. Only high confidence editing sites that are found in all replicates were retained.

Determining the potential target transcripts of *Drosophila* Hrp48 protein

The sample dataset of five sequencing libraries was provided with the HyperTRIBE *Drosophila* protocol: i) S2 Genomic DNA; (ii) S2 WT mRNA; (iii) Hrp48 HyperTRIBE

Replicate 1; (iv) Hrp48 HyperTRIBE Replicate 2; (v) HyperADARcd alone. These files were used to reproduce the HyperTRIBE computational analysis workflow to obtain the RNA targets of *Drosophila* hnRNP protein, Hrp48. We obtained the *Drosophila* (dm6) reference genome sequence in FASTA format and annotation files from the UCSC Genome Browser. Using their *Drosophila* dataset and their editing criterion, we overlapped the final list of edit coordinates found in the gDNA-RNA approach with the list of edit coordinates in the wtRNA-RNA approach to get the precise set of edit sites (Figure 22). In their analysis, the majority of the HyperTRIBE Hrp48 sites from the gDNA-RNA approach were also detected from the wtRNA-RNA approach (9773 edit coordinates) with some non-overlapped sites (1067 unique sites from gDNA-RNA and 516 unique sites from wtRNA-RNA). The remaining non-overlapped editing sites may be from the lack of sequence coverage or single nucleotide polymorphisms. My analysis successfully overlapped datasets using both approaches and obtained all target edit sites that were found from their published data (See Figure 22, left panel).

The editing efficiency of *Drosophila* Hrp48 protein

To address the HyperTRIBE editing efficiency, we investigated the editing frequencies of the sites identified from both gDNA-RNA and wtRNA-RNA approaches (Figure 23). We calculated the number of target genes and the average quantity of edit sites that were found from each target gene. The number of editing sites and genes reported are found in both replicates. As Xu et al. reported [Xu et al., 2018], a larger percentage (about 67%) of the target genes are marked at multiple sites by the HyperTRIBE catalytic domain of the fusion protein. A little over 1000 HyperTRIBE-identified genes had only one edit site detected in the two approaches. Candidate genes with multiple editing events may be due to their protein of interest binding

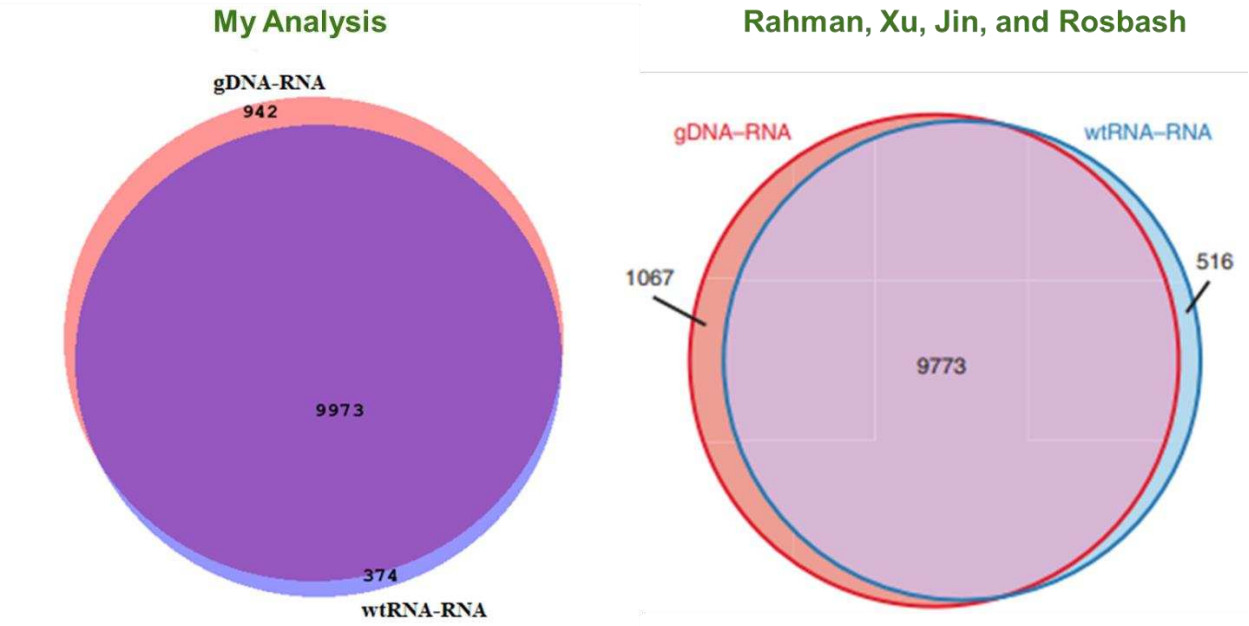
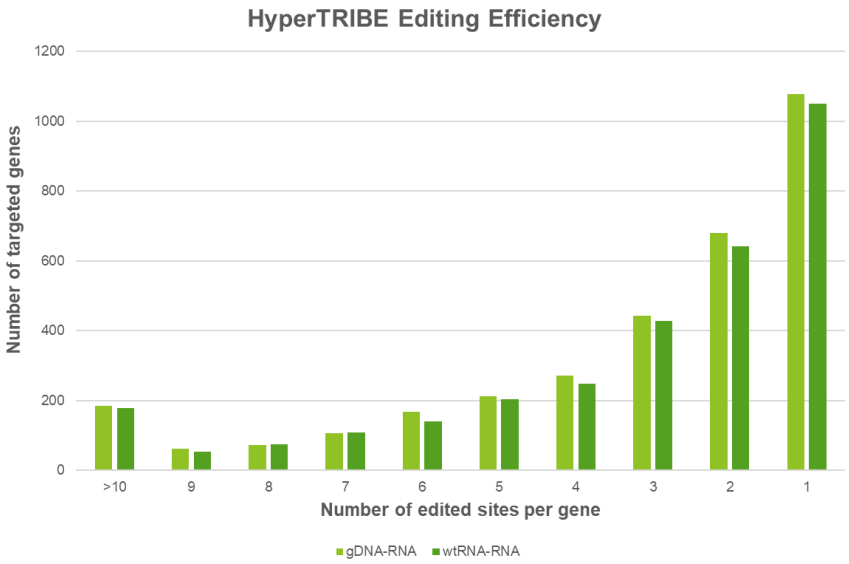


Figure 22. The identification of common edit sites between gDNA-RNA and wtRNA-RNA approaches using the *D. melanogaster* data in my analysis (left panel) and published by Rosbash group (right panel). A Venn diagram to observe high confidence editing sites by comparing datasets from two different approaches. Edit sites were identified by using genomic DNA as the reference to compare with HyperTRIBE RNA (gDNA-RNA) or using wtRNA to compare with HyperTRIBE RNA (wtRNA-RNA) [Rahman et al., Nature Protocols, 2018].



gDNA-RNA	
Parameters	**Count**
>10	185
9	63
8	73
7	106
6	167
5	212
4	271
3	442
2	679
1	1077

wtRNA-RNA	
Parameters	**Count**
>10	179
9	53
8	74
7	108
6	140
5	204
4	249
3	429
2	641
1	1051

Figure 23. HyperTRIBE editing efficiency utilizing *D. melanogaster* data. Using the published HyperTRIBE dataset to identify *Drosophila* Hrp48 protein, the editing efficiency of the HyperTRIBE was observed in the two (gDNA-RNA and wtRNA-RNA) approaches. The histogram displays the number of target genes containing one to more than 10 edit coordinates. A large ratio of the target genes is edited multiple times by the HyperTRIBE Hrp48 fusion construct.

more stably to the transcripts. When we assessed the HyperTRIBE-Hrp48 editing frequencies, similar patterns of editing events were observed from both gDNA-RNA and wtRNA-RNA datasets, which was also seen in their analysis. We obtained a list of the top 10 HyperTRIBE-Hrp48 edited transcripts and summarized the editing results from both approaches and found that almost all of the gDNA-RNA identified genes correspond with the top genes from the wtRNA-RNA approach (Figure 24).

Found in both Xu et al. [Xu et al., 2018] and my analysis, about 90% of gDNA-RNA targeted sites were identical to the results identified by the wtRNA-RNA approach. As mentioned before, a high proportion of its targeted RNA transcripts had more than one conversion of A-to-I (G). All HyperTRIBE-Hrp48 topmost edited genes from the gDNA-RNA approach were also reflected in the HyperTRIBE-Hrp48 results when wtRNA was used as the reference. Based on the HyperTRIBE-Hrp48 results that I have generated using their pipeline, the editing events from both approaches faithfully recapitulated Hrp48 binding specificity as well as the editing efficiency of the HyperTRIBE-Hrp48 fusion protein. This reanalysis of *Drosophila* data with their pipeline and reproducibility of their results has given us the confidence to use this workflow for our data.

Experimental Design

As mentioned above, there are currently no published papers on using the TRIBE/HyperTRIBE technique in plants; this method has been used only in flies and mammals. Recent results demonstrated that applying the HyperTRIBE method to an RBP, 4E-BP, found in both flies and mammals successfully characterized its direct mRNA targets and the editing data reflected the true binding specificity of 4E-BP [Jin et al., 2020]. Hence, it would be desirable to

gDNA-RNA HyperTRIBE edited genes		
Gene name	*# of edit sites*	*Avg. editing (%)*
SCAP	29	23.3
PlexA	27	18.9
hipk	27	20.4
CG11897	26	24.6
CtBP	26	20.4
CG31678	25	27.8
pico	25	25.4
CG42666	24	26.4
Sam-S	24	27.1
crol	23	24.1

wtRNA-RNA HyperTRIBE edited genes		
Gene name	*# of edit sites*	*Avg. editing (%)*
par-6	29	24
PlexA	27	18.9
hipk	27	20.4
pico	26	25
CtBP	25	20.6
SCAP	24	22.8
Sam-S	24	27.1
CG3107	22	22.6
CG31678	22	27.9
crol	22	23.2

Figure 24. Top 10 HyperTRIBE edited genes summary of *D. melanogaster* data. A list of the top 10 genes with the highest number of edit sites and highest average editing percentage from the two (gDNA-RNA and wtRNA-RNA) approaches. Both approaches yielded similar results with a slight variation in the total number of edited sites.

investigate whether TRIBE/HyperTRIBE is an adaptable tool to identify targets of other RBPs in different systems, especially in plants.

For our study, we chose to use both the TRIBE and HyperTRIBE methods for identifying the *in vivo* targets of each SR45 isoform to observe any significant difference between these strategies (Figures 25 and 26). Three biological replicates of each line were analyzed to filter out any transient editing sites and to restrict our consideration to sites consistently present in all triplicates. Wild-type (WT) lines are utilized as a reference to determine the editing sites. This is done by comparing nucleotide frequencies from each TRIBE/HyperTRIBE RNA-seq reads to the wtRNA. Negative controls include mutant (*sr45-1*) lines and lines expressing TRIBE and HyperTRIBE catalytic domain alone. Because of endogenous editing activity that may occur from the catalytic enzyme, editing events from TRIBE and HyperTRIBE ADARcd alone are subtracted from edit sites found in each TRIBE/HyperTRIBE line to remove any background noise. These control lines are significantly crucial to assure editing events are specified by the RBP and that a high confidence set of HyperTRIBE editing sites are being generated.

Generation of TRIBE and HyperTRIBE fusion constructs

The transgenic lines were generated by fusing each SR45 isoform coding region with a short linker to the *Drosophila* ADAR catalytic domain (dADARcd) (Figure 25A, B). To create the HyperTRIBE enzyme domain, a mutation was incorporated to change one amino acid at position 488 (E488Q) of the original SR45 TRIBE construct using Q5 site-directed mutagenesis. The transgenic lines harboring the fusion protein construct (HyperTRIBE-*SR45* short; HyperTRIBE-*SR45* long; TRIBE-*SR45* short; and TRIBE-*SR45* long) and lines expressing TRIBE and HyperTRIBE ADAR catalytic domain alone are expressed in the *sr45-1* mutant background using a constitutive CaMV 35S promoter. WT lines for this experiment were

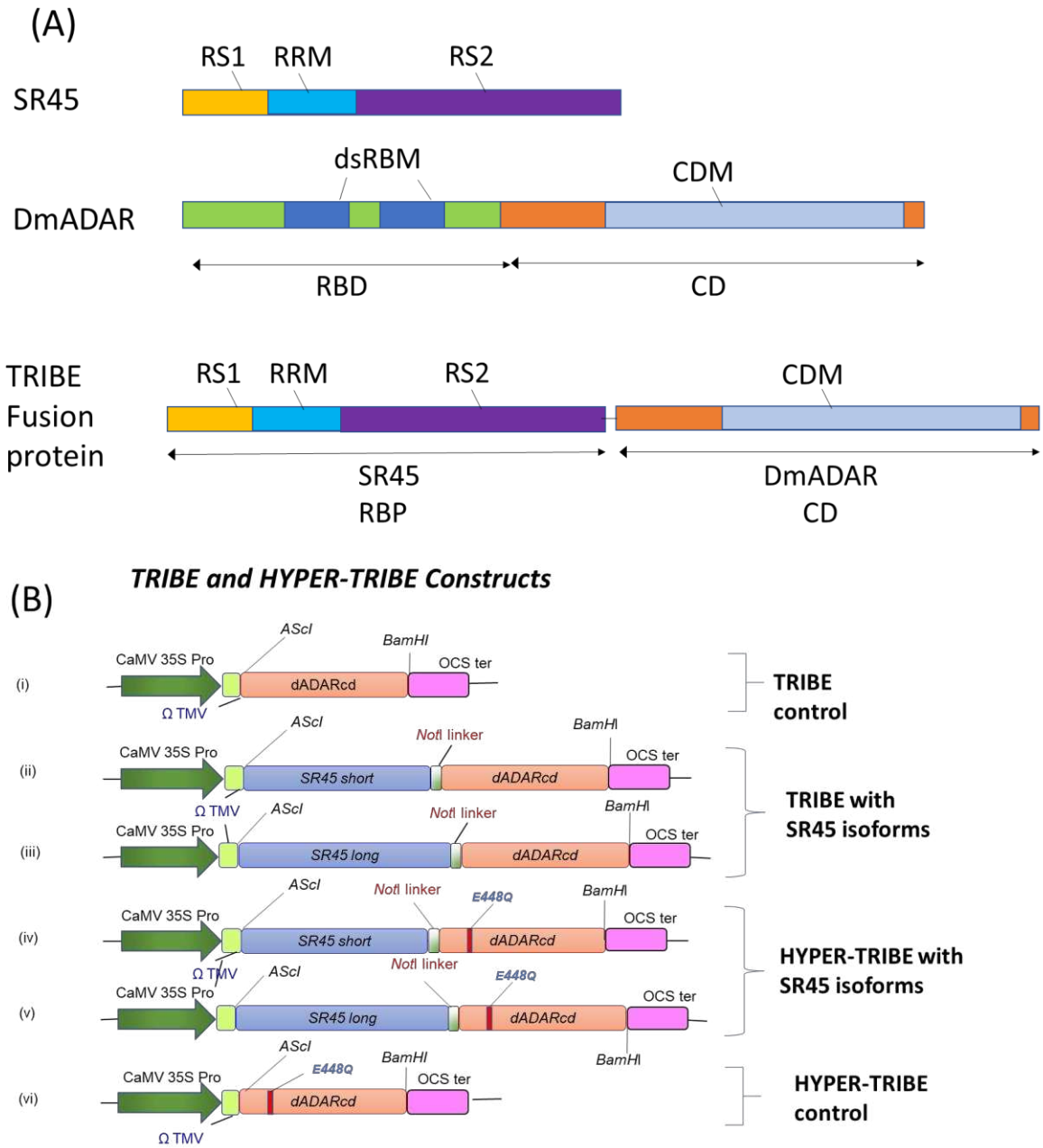


Figure 25. TRIBE and Hyper-TRIBE constructs. A) A schematic diagram showing the domain structure of SR45, DmADAR, and the TRIBE fusion protein. RS1, arginine serine-rich domain 1; RRM, RNA recognition motif; RS2, arginine serine-rich 2 domain; dsRBM; double-stranded RNA binding motif; CDM; catalytic domain motif; DmADAR, *D. melanogaster* ADAR domain. B) Schematic diagram of TRIBE and HyperTRIBE constructs with and without SR45 isoforms driven by CaMV 35S promoter. dADARcd, *Drosophila* ADAR catalytic domain; ΩTMV, tobacco mosaic virus translation enhancer; *Asc*I in forward and BamHI in reverse primer; OCS, octopine synthase.

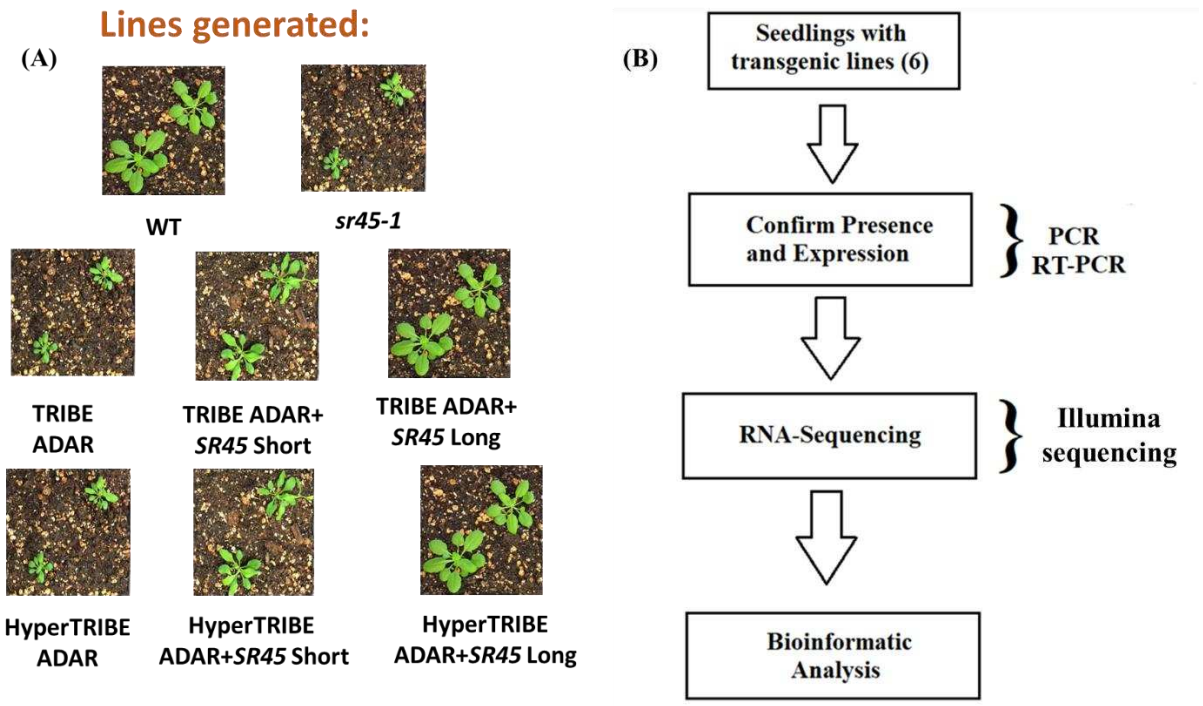


Figure 26. The phenotype of transgenic lines and the workflow for the analysis of the RNA-seq reads from TRIBE/HyperTRIBE transgenic lines. A) Here shown are the phenotypes of different transgenic lines that were generated: wildtype, *sr45-1*, TRIBE-ADAR, TRIBE ADAR + *SR45* Short, TRIBE ADAR + *SR45* Long, HyperTRIBE ADAR, HyperTRIBE ADAR + *SR45* Short, and HyperTRIBE ADAR + *SR45* Long. B) The workflow for RNA-seq analysis to identify the RNA binding targets of each SR45 isoform.

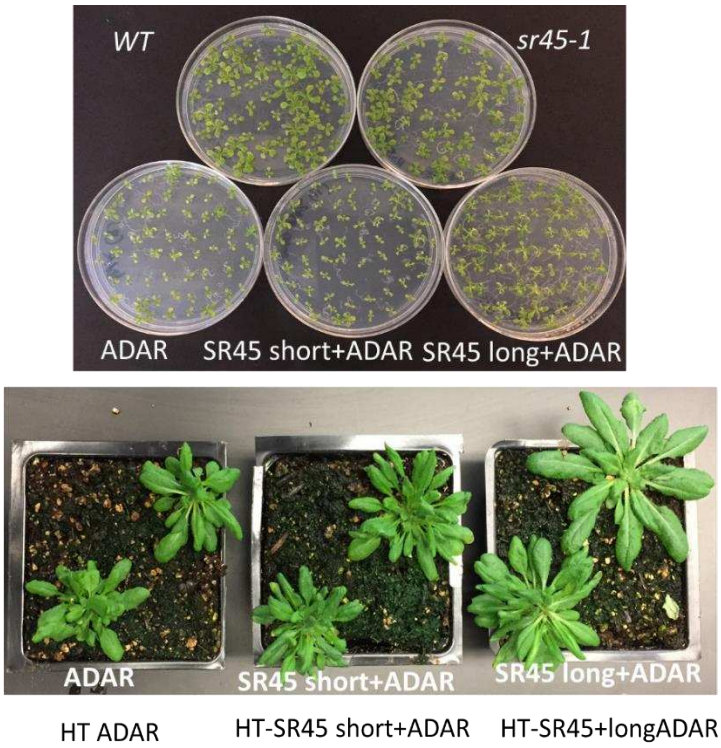
Arabidopsis thaliana ecotype Columbia-0 (Col-0) and homozygous mutants were used for the controls. Homozygous transformed lines of independent TRIBE and HYPERTRIBE lines were used throughout the experiment. Expression of SR45 was detected in all plant organs and tissues with high levels in inflorescence tissues, imbibed seeds, shoot apex, and root tips [Zimmermann et al., 2004]. We expressed each transgenic construct in thirty-day-old T2 plants of *sr45-1* and used total RNA as the source for our RNA-sequencing, so the tissues we used may not reveal binding targets in inflorescence tissues.

Verification of the expression of TRIBE/HyperTRIBE-SR45 isoforms in transgenic lines

Plants lacking SR45 protein display previously reported pleiotropic traits, suggesting that this splicing factor affects multiple genes involved in different growth and developmental processes. As shown in Figure 27, the *sr45-1* plants were distinctively smaller in size with pointy, narrow leaves compared to the larger, round leaves of the wild-type plants (Figure 27A, B). Exhibiting similar mutant features were also found in the TRIBE/HyperTRIBE ADAR alone plants. These similarities are due to the lack of the SR45 protein in the TRIBE/HyperTRIBE ADAR alone lines. The HyperTRIBE-SR45 long plants fully complemented the plant size and petal phenotype of the mutant, while the HyperTRIBE-SR45 short lines partially rescued those features. The HyperTRIBE-SR45 short plants were half of the size of the wild-type and showed abnormal number of floral organs. These observations further verify that SR45 long isoform does contribute to plant flower and petal development and SR45 short isoform does not fully complement these attributes.

Expression analysis in transgenic lines of TRIBE and HyperTRIBE constructs was performed using reverse transcriptase-PCR (RT-PCR) (Figure 28A, B). Three different sets of

(A)



(B)

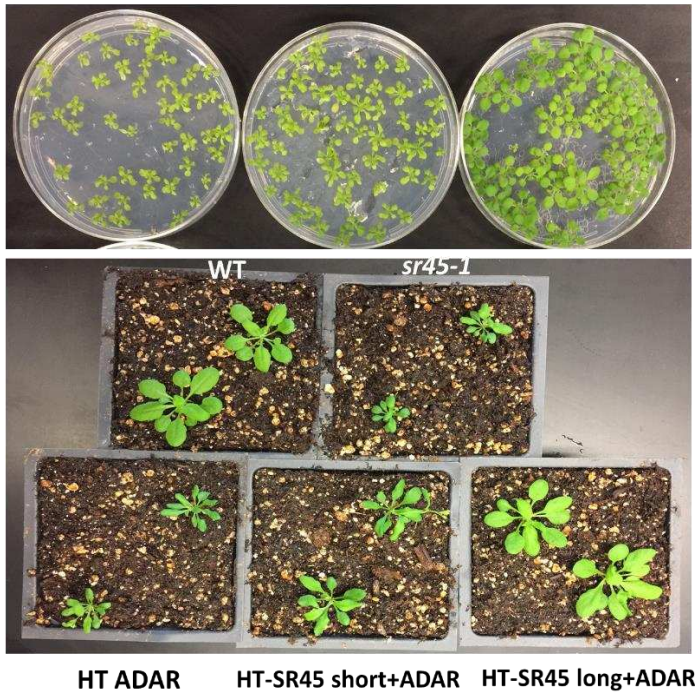


Figure 27. Phenotypes of wildtype, mutant, and TRIBE/HyperTRIBE-SR45 lines. Plant images of 30-day-old day seedlings. A) wildtype, mutant, TRIBE and B) HyperTRIBE lines are shown to observe the plant growth and development in different stages.

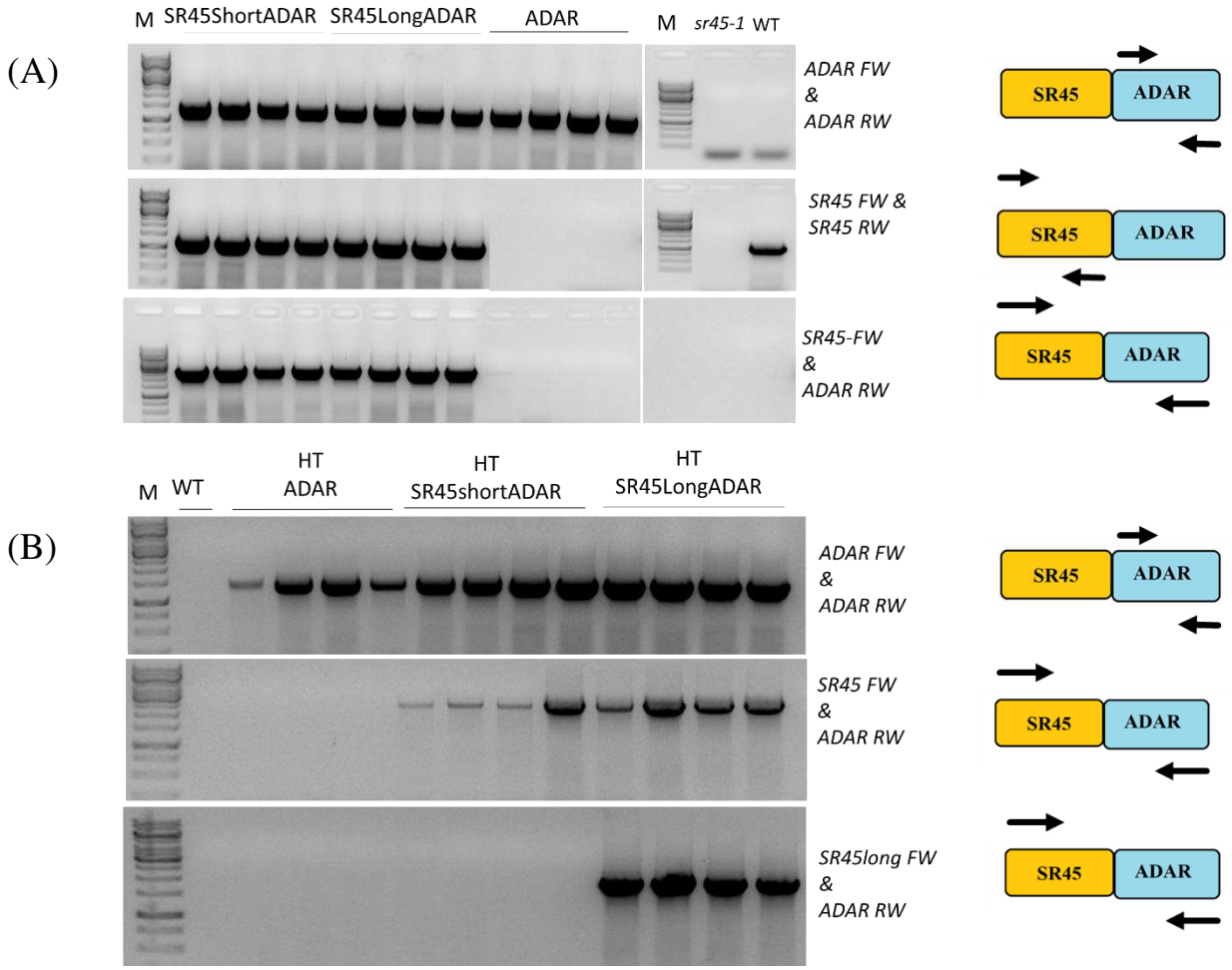


Figure 28. Expression analysis in transgenic TRIBE and HyperTRIBE lines. RT-PCR expression analysis was used to verify that the A) TRIBE and B) HyperTRIBE lines were expressing the introduced gene constructs. Specific primers were utilized to amplify the full length of the fusion protein (SR45 FW + ADAR RW), ADAR alone (ADAR FW + ADAR RW), SR45 (SR45 FW + SR45 RW), and specific isoforms of SR45 (SR45 long FW + ADAR RW). Four independent lines of TRIBE/HyperTRIBE *SR45.1*+ ADAR, TRIBE/HyperTRIBE *SR45.2* + ADAR, and ADAR alone were analyzed; these results confirmed expression of all constructs in transgenic lines. FW, forward primers; RW, reverse primers. Arrows represent the direction of the primers.

primers were used to either amplify the full length or regions of the chimeric TRIBE/HyperTRIBE protein. Utilizing the set of primers to amplify the full length of the fusion, amplification was observed in TRIBE/HyperTRIBE-*SR45* fusion transgenic lines complemented with *SR45.1* or *SR45.2* because these lines are expected to express the full fusion protein. All TRIBE/HyperTRIBE transgenic lines amplified the expression of ADARcd with the ADARcd-specific primers, which indicates that these lines contain the ADAR domain. With the SR45-specific primers, TRIBE/HyperTRIBE-SR45 short, TRIBE/HyperTRIBE-SR45 long, and the wildtype lines displayed expression of the *SR45*, while the TRIBE and HyperTRIBE ADARcd alone showed no amplification using these primers. Overall, these results demonstrated that the transgenic lines successfully expressed the introduced constructs, and the control lines were also showing the expected results.

A differential gene analysis using the bioinformatic package Deseq2 was performed to evaluate the expression of SR45 (gene ID: AT1G16610) from each of the lines (Figure 29). After normalizing counts and implementing a pseudocount of 0.5 to allow for log scale plotting, the count of reads for *SR45* was calculated for each sample. As expected, the expression levels of SR45 protein are sufficiently present in the TRIBE- and HyperTRIBE- lines due to utilizing the constitutive CaM35SV promoter to express the TRIBE/HyperTRIBE fusion constructs. However, the HyperTRIBE-SR45 lines (short and long) have a higher level of *SR45* transcript level compared to the TRIBE lines. The HyperTRIBE-SR45 short has about over 20,000 counts and HyperTRIBE-SR45 long has approximately 18,000 counts; the TRIBE-SR45 short has around 9000 counts and the TRIBE-SR45 long has roughly 7500 counts. Whereas the wildtype line has an average of ~1000 counts of *SR45* expression, which is almost 10-20X less compared to the expression levels found in the *SR45* isoform lines. The *sr45-1* line and the TRIBE and

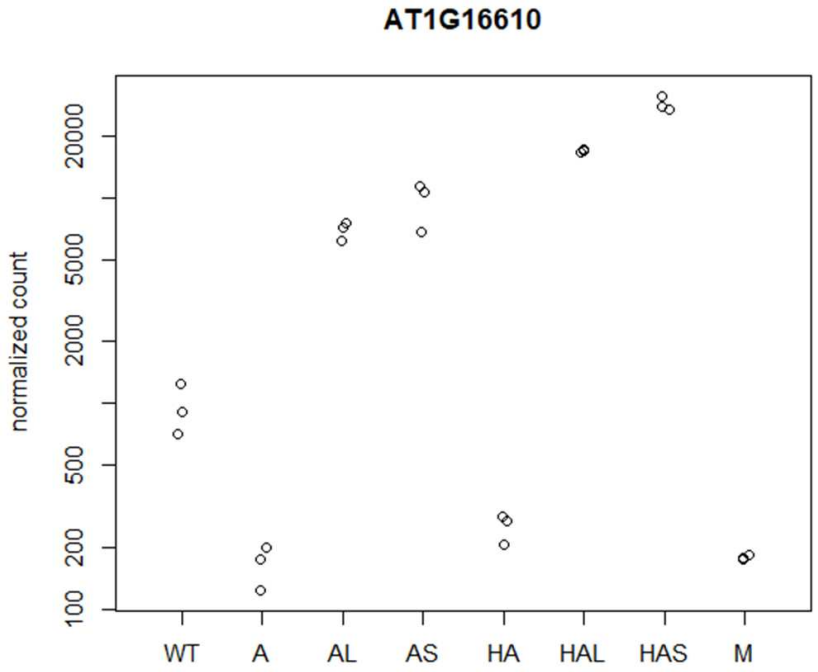


Figure 29. Verification of *SR45* expression in transgenic lines. A differential gene expression analysis was performed to detect the level of *SR45* expression, gene id: AT1G16610, in all our eight samples to verify that the transgenic constructs were expressing the introduced *SR45* in the *sr45-1* background. WT, wildtype, A, TRIBE ADAR; AL, TRIBE-*SR45* long, AS, TRIBE-*SR45* short, HA, HyperTRIBE-ADAR; HAL, HyperTRIBE-*SR45* long; HAS, HyperTRIBE-*SR45* short; M, mutant.

HyperTRIBE ADARcd alone showed marginally low expression of *SR45* (normalized counts ranged from about 150-350) throughout all triplicates. Because the fusion constructs were expressed under a mutant background, it is anticipated to see low levels of *SR45* expression in the lines with constructs that do not have any SR45 isoform linked with the TRIBE/HyperTRIBE ADARcd.

Data analysis pipeline for the identification of RNA editing sites of Arabidopsis SR45 protein

Now that we have confirmed and validated that our transgenic lines have the expected expression and phenotype, RNA from the seedlings was sequenced as paired-end reads using Illumina's NovaSeq 6000 sequencing platform. We sequenced eight lines with three biological replicates for each line: wildtype, *sr45-1*, TRIBE (ADARcd alone, SR45-long, and SR45-short), and HyperTRIBE (ADARcd alone, SR45-long, and SR45-short) lines, a total of 24 samples (Figure 30). We obtained approximately 46 to 52 million reads per sample. Roughly ~95-97.8% of the reads from each library were uniquely mapped to the Arabidopsis genome (Table 1).

For our bioinformatics pipeline, we used our own wild-type as our reference because our RNA libraries were prepared from the same genetic background. We used the Arabidopsis genome sequence and the TAIR10 genome annotation as our transcriptome (Figure 31). The columns of the genome annotation file, originally a GFF format, had to be rearranged properly to be converted into the desired formats, GTF and refFlat, that the protocol required. We attempted to use the same editing parameters that were used to filter the previous *Drosophila* dataset, at least 10% editing and at least a coverage of 20 reads, for our Arabidopsis TRIBE/HyperTRIBE RNA libraries but the majority of the edit sites found in the TRIBE- and HyperTRIBE- lines

HyperTRIBE Transgenic Line	Replicate
HyperTRIBE-ADAR **(HA)**	1
HyperTRIBE-ADAR **(HA)**	2
HyperTRIBE-ADAR **(HA)**	3
HyperTRIBE- ADAR + SR45 Short **(HAS)**	1
HyperTRIBE- ADAR + SR45 Short **(HAS)**	2
HyperTRIBE- ADAR + SR45 Short **(HAS)**	3
HyperTRIBE- ADAR + SR45 Long **(HAL)**	1
HyperTRIBE- ADAR + SR45 Long **(HAL)**	2
HyperTRIBE- ADAR + SR45 Long **(HAL)**	3

TRIBE Transgenic Line	Replicate
TRIBE-ADAR **(A)**	1
TRIBE-ADAR **(A)**	2
TRIBE-ADAR **(A)**	3
TRIBE- ADAR + SR45 Short **(AS)**	1
TRIBE- ADAR + SR45 Short **(AS)**	2
TRIBE- ADAR + SR45 Short **(AS)**	3
TRIBE- ADAR + SR45 Long **(AL)**	1
TRIBE- ADAR + SR45 Long **(AL)**	2
TRIBE- ADAR + SR45 Long **(AL)**	3

Control Line	Replicate
Wildtype **(WT)**	1
Wildtype **(WT)**	2
Wildtype **(WT)**	3
sr45-1 **(M)**	1
sr45-1 **(M)**	2
sr45-1 **(M)**	3
TOTAL NUMBER OF SAMPLES	**24**

Figure 30. A summary of samples used for RNA-seq. The different types and replicates of each HyperTRIBE (HyperTRIBE-*SR45* short, HyperTRIBE-*SR45* long, HyperTRIBE-ADARcd), TRIBE (TRIBE-*SR45* short, TRIBE-*SR45* long, TRIBE-ADARcd), and control (wildtype and *sr45-1*) lines that went through high throughput-sequencing.

Table 1. TRIBE/HyperTRIBE read alignments. The total number of input reads and the percentage that were uniquely mapped to the Arabidopsis genome. AVG, average; WT, wildtype; M, mutant; HA, HyperTRIBE-ADAR; HAL, HyperTRIBE-*SR45* long; HAS, HyperTRIBE-*SR45* short; A, TRIBE-ADAR; AL, TRIBE-*SR45* long, AS, TRIBE-*SR45* short.

	# of Input Reads	Mapped Unique
WT_1	21076192	94.8%
WT_2	22250991	95.4%
WT_3	216146601	95.3%
AVG_WT	24855478	95.2%
M_1	23036364	95.4%
M_2	25267938	95.3%
M_3	24855478	95.2%
AVG_M	24386593	95.3%
HA_1	20882738	96.3%
HA_2	13353746	97.5%
HA_3	1877448	97.2%
AVG_HA	12037977	97.0%
HAL_1	20785816	97.9%
HAL_2	22233765	98.0%
HAL_3	22109161	97.5%
AVG_HAL	21709581	97.8%
HAS_1	22892826	97.8%
HAS_2	24158649	97.3%
HAS_3	23777873	97.3%
AVG_HAS	23609783	97.5%
A_1	22443298	97.5%
A_2	21901625	97.3%
A_3	22399185	95.4%
AVG_A	22248036	96.7%
AL_1	21842726	97.0%
AL_2	21684756	97.3%
AL_3	22208696	96.4%
AVG_AL	21912059	96.9%
AS_1	18368628	92.5%
AS_2	19142808	95.6%
AS_3	21820879	97.1%
AVG_AS	19777438	95.1%

Database	Web Links	Version/Date
Genome	https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_gff3	(TAIR10) 11/2019
DNA sequence	https://plants.ensembl.org/Arabidopsis_thaliana/Info/Index	n/a
Gene Orthology (GO)	http://geneontology.org	05/2019
HyperTRIBE codes	https://github.com/rosbashlab/HyperTRIBE	04/2020

Figure 31. Databases used in our analysis. The sources and the version of the different databases that were utilized for the TRIBE and HyperTRIBE RNA-seq analysis.

were below this threshold. For our datasets, three different thresholds were used to select the editing sites; edit sites were required to have at least 10% editing, 5% editing, or 1% editing.

The comparison between edit sites of TRIBE/HyperTRIBE-*SR45* isoforms

Ideally, we are interested in identifying transcripts with higher occurrences of editing marks to give us more confident targets with a stronger binding interaction to the protein of interest. For every TRIBE/HyperTRIBE isoform line, we overlapped the pool of edited sites between its three replicates to only output consistent and reproducible results. To assure that we are retrieving high confidence sets of HyperTRIBE editing sites, we visualized each potential binding site on the Integrated Genome Browser (IGB) viewer (Figure 32). The less stringent editing threshold of editing sites with at least 1% editing provided about 243 unique edit sites from the HyperTRIBE-SR45 short lines and 110 unique edit sites from the HyperTRIBE-SR45 long lines (Figure 33A). Approximately, 79 sites were edited from both HyperTRIBE-SR45 isoforms. However, utilizing another editing threshold, at least 5% editing, removed a substantial amount of editing sites from the list of potential SR45 binding targets for both HyperTRIBE-SR45 isoform libraries. And the list continues to be reduced considerably when the parameter is more restricted (at least 10% editing).

We also implemented the same three editing thresholds to the TRIBE lines and found that the pool of edit sites for all TRIBE *SR45* isoforms was significantly smaller compared to the HyperTRIBE lines (Figure 33B). With an editing frequency threshold of at least 1%, we identified 55 sites that were edited only from TRIBE-SR45 short, 64 sites marked with edits only from TRIBE-SR45 long, and 4 common edit sites from TRIBE-SR45 short and TRIBE-SR45 long. Only 1 unique edit site from TRIBE-SR45 short satisfied the criterion when using both 5%

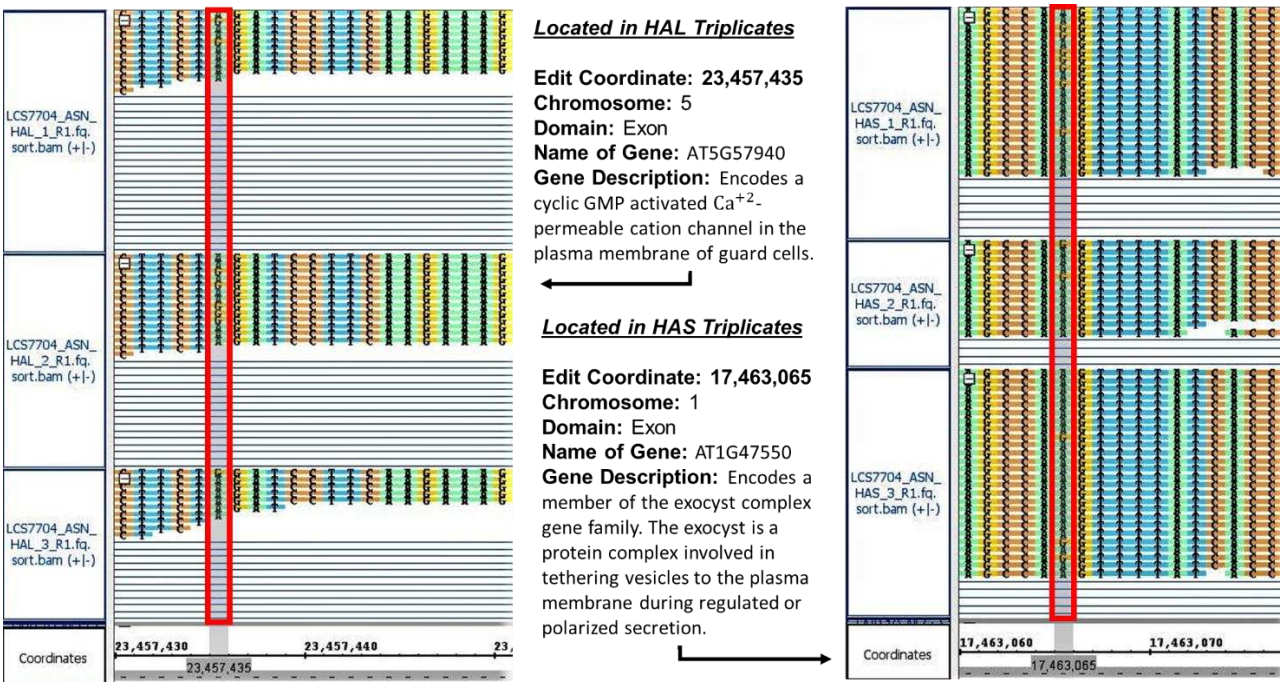


Figure 32. IGB Browser- Verifying Edit Sites i.e., Conversion from A → G. In this study, high confidence editing sites are required to be present in all three replicates. An edit site from each line was verified by visualizing the edit coordinate on the Integrated Genome Browser (IGB). An edit coordinate on the gene, AT5G57940, was located in all HyperTRIBE-*SR45* long triplicates and an edit site on the gene, AT1G47550, was found in all replicates of HyperTRIBE-*SR45* short.

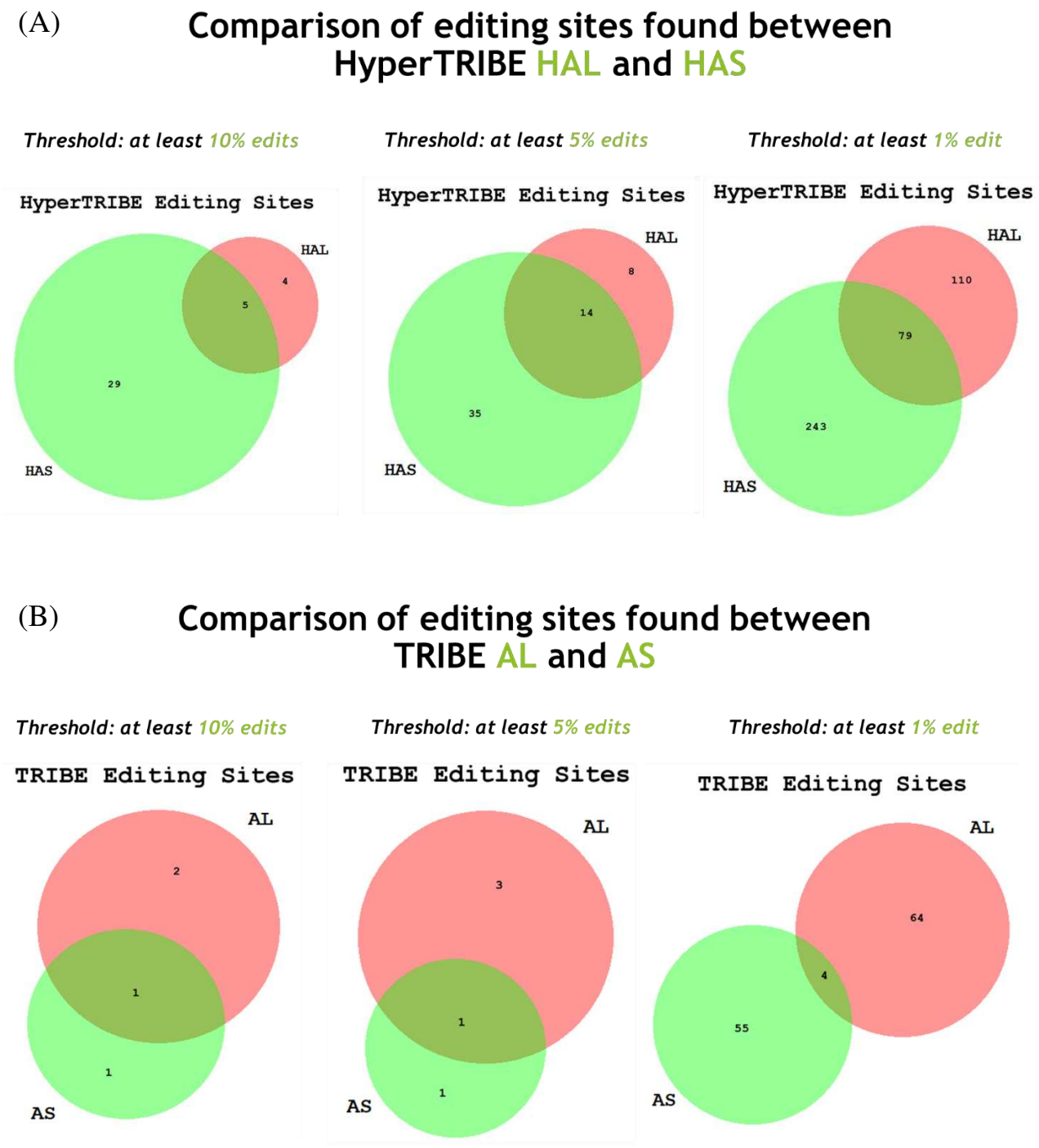


Figure 33. Comparison of editing sites found between HyperTRIBE-and TRIBE-*SR45* lines. Using three different editing thresholds: at least 10%, 5%, or 1% editing to refine the list of editing sites that were found in *SR45* long and *SR45* isoform of the A) HyperTRIBE and B) TRIBE lines. HAL, HyperTRIBE-*SR45* long; HAS, HyperTRIBE-*SR45* short; AL, TRIBE-*SR45* long; AS, TRIBE-*SR45* short.

and 10% editing threshold. The TRIBE-SR45 long also displayed a very low amount of editing events in both thresholds, approximately 32X fold less than using the threshold of 1% editing.

Based on these results, the HyperTRIBE fusion construct seemingly edits more compared to the editing from the TRIBE-expressing cells. The number of editing events when using a less conservative threshold for each TRIBE and HyperTRIBE line demonstrated that the majority of edit sites displayed editing frequencies of 1-5%. Unlike the HyperTRIBE, the TRIBE-expressing cell also revealed that the editing was not as consistent or reproducible due to an extensive amount of edit sites being filtered out from overlapping all replicates of each line. This study proved that the HyperTRIBE ADARcd is a reliable editing enzyme that identifies dramatically more editing sites in plants as compared to the original TRIBE ADARcd.

Gene Ontology (GO) analysis on targeted transcripts of TRIBE/HyperTRIBE-SR45 isoforms

SR45 has been implicated to play an essential role in plant development and modulating stress responses by participating in various regulatory and splicing processes. We performed a gene ontology (GO) enrichment analysis with the TRIBE and HyperTRIBE-identified genes to gain functional insights into these RNA targets (Figure 34, 35). For a threshold of 1% editing, the most GO enriched terms were “catalytic activity” and “binding” under the molecular function and “cellular processes” and “metabolic processes” under the biological processes in all TRIBE/HyperTRIBE lines. Further analysis of the HyperTRIBE-SR45 long targets, about 10% of its RNA targets fell under the category of “response to stimulus”. The HyperTRIBE-SR45 short has about 7% and both common targets of HyperTRIBE long and short lines have 5% that lie in that same GO category. And within that category, HyperTRIBE-SR45 long has GO terms

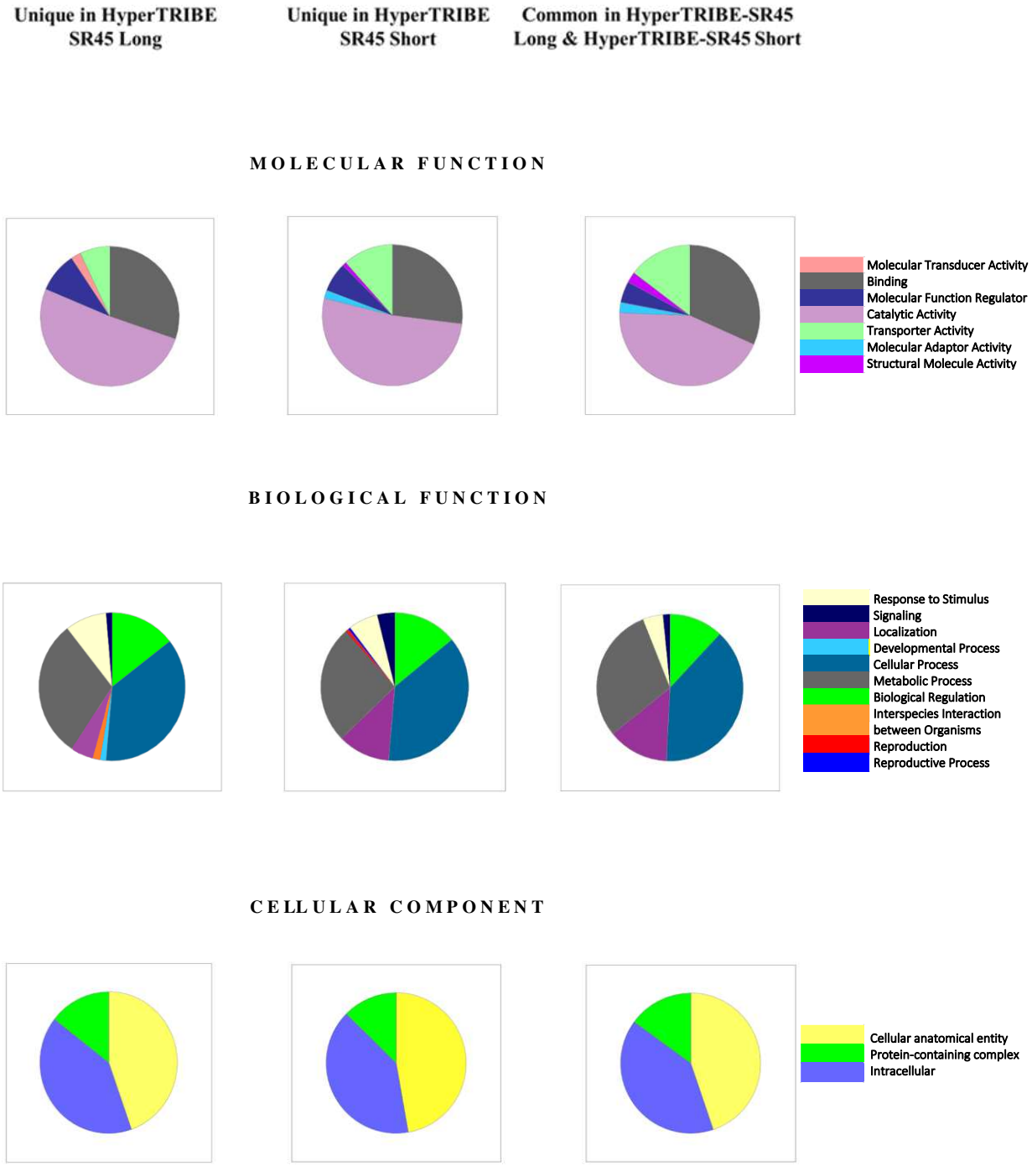


Figure 34. GO enrichment analysis of HyperTRIBE targets with 1% threshold. A gene ontology (GO) enrichment analysis with the HyperTRIBE-identified genes was done using the dataset with the threshold of at least 1% editing. Pie charts depict each of three major GO aspects: molecular function, biological process, and cellular component.

Figure 35. GO enrichment analysis of TRIBE targets with 1% threshold. A gene ontology (GO) enrichment analysis with the TRIBE-identified genes was done using the dataset with the threshold of at least 1% editing. Pie charts depict each of three major GO aspects: molecular function, biological process, and cellular component.

associated with “response to biotic stimulus”, “response to abiotic stimulus”, and “response to external stimulus”. Interestingly, HyperTRIBE-SR45 long has “response to external stimulus” as its most enriched term with a 3.78-fold enrichment and high confidence (p-value = 7.87E-6) among all GO categories. Whereas, HyperTRIBE-SR45 short had a high percentage for the GO term, “response to endogenous stimulus”. This suggests that SR45.1 is strongly associated with stress-responsive genes and has a major regulatory role in external stimulus-response as compared to SR45.2.

“Embryonic meristem development”, “macroautophagy”, and “ion transport” were the three most significant enrichment GO categories shown in HyperTRIBE-SR45 short with a fold of enrichment greater than 2. We also found that unique targets of HyperTRIBE-SR45 short exclusively have GO terms of “molecular adaptor” and “structural molecule activity”. This evidence indicates that SR45.2 potentially interacts with other RNA binding partners or contributes to an assembly of a complex. Common targets between HyperTRIBE-SR45 long and HyperTRIBE-SR45 short shared similar biological processes like “export from cell”, “vacuolar transport”, “transmembrane transport”, and “exocytic process”. Based on these GO terms, this supports the hypothesis that SR45.1 and SR45.2 isoforms both contribute to intracellular trafficking.

About 97% of the GO categories from HyperTRIBE lines were exhibited in the TRIBE lines. But it appears that the HyperTRIBE lines had some GO categories that were not featured in the TRIBE lines. For instance, HyperTRIBE-SR45 short has RNA targets involved in “reproduction” and “reproductive process”, while TRIBE-SR45 short does not have any that were associated with these categories. Even though the enzyme activity of the ADARcd from

TRIBE generated consistent results of the HyperTRIBE, the ADARcd of the HyperTRIBE potentially binds to more target sites thus providing a more comprehensive list of GO terms.

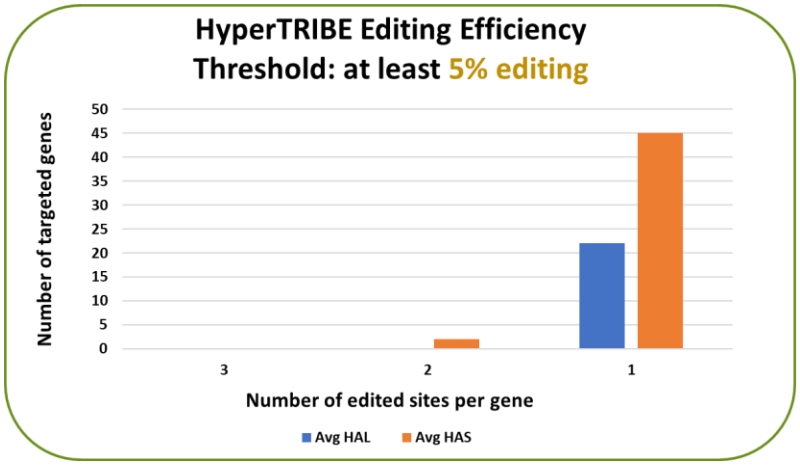
The editing efficiency of TRIBE/HyperTRIBE-SR45 isoforms

We analyzed the editing percentage for each TRIBE and HyperTRIBE-SR45 isoform line by recording the number of editing sites per transcript. We totaled the number of edit sites per transcript from each replicate and averaged them to calculate the editing efficiency in each TRIBE/HyperTRIBE library. We utilized two different thresholds, target sites having at least 5% or at least 1% editing, to generate two datasets and observe any significant differences between them (Figure 36A, B). Both TRIBE-*SR45* short and long lines displayed similar editing event patterns as the HyperTRIBE lines (Figure 37A, B). When reducing the threshold to 1% editing, both TRIBE and HyperTRIBE showed more multiple-edited genes in both TRIBE/HyperTRIBE-SR45 short and long lines. Moreover, the data suggest that TRIBE and HyperTRIBE-*SR45* fusion constructs are capable of editing multiple adenosines on a transcript, but it is most likely that a large percentage of the sites are edited at a low editing frequency.

Target transcripts of TRIBE and HyperTRIBE lines that are alternatively spliced

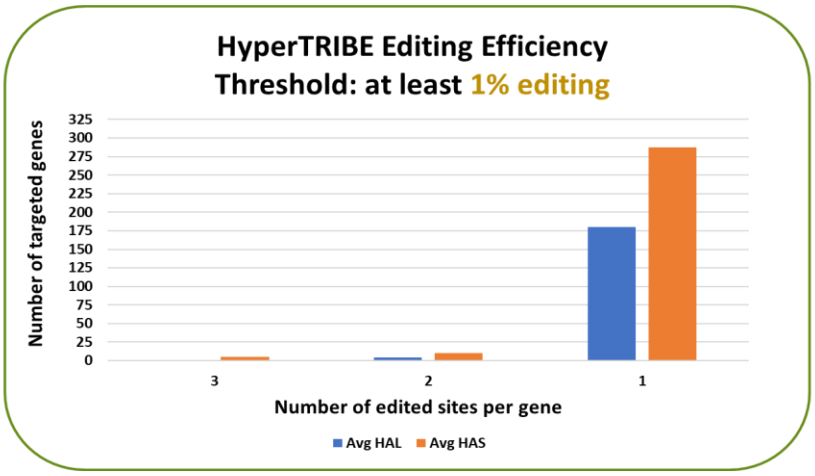
We also investigated whether the targets of SR45 isoforms that we have identified are alternatively spliced (Figure 38). To generate our TRIBE and HyperTRIBE data, at least one edit coordinate on a transcript was utilized as the threshold to consider an editing site as a candidate binding target. Editing sites reproduced in all three replicates were only retained. To obtain high confidence editing sites, we additionally required the sites to be present in both TRIBE and HyperTRIBE lines. About 120 targets of the SR45 long isoform and 87 targets of the SR45 short

(A)



Number of Sites per Gene	Number of Genes	
	Avg HAL	Avg HAS
3	0	0
2	0	2
1	22	45

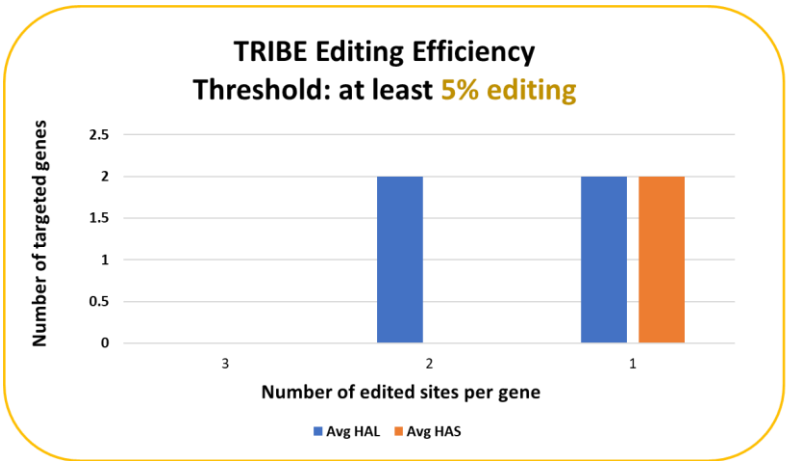
(B)



Number of Sites per Gene	Number of Genes	
	Avg HAL	Avg HAS
3	1	5
2	4	10
1	180	287

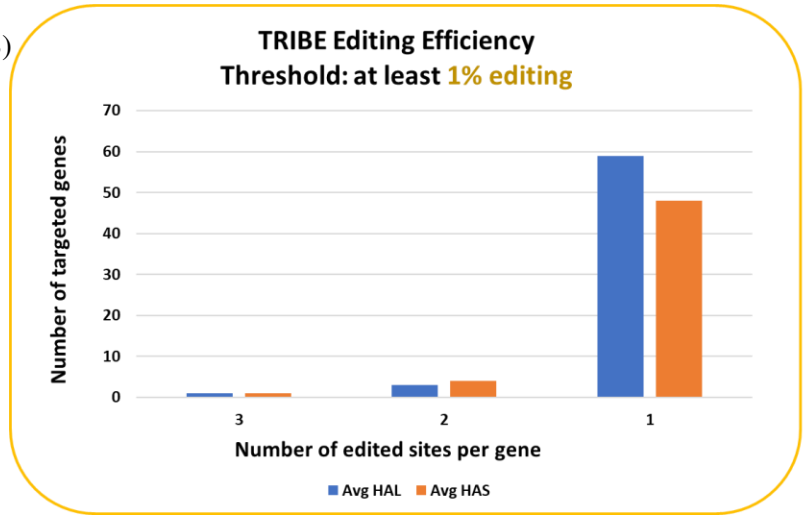
Figure 36. Multiple editing sites within a gene occur at a low editing efficiency in HyperTRIBE lines. The editing efficiency of HyperTRIBE-*SR45.1* and *SR45.2* lines. The histogram displays the number of target genes containing one to more than 10 edit sites. The datasets were generated by using two different editing thresholds: at least 5% and at least 1% editing. The histogram compares the number of edit sites per gene in HyperTRIBE *SR45* long and short lines using A) a threshold of 5% editing and B) 1% editing. Avg HAL, the average number of edit sites per gene in HyperTRIBE-*SR45* long; Avg HAS, the average number of edit sites per gene in HyperTRIBE-*SR45* short.

(A)



Number of Sites per Gene	Number of Genes	
	Avg HAL	Avg HAS
3	0	0
2	2	0
1	2	2

(B)



Number of Sites per Gene	Number of Genes	
	Avg HAL	Avg HAS
3	1	1
2	3	4
1	59	48

Figure 37. Multiple editing sites within a gene occur at a low editing efficiency in TRIBE lines. A histogram that shows the number of edit sites per gene in TRIBE SR45 long and short lines using A) a threshold of 5% editing and B) 1% editing. Avg AL, the average number of edit sites per gene in TRIBE-SR45 long; Avg AS, the average number of edit sites per gene in TRIBE-SR45 short.

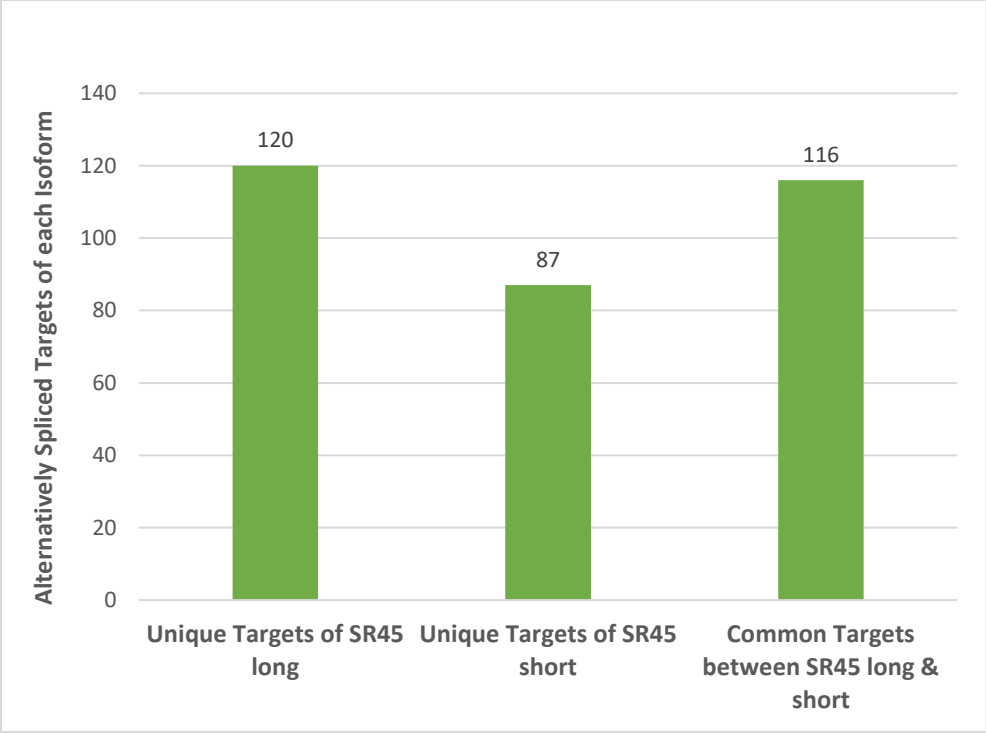


Figure 38. RNA targets of each isoform that are alternatively spliced. The number of target genes from SR45 long and short that are derived from alternatively spliced transcripts. Target genes of SR45 long and short were considered only if they were present in both TRIBE and HyperTRIBE lines.

isoformsare alternatively spliced. Also, we found 116 common targets for both SR45 short and long isoforms. GO analysis on targets with splice isoforms was done to determine if they are enriched in specific functional categories (Figure 39). All lines had similar GO categories but no significant enrichment in any categories was observed. Common targets of SR45 long and SR45 short had unique functions under the annotation, response to biological processes, like "reproductive process", "reproduction", "developmental process", and "multicellular organismal process". These shared GO terms further strengthen the theory that both SR45 long and short are involved in the regulation of plant reproduction and development. Under molecular function, "structural molecule activity" was the only GO term that was shown for unique targets of SR45 short. These findings are consistent with the results that were also observed from our previous GO analysis with the dataset of a 1% editing threshold.

The overlap of target genes from HyperTRIBE lines associated with the SR45 RIP-seq dataset

In Xing et al. 's paper [Xing et al., 2015], they performed RNA immunoprecipitation (RIP) and high-throughput sequencing for a transcriptome-wide identification of RNA targets of SR45 in Arabidopsis. They provided an extensive list of SR45-associated RNAs (SARs) of over 4,000 genes that are directly or indirectly associated with SR45. In their findings, they discovered that SR45 participates in an important function in regulating the expression of numerous abscisic acid (ABA) genes and also an unexpected role in mRNA processing of intronless genes. To test if the direct RNA targets of SR45 identified in this study overlap with the previously identified direct and indirect targets, we compared these two datasets using a 1% editing threshold to identify overlap of targets between these two datasets (Figure 40).

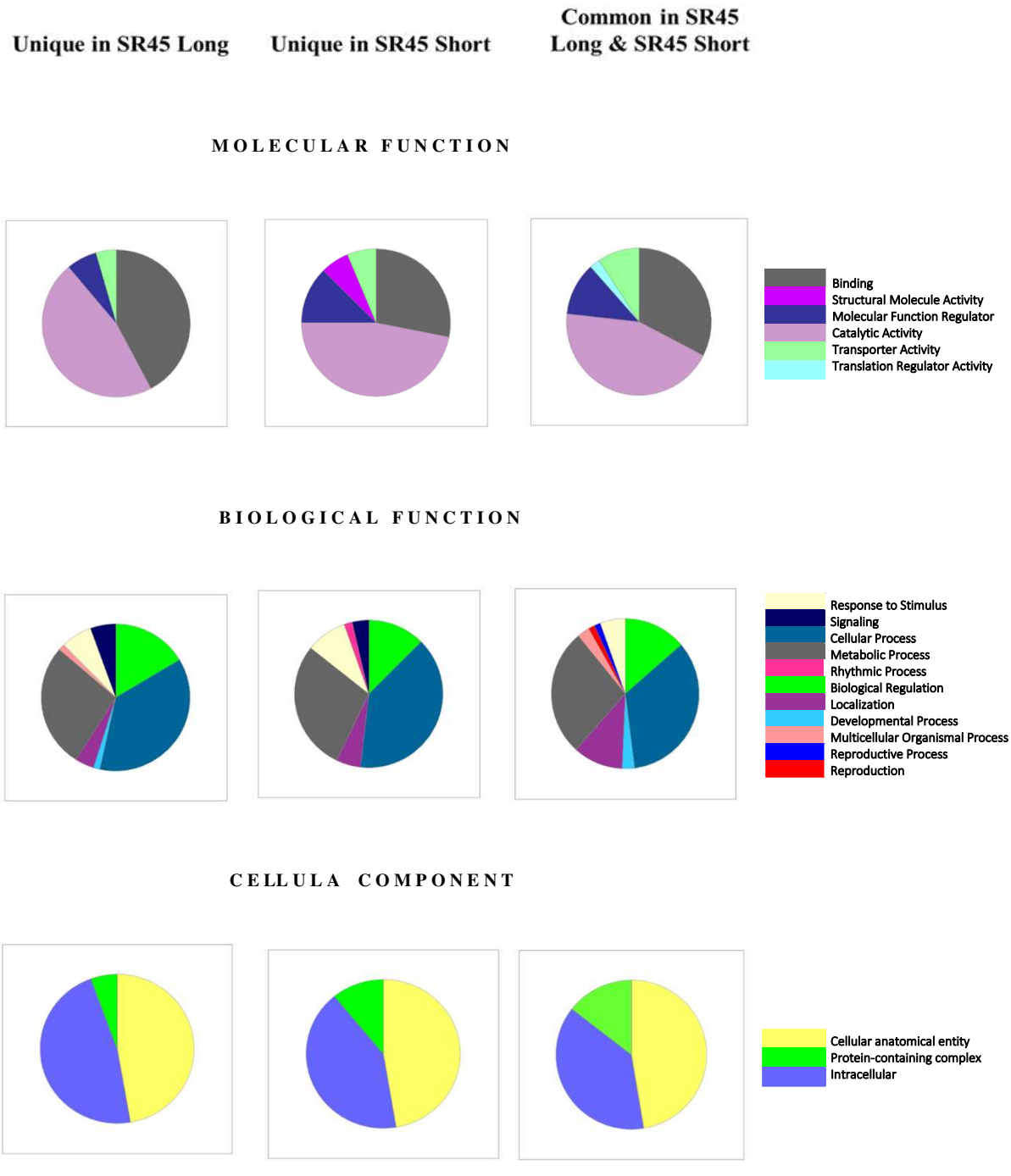


Figure 39. GO enrichment analysis of SR45 targets that are alternatively spliced. A gene ontology (GO) enrichment analysis with alternatively spliced targets of each SR45 isoform and from both isoforms was performed. Pie charts depict each of three major GO aspects: molecular function, biological process, and cellular component.

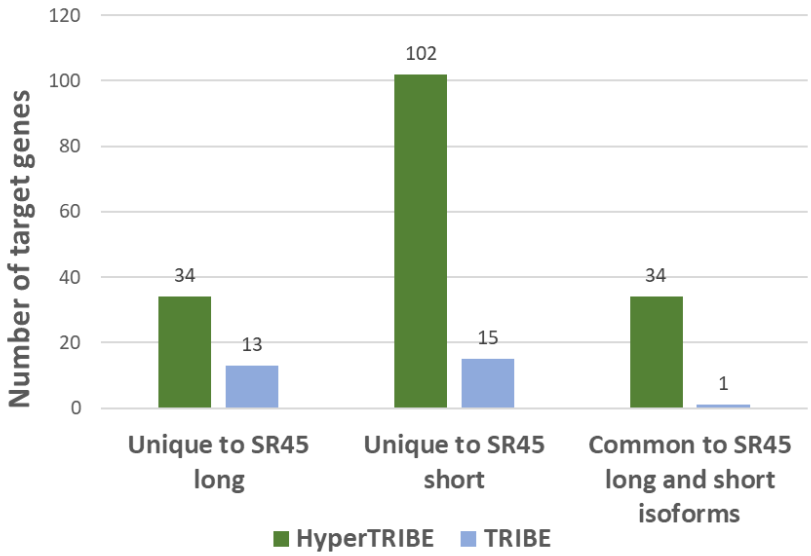


Figure 40. Overlap between the RNA targets identified in TRIBE/HyperTRIBE approach and SR45 RIP-seq method. The number of targets of the TRIBE and HyperTRIBE lines that overlapped with the Xing et al. [Xing et al., 2015] dataset of SR45-associated RNA transcripts (SARs) that are either directly or indirectly associated with SR45.

We then performed a GO analysis with targets from the HyperTRIBE line that overlapped with the RIP-seq. The TRIBE line had fewer overlapped targets hence did not provide many GO terms so the results will not be discussed. Most enriched terms in all HyperTRIBE lines are related to “cellular processes” and “catalytic activity” (Figure 41). Unique targets of HyperTRIBE-SR45 long showed the same GO terms as the previous GO analysis of “cell death” and “chromosome segregation” which implicate SR45 long may serve a potential role in cell cycle processes. The targets of HyperTRIBE SR45-short contain various unique GO terms such as “process utilizing autophagic mechanism”, “microtubule-based process, and “actin filament-based process”. These GO categories were shown in the previous GO analyses, suggesting that SR45.2 also may regulate different aspects of the cell cycle compared to SR45.1. Even though 13% of the unique targets from HyperTRIBE-SR45 long and 9.9% of the unique targets from HyperTRIBE-SR45 short isoform fall under the *“*response to stimulus” GO term, there were no GO terms that were commonly found in both lines, suggesting that SR45 long and SR45 short isoforms regulate responses to different stress signals.

We also inspected whether the identified RNA targets are derived from intron-containing/less genes and if the transcript itself is alternatively spliced. Most of the RNA transcripts are from intron-containing genes, and, surprisingly, ⩾50% are alternatively spliced in the HyperTRIBE lines (long, short, and both). In the TRIBE lines, among the unique targets found in the TRIBE-SR45 short isoform line, 67% were alternatively spliced and 24% of the RNA targets unique to the TRIBE-SR45 long line were alternatively spliced. These results reflect the high likelihood that these intron-containing genes are potentially direct targets of SR45, thus implying the importance of SR45 in the regulation of splicing.

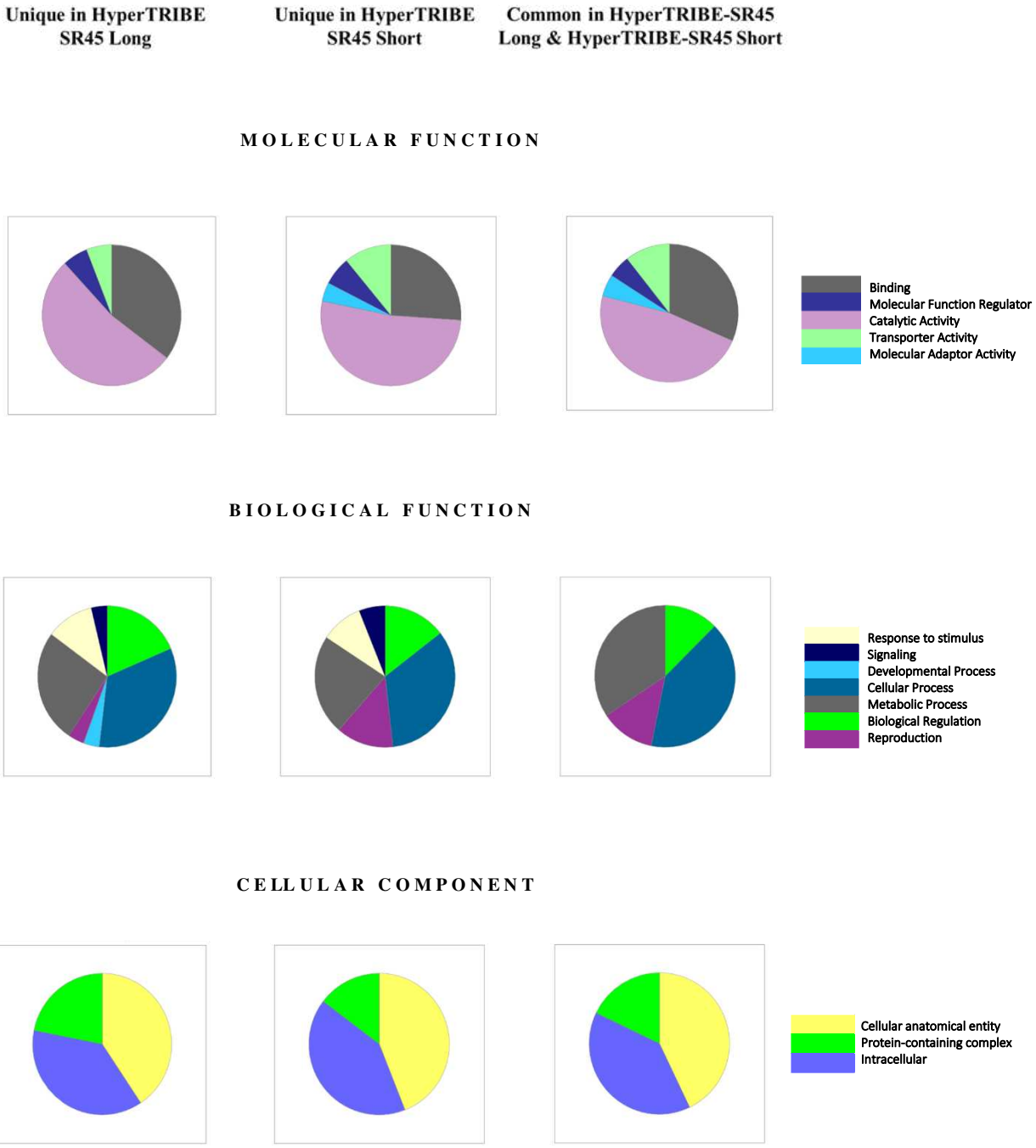


Figure 41. GO enrichment analysis of targets identified using the HyperTRIBE method that overlapped with SR45 RIP-seq. Pie charts depict each of three major GO aspects: molecular function biological process, and cellular component.

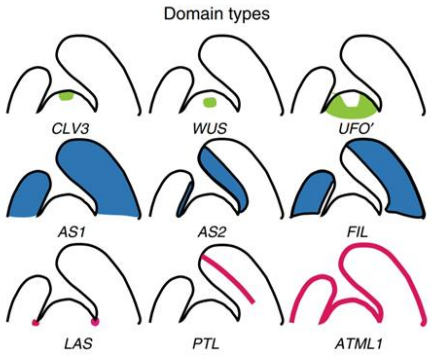
TRIBE lines also displayed a few targets that were from intronless genes. Thus, our studies strengthen the point that SR45's major involvement is in regulating splicing and post-transcriptional events in intronless genes.

Target genes from TRIBE and HyperTRIBE lines are expressed in meristem tissues

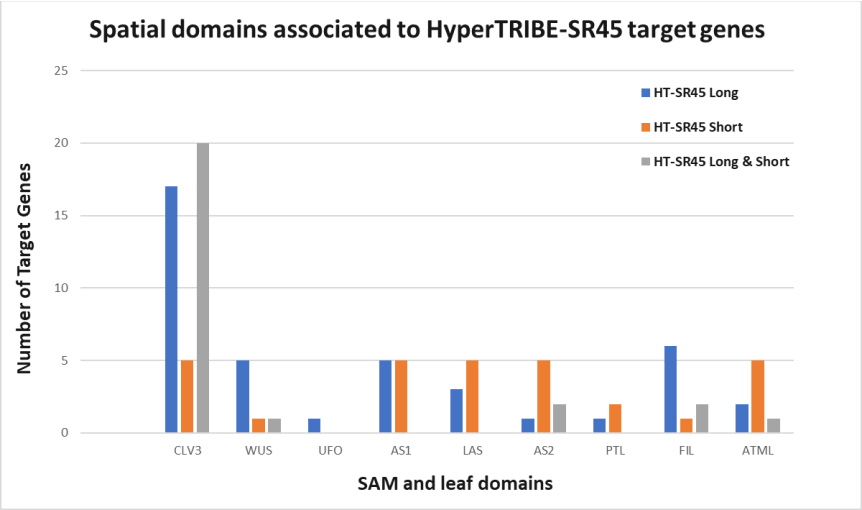
Meristems are the main reservoirs for undifferentiated stem cells. These daughter cells from these tissues continuously divide and the descendent cells differentiate into different cell types, tissues, and organs of the plant [Perelli et al., 2012]. *CLAVATA3 (CLV3)* is known for limiting the size of the stem cell niche, hence studies have shown that *CLV3* mutants exhibit a bigger stem cell zone and larger meristem tissues. This effect leads to larger organ primordial size and an increase in the number of flower organs [Szczęsny et al., 2009]. In Ali et al.'s paper, they found that growing *sr45-1 mutants* under short-day conditions resulted in a range of flowers with an altered number of petals and stamens [Ali et al., 2007]. Evidence has indicated that alternative splicing heavily shapes and regulates early plant development stages, so there is a possibility that SR45 protein, a crucial splicing regulator, may directly or indirectly come in contact with transcripts that express in meristematic cells, which could lead to these flowering defects that *CLV3* mutants are also displaying [Szakonyi and Duque, 2018].

To further identify if any of the SR45 RNA targets identified in our study are meristem specific, we compared each HyperTRIBE-SR45 isoform dataset with a published dataset of genes that are enriched in different domains of meristem [Tian et al., 2019]. Tian et al. released a list of domain-specifically expressed genes from the shoot apical meristem (SAM) and leaf domains. *CLAVATA3 (CLV3), UNUSUAL FLORAL ORGANS (UFO),* and *WUSCHEL (WUS)* promoters were used to label major components of the SAM (Figure 42A). For leaf domains, the

(A)



(B)



(C)

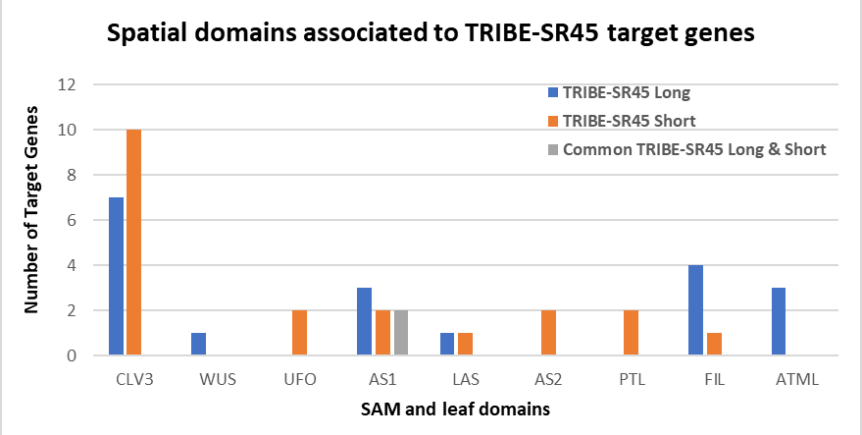


Figure 42. Expression of the number of SR45 direct targets identified in HyperTRIBE and TRIBE lines in different domains of shoot apical meristem. A) A schematic diagram showing the different spatial domains in the shoot apical meristem of plants [Tian et al., Nature Communications, 2019]. B) Target genes from (B) HyperTRIBE and (C) RNA targets expressed in specific domains of the shoot apical meristem (SAM) and leaf domains. These were identified by comparing SR45 RNA targets to Tian et al.'s dataset.

promoters used to assign the adaxial and abaxial cells were ASYMMETRIC LEAVES2 (AS2) and FILAMENTOUS FLOWER (FIL). The ASYMMETRIC LEAVES1 (AS1) promoters were utilized to label the entire leaf. Moreover, MERISTEM LAYER1 (ATML1) promoters were used to characterize the epidermal cells, LATERAL SUPPRESSOR (LAS) primers to define the domains at the adaxial boundary of leaf primordia, and the PETAL LOSS (PTL) promoters to express leaf margin cells.

The dataset from the TRIBE and HyperTRIBE-SR45 lines with a threshold of 1% editing were analyzed for their expression in different SAM and leaf domains (Figure 42B, C). Interestingly, 17 RNA targets found only in SR45.1 and 20 genes found in both SR45.1 and SR45.2 in the HyperTRIBE are expressed in the CLV3 domain (Figure 42B). A high number of targets identified in TRIBE are also expressed in the CLV3 domain (Figure 42C). Targets identified in both lines were expressed in other domains also (Figure 42B and C). However, *WUS* and *UFO* domains displayed the expression of few SR45 targets identified in both HyperTRIBE-SR45 long and HyperTRIBE-SR45 short lines. Tian et al. have reported enrichment of domain-specific alternative splicing events in the *CLV3, WUS*, and *LAS* domains, thus implying roles of alternative splicing in these specific domains. Based on our analysis, SR45.1 and SR45.2 may regulate alternative splicing of specific genes in the SAM domains, more specifically in *CLV3*, which also further supports SR45 role in regulating shoot growth, proper leaf development, and floral meristem development.

DISCUSSION

Application of TRIBE/HyperTRIBE method to identify targets of an RBP in plants

Genetic studies have shown that the SR45 splicing factor plays a key role in regulating plant development and stress responses [Ali et al., 2007; Albaqami et al., 2019; Zhang et al., 2017]. Using genetic studies and an *in vitro* splicing assay, it has been shown that SR45 is a bona fide splicing factor. The pre-mRNA of the *SR45* gene undergoes alternative splicing and generates two mRNAs (*SR45.*1 -long isoform and *SR45.2* – short isoform) that encode proteins differing in 8 amino acids [Palusa et al., 2007; Ali et al., 2007; Zhang and Mount, 2009]. Complementation studies with *sr45* mutant have revealed that each isoform has distinct functions in development and stress responses [Zhang and Mount, 2009; Albaqami et al., 2019]. The biological importance of each SR45 isoform has been shown during different stages of plant development and in biotic and abiotic stress responses, but the mechanistic understanding of isoform functions remains unknown. One approach to gain insights into the function of SR45 splice isoforms is to identify the direct RNA targets of each splice isoform. Towards this goal, I have employed a recently developed novel *in vivo* assay called TRIBE/HyperTRIBE and analyzed RNA-seq to identify direct RNA targets of SR45 isoforms. As discussed in the introduction this method is expected to overcome many of the limitations, especially with plants, with other mainstream methods to identify *in vivo* targets of RBPs.

We were able to successfully generate TRIBE and HyperTRIBE constructs with each SR45 isoform attached to the dADARcd by a linker sequence (Figure 25) and introduced them into an *sr45* mutant to generate transgenic lines. The transgenic lines were confirmed to express the introduced constructs in transgenic lines (Figure 28). The functionality of SR45-dADAR fusions was verified by the complementation of *sr45* mutant phenotypes (Figures 26 and 27).

Predicted phenotypes were observed for each construct and the expression analysis in transgenic lines of TRIBE and HyperTRIBE constructs were confirmed by RT-PCR using primers to amplify specific regions of the introduced genes. Overall, the RT-PCR analyses demonstrated that we are using a null mutant hence the observed *sr45-1* phenotypes result from the full loss of function of SR45 and that heterologous expression of SR45 does indeed restore that function.

TRIBE/HyperTRIBE methods work in plants but with low editing efficiency

For the bioinformatic analyses, we tailored the original HyperTRIBE pipeline for the Arabidopsis dataset by adjusting the parameters and implementing Arabidopsis-specific databases in the workflow. Consequently, we were able to effectively integrate the pipeline with our RNA-seq data from our plant lines. Utilizing this bioinformatic pipeline, we used the edit sites present in all triplicates to obtain a high confidence set of RNA targets of SR45 isoforms from TRIBE and HyperTRIBE lines. We also removed any background noise that may have derived from the ADARcd by subtracting any editing found in the controls (i.e., transgenic lines expressing only the dADARcd) from the rest of the TRIBE and HyperTRIBE-*SR45* fusion lines.

We attempted to use the original editing threshold that was initially used to filter the *Drosophila* dataset with our Arabidopsis dataset to create a list of bona fide edit sites but found that the list of transcripts was much smaller compared to the list of targeted genes from the *Drosophila* dataset. We utilized three different editing thresholds (at least 10%, at least 5% editing, or at least 1% editing) since a substantial proportion of editing events met these parameters. We found a significant increase in the editing sites when implementing the least stringent threshold, at least 1% editing, indicating that the majority of these edit coordinates seem to have an editing efficiency between 1-5%.

When we compared the edited transcripts from SR45 isoforms from the HyperTRIBE lines to the TRIBE lines, we found that the HyperTRIBE showed a larger pool of editing sites. The reason is that the HyperTRIBE ADARcd was able to edit more nearby adenosines of the RBP targets and were consistently present in all triplicates whereas the TRIBE lines had more transient editing sites that were not reproducible thus creating a smaller pool of editing events. Based on our results, we conclude that TRIBE/HyperTRIBE method can be applied to plants and that, as in animals, HyperTRIBE has higher efficiency in editing as compared to TRIBE. Hence, the HyperTRIBE method could be used to identify targets of RBP in plants also. However, the efficiency of HyperTRIBE was found to be not as high as in *Drosophila*. This could be due to multiple reasons. First, the RBP used in *Drosophila* is known to bind directly to a large number of targets and the RBP we used (SR45) may have fewer direct targets naturally. Part of the SR45 impact on splicing and other post-transcriptional processes could be indirect as SR45 is known to interact with multiple RBPs including several SR proteins [Reddy, 2007; Golovkin and Reddy, 1999]. Second, it is possible that *Drosophila* ADAR used in our study is not as efficient in editing heterologous systems as in plants. The use of a plant ADAR could improve editing efficiency. Although plants have at least four ADARs, all characterized ones are known to target tRNAs [Delannoy et al., 2009; Zhou et al., 2014; Zhou et al., 2013]. Finally, the level of fusion protein level could be different between *Drosophila* and in our transgenic lines. The differences in editing efficiency of dADAR between *Drosophila* and plants is not due to RNA-seq read depth as the number of reads per sample in Arabidopsis (46-52 million) are higher than what was used in Drosophila (20 million/sample).

SR45 isoforms have common and unique RNA targets

Previous studies to complement the mutant phenotypes have shown that some phenotypes (e.g., glucose sensitivity) can be rescued by both splice isoforms whereas other phenotypes such as root growth, flower developmental defects, and salt sensitivity can be rescued by only one isoform [Carvalho et al., 2010; Albaqami et al., 2019; Zhang and Mount, 2009]. As described in the "Results" section, each SR isoform has shared and unique RNA targets irrespective of the editing threshold (Figure 33). Based on the complementation studies, this is what is expected in terms of isoform targets. The unique targets of each isoform are likely to function in processes that are complemented with that isoform. Based on the GO results with targets of long and short isoform, SR45 long was found to play a major role in fine-tuning different stress responses while SR45 short was shown to have an association with "structural molecule activity" term. Interestingly, we find that both SR45 long and short are heavily involved in the cell cycle but most likely regulating different aspects of the process. However, both SR45 long and short isoforms shared targets involved in cell transport and plant development/reproduction. Further biochemical studies such as *in vitro* RNA binding with isoform-unique and isoform-common targets are needed to further validate these editing results. In addition, future functional studies with some of these SR45 targets are warranted to further elucidate their role in SR45 isoform-mediated responses. We will also be performing motif analysis with unique and common targets to identify putative *cis*-elements (motifs) involved in the binding of SR45 isoforms.

SR45 bound mostly to alternatively spliced transcripts with some transcripts from intron-less genes

Many of the targets identified in HyperTRIBE lines were also found in the RNA targets list identified in a RIP-seq study with SR45 [Xing et al., 2015], confirming that these are likely

direct targets of SR45 (Figure 33). After identifying the target transcripts that matched with RNA targets identified in a previous RIP-seq dataset, we investigated whether these identified transcripts are alternatively spliced or if they are from intron-less/containing genes. Interestingly, over 50% of the HyperTRIBE SR45 long and short were derived from spliced transcripts. Even though the majority of the transcripts were from intron-containing genes, there were few targets from intronless genes, suggesting a direct involvement of SR45 in processes other than splicing. Overall, our studies have broadened our understanding of differences in RNA targets between SR45 splice isoforms and demonstrated the utility of the TRIBE/HyperTRIBE method for identifying RNA targets of RBPs in plants.

The potential connection between meristem activity and SR45 function

Plants have high developmental plasticity and have the capability to cope with diverse stressful conditions, partly because they contain stem cell niches that respond to environmental signals [Aichinger et al., 2012]. This causes the stem cells to reprogram and commit to a cell lineage that will sustain the plant growth. Thus, identifying the vital regulators that modulate the homeostasis of these stem cell niches can decode another layer of regulating meristem growth and tissue specialization. The status of the chromatin landscape also has a crucial impact on regulating cell divisions and defining cellular tissue fates. *WUS*, a major transcription factor that induces stem cells, can repress certain differentiation genes, and ultimately affect chromatin remodeling and stem cell fate [Yadav et al., 2011]. So, activities that control and regulate these important transcription factors may have an indirect influence on stem cell activity.

CLV3 expression contributes to the patterning of a stem cell zone within the meristem, and because SR45 facilitates RNA splicing and has a prominent influence in early plant developmental stages, there may be a possibility that SR45 may mediate some of the *CLV3*

functions. Utilizing the TRIBE and HyperTRIBE method, we showed that both SR45 isoforms had more targets that are expressed in meristem-specific cells, specifically in the *CLV3* domain, and other domains involved in leaf development. These results support the known role of SR45 in seedling growth, the lack of which results in a drastic reduction in growth and abnormal small-sized leaves.

However, further analysis with other approaches such as direct *in vitro* binding assays with these targets and/or HyperTRIBE RNA-editing analysis with RNA from the *CLV3* domain cells (and other SAM domains) are needed to illuminate the role of SR45 targets in meristem and to understand the effect of this interaction on the function of these RNAs. To demonstrate whether SR45 and meristem activity are interrelated, we can utilize different endogenous tissue-specific promoters, including a promoter that will express only in meristem tissues, to express our TRIBE and HyperTRIBE fusion proteins. By obtaining editing results from each domain, we can comparatively observe whether the mapping of the SR45 target sites is drastically distinguishable among cells in different niches of the meristem. We can also perform an *in-vitro* protein-RNA interaction assay like EMSA (RNA electrophoretic mobility assay) to observe whether SR45 directly binds to *CLV3* RNA transcripts.

Another approach is to identify the target genes in our dataset that have a role in cell fate or meristem patterning. These targets would be considered candidate meristem genes to further our studies and prove our theory that SR45 functions in meristem tissues. A differential gene expression analysis on *sr45-1* mutants and *clv3-1* mutants can be done to detect whether the candidate genes are a potential target of SR45 and associated with *CLV3* if the expression level of these candidate genes is significantly affected. To examine if the candidate genes are meristem-specific, we should expect a loss of function mutation of the candidate genes to

produce the specific phenotypes related to meristem. We can also investigate the association of these candidate genes with SR45 or *CLV3* by observing if affected phenotypes were rescued in *sr45-1* or *clv3-1* mutant line when complemented with the expression of the candidate genes.

Does the chimeric nature of fusion protein (RBP-cADARd) impact the RBP function?

The fusion of ADARcd to an RBP has raised some concern that this fusion would adversely impact the function of RBP, ADARcd, or both. This is a valid concern that needs to be addressed with some functional assays with the fusion (e.g., complementing a mutant phenotype or performing biochemical assays with RBD with and without fusion) to rule out this possibility. Xu et al. proved that this method works successfully in editing in S2 tissue culture cells and in identifying cell-context dependent RBP-RNA interactions in fly neurons [Xu et al., 2018]. To deliver similar functional effects of *Drosophila* HyperTRIBE ADARcd (dADARcd) in humans, they incorporated the mutational residue corresponding to the E488 glutamate of dADARcd onto the human ADARcd (hADARcd). Overall, HyperTRIBE has been tested in mammalian systems like mouse hematopoietic and progenitor cell lines and human AML cell lines and reported promising results [Nguyen et al., 2020]. However, in one study, they attempted to use *Drosophila* ADARcd as its editing component and linked it with their human RBP in human prostate cancer cell lines. Unfortunately, they found that even though this fusion protein displayed normal expression levels, but poor editing efficiency was reported [Jin et al, 2020]. One possibility for this is that the dADARcd may not have reacted well with the mammalian cellular environment where different temperature, acidity, and different interacting proteins are introduced to the domain, and as a consequence causing editing to be not as effective.

Comparing the editing events that occurred from the TRIBE and HyperTRIBE constructs, we did find that the HyperTRIBE edited more adenosines as well, providing more reproducible

and reliable results in their triplicates than the TRIBE lines. In our study, we did use the dADARcd as our editing enzyme for our chimeric protein which may address the low level of editing marks found in both of our Arabidopsis TRIBE and HyperTRIBE lines. The sequence and the mutated glutamate of dADARcd and hADARcd are highly conserved which makes this approach feasible for mammalian systems, yet there is currently no equivalent that resembles in plants. Identifying Arabidopsis gene homologues or counterparts to the dADARcd in plants is the initial step in optimizing the editing efficiency and ultimately successfully adapting the HyperTRIBE approach with high editing efficiency in plant systems.

The human ADARcd and the *Drosophila* ADARcd both has a preference to only edit adenosines by neighboring UAG sequence elements, but the HyperTRIBE ADARcd was able to lose that sequence bias and reduce false negative resutlts [Rahman et al., 2018]. We do not know if the deaminating adenosines that we detected were enriched in UAG elements.

Our results suggest that the HyperTRIBE method may be more optimal in *Drosophila* compared to plants. For each RNA library in our TRIBE and HyperTRIBE lines, we found a drastic reduction of edits per mRNA compared to the published data using *Drosophila* cells. However, our study does not serve as a proper bona fide comparison in concluding that the HyperTRIBE technique has a better editing performance in *Drosophila* than plants. SR45 is known to regulate splicing and other post-transcriptional processes [Xu et al., 2018]. Because these RBPs have different biological roles and are involved in different pathways, it's most likely that these proteins have a different number of binding targets and therefore different editing patterns. A protein with diversified roles is more likely to obtain more binding substrates, so the editing efficiency depends on the role of that protein. Thus, it may not be wise to compare their data to our protein of interest.

Jin et al. applied the HyperTRIBE technique in flies and mammals, using 4E-BP [eukaryotic initiation factor 4E (eIF4E)–binding protein] as the RBP, and found 711 h4E-BP1 targets in human cells and 968 d4E-BP targets in fly cells and overall identified 180 sets of targeted homologs between them [Jin et al., 2020]. Because the 4E-BP protein interacted with similar target transcripts in *Drosophila* and mammals, the HyperTRIBE method successfully characterized the 4E-BP protein and affirmed that this RBP had conserved mechanisms and functions in both systems. To address if HyperTRIBE works efficiently in plants, a HyperTRIBE fusion construct with a plant counterpart of a *Drosophila* RBP should be developed and expressed in plants and be performed in a head-to-head comparison with the *Drosophila* RBP chimeric protein expressed in fly cells. This experiment will ultimately deduce whether HyperTRIBE can be an adaptive, accessible tool to identify RBP-RNA interactions efficiently in different cells and systems.

Potential problems in using a constitutive promoter for expression of the fusion constructs

In this study, a constitutive CaMV35S promoter was used to drive the expression of each of our TRIBE and HyperTRIBE fusion constructs. Originally in the TRIBE and HyperTRIBE protocols, they performed their experiments with two types of promoters for the cultured neuron cells and in *Drosophila*: a copper inducible metallothionein (MT) promoter and the Gal4–upstream activating sequence (UAS) system [McMahon et al., 2016; Rosbash et al., 2018]. However, the overexpression of the fusion protein constructs may raise challenges of perturbing different regulatory pathways or affecting the typical cell behavior. A higher expression of the protein of interest could potentially misconstrue results since an oversaturation of the RBP may bind to non-specific targets when the actual RBP substrates are being occupied.

But in a previous analysis, studying MSI2, a regulator protein for leukemia stem cells, addressed this issue by performing a differential gene expression analysis of the cells expressing the MSI2-dADARcd fusion and comparing the results with a control vector [Nguyen et al., 2020]. Based on their investigation, they found little change in the transcriptome expression level of its target genes (MOLM13, LSKs, and LSCs) in MSI2-ADA cells. For future approaches, a native promoter of Arabidopsis can be used to express each of the TRIBE and HyperTRIBE lines to avoid any complications that can occur with a constitutive promoter. The expression of the *SR45* fusion constructs from the endogenous promoter can be further verified by comparing gene expression between the lines with a constitutive promoter and a native promoter.

The potential connection between circRNAs and SR45 function

As mentioned above, circRNA plays an important regulatory role in post-transcriptional RNA silencing by sequestering binding sites of mRNA substrates. This miRNA-mediated gene regulation modulates gene expression and alternative splicing by altering the abundance of functional transcripts and also regulates the expression of stimulus-responsive genes when external stress signals are introduced [Zhang et al., 2020]. Like the SR45 protein, circRNAs contribute to plant growth and development by altering gene expression in a temporal-, spatial-, and developmental stage. This leads to the question of interplay between these two components i.e., does SR45, a known splicing factor, have any role in backsplicing and regulating the levels of circRNAs. One way to address this is to perform circRNAs analysis globally in wild-type, *sr45* mutant, and *sr45* mutant complemented with each isoform. This is likely to provide the role of SR45 in general and individual splice isoforms in particular in the biogenesis of circRNAs. Although this is a straightforward and feasible experiment, such studies have not been performed so far.

A more direct approach to detect a modified base in direct RNA reads obtained with Oxford Nanopore sequencing

Alternatively, we can utilize the same concept of the HyperTRIBE method, where we incorporate some type of editing enzyme domain to our fusion protein construct and use high-throughput sequencing and bioinformatics tools to identify the edited sites of the RBP. Due to inefficient cross-linking from immunoprecipitation assays in plant studies, it is desirable to seek an alternative method that is antibody independent similar to the HyperTRIBE method. There have been pioneering studies in performing nanopore direct RNA (dRNA) sequencing to detect modified base signals directly on RNA. *De novo* identification of methyl-adenosine was found attainable when Gao et al. constructed a novel pipeline, Nanom6A, using XGBoost algorithm, to identify the quantification of m6A sites on individual transcripts [Gao et al., 2021]. Using known modified m6A sites in stem-differentiated xylems of Arabidopsis, Nanom6A successfully predicted about 91%-96% modified and unmodified sites from individual transcripts and also discovered that transcripts with different polyadenylation lengths differ in the distribution of m6A ratio. Because of these advances, we can potentially construct a chimeric protein with our protein of interest and an adenine methylase enzyme in replacement to the deaminase enzyme. Instead of seeking for conversions of adenine to guanine, this method can present a transcriptome-wide identification and quantification of methylated marks to map the RBP sites on the target RNAs. In addition, several groups including Dr. Asa's group at CSU are working on developing computational tools to identify different modified RNA bases including inosine in dRNA reads, which will permit the identification of edited RNAs in HyperTRIBE lines as the sequences are read [Zhao et al., 2019; Reddy et al., 2020].

REFERENCES

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J.M., del Pozo, A., Vázquez, J., Valencia, A., and Tress, M.L. (2015). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Comput Biol **11**: e1004325.

Aichinger, E., Kornet, N., Friedrich, T., and Laux, T. (2012). Plant Stem Cell Niches. Annual Review of Plant Biology **63**: 615–636.

Akua, T. and Shaul, O. (2013). The Arabidopsis thaliana MHX gene includes an intronic element that boosts translation when localized in a 5' UTR intron. J Exp Bot **64**: 4255–4270.

Albaqami, M., Laluk, K., and Reddy, A.S.N. (2019). The Arabidopsis splicing regulator SR45 confers salt tolerance in a splice isoform-dependent manner. Plant Mol Biol **100**: 379–390.

Ali, G.S., Golovkin, M., and Reddy, A.S.N. (2003). Nuclear localization and in vivo dynamics of a plant-specific serine/arginine-rich protein. Plant J **36**: 883–893.

Ali, G.S., Palusa, S.G., Golovkin, M., Prasad, J., Manley, J.L., and Reddy, A.S.N. (2007). Regulation of plant developmental processes by a novel splicing factor. PLoS One **2**: e471.

Ali, G.S., Prasad, K.V.S.K., Hanumappa, M., and Reddy, A.S.N. (2008). Analyses of in vivo interaction and mobility of two spliceosomal proteins using FRAP and BiFC. PLoS One **3**: e1953.

Ali, G.S. and Reddy, A.S.N. (2006). ATP, phosphorylation and transcription regulate the mobility of plant splicing factors. J Cell Sci **119**: 3527–3538.

Ausin, I., Greenberg, M.V.C., Li, C.F., and Jacobsen, S.E. (2012). The splicing factor SR45 affects the RNA-directed DNA methylation pathway in Arabidopsis. Epigenetics **7**: 29–33.

Barta, A., Kalyna, M., and Lorković, Z.J. (2008). Plant SR proteins and their functions. Curr Top Microbiol Immunol **326**: 83–102.

Barta, A., Kalyna, M., and Reddy, A.S.N. (2010). Implementing a rational and consistent nomenclature for serine/arginine-rich protein splicing factors (SR proteins) in plants. Plant Cell **22**: 2926–2929.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell **136**: 215–233.

Beyer, A.L., Christensen, M.E., Walker, B.W., and LeStourgeon, W.M. (1977). Identification and characterization of the packaging proteins of core 40S hnRNP particles. Cell **11**: 127–138.

Biswas, J., Rahman, R., Gupta, V., Rosbash, M., and Singer, R.H. (2020). MS2-TRIBE Evaluates Both Protein-RNA Interactions and Nuclear Organization of Transcription by RNA Editing. iScience **23**: 101318.

Boothby, T.C., Zipper, R.S., van der Weele, C.M., and Wolniak, S.M. (2013). Removal of Retained Introns Regulates Translation in the Rapidly Developing Gametophyte of Marsilea vestita. Developmental Cell **24**: 517–529.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. **24**: 1774–1786.

Braunschweig, U., Gueroussov, S., Plocik, A.M., Graveley, B.R., and Blencowe, B.J. (2013). Dynamic integration of splicing within gene regulatory pathways. Cell **152**: 1252–1269.

Brown, J.W., Feix, G., and Frendewey, D. (1986). Accurate in vitro splicing of two pre-mRNA plant introns in a HeLa cell nuclear extract. EMBO J **5**: 2749–2758.

Burjoski, V. and Reddy, A.S.N. (2021). The Landscape of RNA-Protein Interactions in Plants: Approaches and Current Status. Int J Mol Sci **22**: 2845.

Carvalho, R.F., Carvalho, S.D., and Duque, P. (2010). The plant-specific SR45 protein negatively regulates glucose and ABA signaling during early seedling development in Arabidopsis. Plant Physiol **154**: 772–783.

Carvalho, R.F., Szakonyi, D., Simpson, C.G., Barbosa, I.C.R., Brown, J.W.S., Baena-González, E., and Duque, P. (2016). The Arabidopsis SR45 Splicing Factor, a Negative Regulator of Sugar Signaling, Modulates SNF1-Related Protein Kinase 1 Stability. The Plant Cell **28**: 1910–1925.

Chodavarapu, R.K. et al. (2010). Relationship between nucleosome positioning and DNA methylation. Nature **466**: 388–392.

Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein–RNA regulation in living cells. WIREs RNA **1**: 266–286.

Day, I.S., Golovkin, M., Palusa, S.G., Link, A., Ali, G.S., Thomas, J., Richardson, D.N., and Reddy, A.S.N. (2012). Interactions of SR45, an SR-like protein, with spliceosomal proteins and an intronic sequence: insights into regulated splicing. Plant J **71**: 936–947.

Delannoy, E., Le Ret, M., Faivre-Nitschke, E., Estavillo, G.M., Bergdoll, M., Taylor, N.L., Pogson, B.J., Small, I., Imbault, P., and Gualberto, J.M. (2009). Arabidopsis tRNA Adenosine Deaminase Arginine Edits the Wobble Nucleotide of Chloroplast tRNAArg(ACG) and Is Essential for Efficient Chloroplast Translation. The Plant Cell **21**: 2058–2071.

Duque, P. (2011). A role for SR proteins in plant stress responses. Plant Signal Behav **6**: 49–54.

Eversole, A. and Maizels, N. (2000). In Vitro Properties of the Conserved Mammalian Protein hnRNP D Suggest a Role in Telomere Maintenance. Mol Cell Biol **20**: 5425–5432.

Ezkurdia, I., Rodriguez, J.M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A., and Tress, M.L. (2015). Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res **14**: 1880–1887.

Filichkin, S., Priest, H.D., Megraw, M., and Mockler, T.C. (2015). Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. Curr Opin Plant Biol **24**: 125–135.

Filichkin, S.A. and Mockler, T.C. (2012). Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. Biol Direct **7**: 20.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.-K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res **20**: 45–58.

Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M.V., Reddy, A.S.N., and Gu, L. (2021). Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. Genome Biol **22**: 22.

Giono, L.E. and Kornblihtt, A.R. (2020). Linking transcription, RNA polymerase II elongation and alternative splicing. Biochemical Journal **477**: 3091–3104.

Göhring, J., Jacak, J., and Barta, A. (2014). Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in Arabidopsis. Plant Cell **26**: 754–764.

Golovkin, M. and Reddy, A.S. (1999). An SC35-like protein and a novel serine/arginine-rich protein interact with Arabidopsis U1-70K protein. J Biol Chem **274**: 36428–36438.

Gomes, E. and Shorter, J. (2019). The molecular language of membraneless organelles. Journal of Biological Chemistry **294**: 7115–7127.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature **495**: 384–388.

Harada, K., Yamada, A., Yang, D., Itoh, K., and Shichijo, S. (2001). Binding of a SART3 tumor-rejection antigen to a pre-mRNA splicing factor RNPS1: a possible regulation of splicing by a complex formation. Int J Cancer **93**: 623–628.

Hartmann, L., Wießner, T., and Wachter, A. (2018). Subcellular Compartmentation of Alternatively Spliced Transcripts Defines SERINE/ARGININE-RICH PROTEIN30 Expression. Plant Physiol **176**: 2886–2903.

Hartmuth, K. and Barta, A. (1986). In vitro processing of a plant pre-mRNA in a HeLa cell nuclear extract. Nucleic Acids Res **14**: 7513–7528.

Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol **18**: 637–650.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein–RNA interactions at nucleotide resolution. Methods **65**: 274–287.

Jabre, I., Chaudhary, S., Guo, W., Kalyna, M., Reddy, A.S.N., Chen, W., Zhang, R., Wilson, C., and Syed, N.H. (2021). Differential nucleosome occupancy modulates alternative splicing in Arabidopsis thaliana. New Phytologist **229**: 1937–1945.

Jeck, W.R. and Sharpless, N.E. (2014). Detecting and characterizing circular RNAs. Nature Biotechnology **32**: 453–461.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA **19**: 426.

Kalyna, M. et al. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. Nucleic Acids Res **40**: 2454–2469.

Kalyna, M. and Barta, A. (2004). A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? Biochem Soc Trans **32**: 561–564.

Kalyna, M., Lopato, S., and Barta, A. (2003). Ectopic Expression of atRSZ33 Reveals Its Function in Splicing and Causes Pleiotropic Changes in Development. MBoC **14**: 3565–3577.

Keegan, L.P., Leroy, A., Sproul, D., and O'Connell, M.A. (2004). Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. Genome Biol **5**: 209.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene **514**: 1–30.

Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P., Lührmann, R., Garcia-Blanco, M.A., and Manley, J.L. (1994). Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. NATURE **368**: 119–124.

Lapointe, C.P., Wilinski, D., Saunders, H.A.J., and Wickens, M. (2015). Protein-RNA networks revealed through covalent RNA marks. Nature Methods **12**: 1163–1170.

Lazar, G. and Goodman, H.M. (2000). The Arabidopsis splicing factor SR1 is regulated by alternative splicing. Plant Mol Biol **42**: 571–581.

Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001). The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J **20**: 4987–4997.

Li, S., Wang, Y., Zhao, Y., Zhao, X., Chen, X., and Gong, Z. (2020). Global Co-transcriptional Splicing in Arabidopsis and the Correlation with Splicing Regulation in Mature RNAs. Molecular Plant **13**: 266–277.

Ling, Y. et al. (2018). Thermopriming triggers splicing memory in Arabidopsis. Journal of Experimental Botany **69**: 2659–2675.

Listerman, I., Sapra, A.K., and Neugebauer, K.M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nature Structural & Molecular Biology **13**: 815–822.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R., and Barta, A. (1999). atSRp30, one of two SF2/ASF-like proteins from Arabidopsis thaliana, regulates splicing of specific plant genes. Genes Dev **13**: 987–1001.

Lopez-Molina, L., Mongrand, S., and Chua, N.H. (2001). A postgermination developmental arrest checkpoint is mediated by abscisic acid and requires the ABI5 transcription factor in Arabidopsis. Proc Natl Acad Sci U S A **98**: 4782–4787.

Lorković, Z.J. and Barta, A. (2004). Compartmentalization of the splicing machinery in plant cell nuclei. Trends in Plant Science **9**: 565–568.

Lorković, Z.J., Hilscher, J., and Barta, A. (2004). Use of Fluorescent Protein Tags to Study Nuclear Organization of the Spliceosomal Machinery in Transiently Transformed Living Plant Cells. Mol Biol Cell **15**: 3233–3243.

Loyer, P., Trembley, J.H., Lahti, J.M., and Kidd, V.J. (1998). The RNP protein, RNPS1, associates with specific isoforms of the p34cdc2-related PITSLRE protein kinase in vivo. J Cell Sci **111 (Pt 11)**: 1495–1506.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell **144**: 16–26.

Lykke-Andersen, J., Shu, M.-D., and Steitz, J.A. (2001). Communication of the position of Exon-Exon junctions to the mRNA surveillance machinery by the protein RNPS1. Science **293**: 1836–1839.

M, L., Cf, M., and Mc, T. (2008). Genetic and epigenetic regulation of stem cell homeostasis in plants. Cold Spring Harb Symp Quant Biol **73**: 243–251.

Manley, J.L. and Krainer, A.R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev **24**: 1073–1074.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv Exp Med Biol **623**: 123–147.

Martins, S.B., Rino, J., Carvalho, T., Carvalho, C., Yoshida, M., Klose, J.M., de Almeida, S.F., and Carmo-Fonseca, M. (2011). Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3’ end of human genes. Nat Struct Mol Biol **18**: 1115–1123.

McMahon, A.C., Rahman, R., Jin, H., Shen, J.L., Fieldsend, A., Luo, W., and Rosbash, M. (2016). TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. Cell **165**: 742–753.

Memczak, S. et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature **495**: 333–338.

Meng, F., Zhao, H., Zhu, B., Zhang, T., Yang, M., Li, Y., Han, Y., and Jiang, J. (2021). Genomic Editing of Intronic Enhancers Unveils Their Role in Fine-Tuning Tissue-Specific Gene Expression in Arabidopsis thaliana. Plant Cell: koab093. **Meng, X., Chen, Q., Zhang, P., and Chen, M.** (2017). CircPro: an integrated tool for the identification of circRNAs with protein-coding potential. Bioinformatics **33**: 3314–3316.

Meng, X., Zhang, P., Chen, Q., Wang, J., and Chen, M. (2018). Identification and characterization of ncRNA-associated ceRNA networks in Arabidopsis leaf development. BMC Genomics **19**: 607.

Montiel-Gonzalez, M.F., Vallecillo-Viejo, I., Yudowski, G.A., and Rosenthal, J.J.C. (2013). Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed RNA editing. PNAS **110**: 18285–18290.

Moore, M.J., Zhang, C., Gantman, E.C., Mele, A., Darnell, J.C., and Darnell, R.B. (2014). Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. Nat Protoc **9**: 263–293.

Morton, M., AlTamimi, N., Butt, H., Reddy, A.S.N., and Mahfouz, M. (2019). Serine/Arginine-rich protein family of splicing regulators: New approaches to study splice isoform functions. Plant Science **283**: 127–134.

Nguyen, D.T.T. et al. (2020). HyperTRIBE uncovers increased MUSASHI-2 RNA binding activity and differential regulation in leukemic stem cells. Nat Commun **11**: 2026.

Palusa, S.G., Ali, G.S., and Reddy, A.S.N. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. Plant J **49**: 1091–1107.

Pandya-Jones, A. and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA **15**: 1896–1908.

Perilli, S., Di Mambro, R., and Sabatini, S. (2012). Growth and development of the root apical meristem. Current Opinion in Plant Biology **15**: 17–23.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., and Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports 8: 10950.

Reddy, A.S.N. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol **58**: 267–294.

Reddy, A.S.N. (2001). Nuclear Pre-mRNA Splicing in Plants. Critical Reviews in Plant Sciences **20**: 523–571.

Reddy, A.S.N. (2004). Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. Trends in Plant Science **9**: 541–547.

Reddy, A.S.N., Day, I.S., Göhring, J., and Barta, A. (2012a). Localization and Dynamics of Nuclear Speckles in Plants1. Plant Physiol **158**: 67–77.

Reddy, A.S.N., Huang, J., Syed, N.H., Ben-Hur, A., Dong, S., and Gu, L. (2020). Decoding co-/post-transcriptional complexities of plant transcriptomes and epitranscriptome using next-generation sequencing technologies. Biochem Soc Trans **48**: 2399–2414.

Reddy, A.S.N., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. Plant Cell **25**: 3657–3683.

Reddy, A.S.N., Rogers, M.F., Richardson, D.N., Hamilton, M., and Ben-Hur, A. (2012b). Deciphering the Plant Splicing Code: Experimental and Computational Approaches for Predicting Alternative Splicing and Splicing Regulatory Elements. Front. Plant Sci. **3:** 18.

Ren, Y. et al. (2018). Identification and characterization of circRNAs involved in the regulation of low nitrogen-promoted root growth in hexaploid wheat. Biological Research **51:** 43.

Rogers, K. and Chen, X. (2013). Biogenesis, Turnover, and Mode of Action of Plant MicroRNAs[OPEN]. Plant Cell **25**: 2383–2399.

Rose, A.B. (2008). Intron-mediated regulation of gene expression. Curr Top Microbiol Immunol **326**: 277–290.

Sakashita, E., Tatsumi, S., Werner, D., Endo, H., and Mayeda, A. (2004). Human RNPS1 and Its Associated Factors: a Versatile Alternative Pre-mRNA Splicing Regulator In Vivo. Molecular and Cellular Biology **24**: 1174–1187.

van Santen, V.L. and Spritz, R.A. (1987). Splicing of plant pre-mRN As in animal systems and vice versa. Gene **56**: 253–265.

Schwartz, A.M., Komarova, T.V., Skulachev, M.V., Zvereva, A.S., Dorokhov, I.L., and Atabekov, J.G. (2006). Stability of plant mRNAs depends on the length of the 3'-untranslated region. Biochemistry (Mosc) **71**: 1377–1384.

Szczęsny, T., Routier-Kierzkowska, A.-L., and Kwiatkowska, D. (2009). Influence of clavata3-2 mutation on early flower development in Arabidopsis thaliana: quantitative analysis of changing geometry. Journal of Experimental Botany **60**: 679–695.

Simpson, C.G., Thow, G., Clark, G.P., Jennings, S.N., Watters, J.A., and Brown, J.W.S. (2002). Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon. RNA **8**: 47–56.

Spector, D.L. and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb Perspect Biol **3**.

Sun, X. et al. (2016). Integrative analysis of Arabidopsis thaliana transcriptomics reveals intuitive splicing mechanism for circular RNA. FEBS Lett **590**: 3510–3516.

Szakonyi, D. and Duque, P. (2018). Alternative Splicing as a Regulator of Early Plant Development. Front. Plant Sci. **9:** 1174.

Tange, T.Ø., Nott, A., and Moore, M.J. (2004). The ever-increasing complexities of the exon junction complex. Curr Opin Cell Biol **16**: 279–284.

Thiebaut, F., Hemerly, A.S., and Ferreira, P.C.G. (2019). A Role for Epigenetic Regulation in the Adaptation and Stress Responses of Non-model Plants. Front Plant Sci **10:** 246.

Tian, C., Wang, Y., Yu, H., He, J., Wang, J., Shi, B., Du, Q., Provart, N.J., Meyerowitz, E.M., and Jiao, Y. (2019). A gene expression map of shoot domains reveals regulatory mechanisms. Nature Communications **10**: 141.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., and Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nature Structural and Molecular Biology **16**: 996–1003.

Tillemans, V., Dispa, L., Remacle, C., Collinge, M., and Motte, P. (2005). Functional distribution and dynamics of Arabidopsis SR splicing factors in living plant cells. Plant J **41**: 567–582.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci **42**: 98–110.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: A method for identifying protein–RNA interaction sites in living cells. Methods **37**: 376–386.

Vinocur, B. and Altman, A. (2005). Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Curr Opin Biotechnol **16**: 123–132.

Vogel, P., Schneider, M.F., Wettengel, J., and Stafforst, T. (2014). Improving Site-Directed RNA Editing In Vitro and in Cell Culture by Chemical Modification of the GuideRNA. Angewandte Chemie International Edition **53**: 6267–6271.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell **136**: 669–687.

Wachter, A., Rühl, C., and Stauffer, E. (2012). The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. Front Plant Sci **3**: 81.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell **136**: 701–718.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative Isoform Regulation in Human Tissue Transcriptomes. Nature **456**: 470–476.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature **402**: 832–835.

Xiao, M.-S., Ai, Y., and Wilusz, J.E. (2020). Biogenesis and Functions of Circular RNAs Come into Focus. Trends Cell Biol **30**: 226–240.

Xing, D., Wang, Y., Hamilton, M., Ben-Hur, A., and Reddy, A.S.N. (2015). Transcriptome-Wide Identification of RNA Targets of Arabidopsis SERINE/ARGININE-RICH45 Uncovers the Unexpected Roles of This RNA Binding Protein in RNA Processing. Plant Cell **27**: 3294–3308.

Xu, W., Rahman, R., and Rosbash, M. (2018). Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. RNA **24**: 173–182.

Yadav, R.K., Perales, M., Gruel, J., Girke, T., Jönsson, H., and Reddy, G.V. (2011). WUSCHEL protein movement mediates stem cell homeostasis in the Arabidopsis shoot apex. Genes Dev. **25**: 2025–2030.

Ye, C.-Y., Chen, L., Liu, C., Zhu, Q.-H., and Fan, L. (2015). Widespread noncoding circular RNAs in plants. New Phytol **208**: 88–95.

Zhang, X.-N., Mo, C., Garrett, W.M., and Cooper, B. (2014). Phosphothreonine 218 is required for the function of SR45.1 in regulating flower petal development in Arabidopsis. Plant Signal Behav **9**: e29134.

Zhang, X.-N. and Mount, S.M. (2009). Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. Plant Physiol **150**: 1450–1458.

Zhang, X.-N., Shi, Y., Powers, J.J., Gowda, N.B., Zhang, C., Ibrahim, H.M.M., Ball, H.B., Chen, S.L., Lu, H., and Mount, S.M. (2017). Transcriptome analyses reveal SR45 to be a neutral splicing regulator and a suppressor of innate immunity in Arabidopsis thaliana. BMC Genomics **18**: 772.

Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., Wang, J., and Fan, J. (2019). Crop Phenomics: Current Status and Perspectives. Front. Plant Sci. **10:** 714.

Zhou, W., Karcher, D., and Bock, R. (2014). Identification of Enzymes for Adenosine-to-Inosine Editing and Discovery of Cytidine-to-Uridine Editing in Nucleus-Encoded Transfer RNAs of Arabidopsis. Plant Physiology **166**: 1985–1997.

Zhou, W., Karcher, D., and Bock, R. (2013). Importance of adenosine-to-inosine editing adjacent to the anticodon in an Arabidopsis alanine tRNA under environmental stress. Nucleic Acids Research **41**: 3362–3372.

Zhu, D., Mao, F., Tian, Y., Lin, X., Gu, L., Gu, H., Qu, L., Wu, Y., and Wu, Z. (2020). The Features and Regulation of Co-transcriptional Splicing in Arabidopsis. Molecular Plant **13**: 278–294.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol **136**: 2621–2632.

APPENDIX

	UNIQUE TARGET GENES OF HAL		
GENE ID	**GENE DESCRIPTION**	**INTRONS (Y/N)**	**ALT. SPLICED (Y/N)**
AT1G05500	Calcium-dependent lipid-binding (CaLB domain) family protein(NTMC2T2.1)	Y	N
AT1G06240	diiron containing four-helix bundle family ferritin protein, putative (Protein of unknown function DUF455)(AT1G06240)	Y	N
AT1G13390	translocase subunit seca(AT1G13390)	Y	Y
AT1G16970	ATP-dependent DNA helicase 2 subunit Ku70-like protein(KU70)	Y	N
AT1G19330	histone deacetylase complex subunit(AT1G19330)	Y	Y
AT1G27150	Tetratricopeptide repeat (TPR)-like superfamily protein(AT1G27150)	Y	N
AT1G64670	alpha/beta-Hydrolases superfamily protein(BDG1)	Y	N
AT1G69850	nitrate transporter 1:2(NRT1:2)	Y	N
AT1G77930	Chaperone DnaJ-domain superfamily protein(AT1G77930)	Y	Y
AT1G79880	RNA recognition motif (RRM)-containing protein(La2)	Y	Y
AT2G17520	Endoribonuclease/protein kinase IRE1-like protein(IRE1A)	Y	N
AT2G21470	SUMO-activating enzyme 2(SAE2)	Y	Y
AT2G39340	SAC3/GANP/Nin1/mts3/eIF-3 p25 family(SAC3A)	Y	N
AT2G39930	isoamylase 1(ISA1)	Y	N
AT2G40960	Single-stranded nucleic acid binding R3H protein(AT2G40960)	Y	N
AT2G46930	Pectinacetylesterase family protein(AT2G46930)	Y	N
AT3G02710	ARM repeat superfamily protein(AT3G02710)	Y	N
AT3G09090	defective in exine formation protein (DEX1)(DEX1)	Y	Y
AT3G57170	N-acetylglucosaminyl transferase component family protein / Gpi1 family protein(AT3G57170)	Y	N
AT3G60410	hypothetical protein (DUF1639)(AT3G60410)	Y	Y
AT4G16442	Uncharacterized protein family (UPF0497)(AT4G16442)	Y	N
AT4G19710	aspartate kinase-homoserine dehydrogenase ii(AK-HSDH II)	Y	Y
AT4G20130	plastid transcriptionally active 14(PTAC14)	Y	N
AT4G30210	P450 reductase 2(ATR2)	Y	Y
AT5G11030	aberrant root formation protein(ALF4)	Y	Y
AT5G15540	PHD finger family protein(EMB2773)	Y	Y
AT5G16030	mental retardation GTPase activating protein(AT5G16030)	Y	Y
AT5G17440	LUC7 related protein(AT5G17440)	Y	N
AT5G18230	transcription regulator NOT2/NOT3/NOT5 family protein(AT5G18230)	Y	Y
AT5G18410	transcription activator(PIR121)	Y	Y
AT5G21930	P-type ATPase of Arabidopsis 2(PAA2)	Y	Y
AT5G22450	spectrin beta chain, brain(AT5G22450)	Y	N
AT5G23080	SWAP (Suppressor-of-White-APricot)/surp domain-containing protein(TGH)	Y	Y
AT5G54710	Ankyrin repeat family protein(AT5G54710)	Y	N
	Total of "Y":	34	16
	Total of "N":	0	18

	UNIQUE TARGET GENES OF HAS		
GENE ID	GENE DESCRIPTION	INTRONS (Y/N)	ALT. SPLICED (Y/N)
AT1G03550	Secretory carrier membrane protein (SCAMP) family protein(SCAMP4)	Y	N
AT1G04300	TRAF-like superfamily protein(AT1G04300)	Y	N
AT1G06700	Protein kinase superfamily protein(AT1G06700)	Y	Y
AT1G08190	vacuolar protein sorting 41(VPS41)	Y	N
AT1G08420	BRI1 suppressor 1 (BSU1)-like 2(BSL2)	Y	Y
AT1G08810	myb domain protein 60(MYB60)	Y	Y
AT1G09140	SERINE-ARGININE PROTEIN 30(SR30)	Y	Y
AT1G13450	Homeodomain-like superfamily protein(GT-1)	Y	Y
AT1G19660	Wound-responsive family protein(AT1G19660)	Y	Y
AT1G27000	GRIP/coiled-coil protein, putative (DUF1664)(AT1G27000)	Y	N
AT1G30400	multidrug resistance-associated protein 1(ABCC1)	Y	Y
AT1G34340	alpha/beta-Hydrolases superfamily protein(AT1G34340)	Y	N
AT1G47550	exocyst complex component sec3A(SEC3A)	Y	Y
AT1G48635	peroxin 3(PEX3)	Y	Y
AT1G50030	target of rapamycin(TOR)	Y	Y
AT1G53780	26S proteasome regulatory complex ATPase(AT1G53780)	Y	Y
AT1G54030	GDSL-like Lipase/Acylhydrolase superfamily protein(MVP1)	Y	N
AT1G54370	sodium hydrogen exchanger 5(NHX5)	Y	N
AT1G60160	Potassium transporter family protein(AT1G60160)	Y	N
AT1G61100	disease resistance protein (TIR class)(AT1G61100)	Y	Y
AT1G65580	Endonuclease/exonuclease/phosphatase family protein(FRA3)	Y	N
AT1G67300	Major facilitator superfamily protein(AT1G67300)	Y	Y
AT1G68100	ZIP metal ion transporter family(IAR1)	Y	N
AT1G68890	2-oxoglutarate decarboxylase/hydro-lyase/magnesium ion-binding protein(PHYLLO)	Y	N
AT1G69340	appr-1-p processing enzyme family protein(AT1G69340)	Y	N
AT1G73920	alpha/beta-Hydrolases superfamily protein(AT1G73920)	Y	Y
AT1G73990	signal peptide peptidase(SPPA)	Y	N
AT1G74960	fatty acid biosynthesis 1(FAB1)	Y	Y
AT1G76990	ACT domain repeat 3(ACR3)	Y	Y
AT1G80910	vacuolar fusion CCZ1-like protein (DUF1712)(AT1G80910)	Y	N
AT2G01450	MAP kinase 17(MPK17)	Y	Y
AT2G03730	ACT domain repeat 5(ACR5)	Y	Y
AT2G13370	chromatin remodeling 5(CHR5)	Y	N
AT2G19880	Nucleotide-diphospho-sugar transferases superfamily protein(AT2G19880)	Y	N
AT2G23000	serine carboxypeptidase-like 10(scpl10)	Y	N
AT2G28260	cyclic nucleotide-gated channel 15(CNGC15)	Y	N
AT2G34410	O-acetyltransferase family protein(RWA3)	Y	Y
AT2G36340	DNA-binding storekeeper protein-related transcriptional regulator(AT2G36340)	Y	N
AT2G41210	phosphatidylinositol- 4-phosphate 5-kinase 5(PIP5K5)	Y	N
AT2G44950	histone mono-ubiquitination 1(HUB1)	Y	N
AT2G45920	U-box domain-containing protein(AT2G45920)	Y	N
AT3G01090	SNF1 kinase homolog 10(KIN10)	Y	Y
AT3G09090	defective in exine formation protein (DEX1)(DEX1)	Y	Y
AT3G09410	Pectinacetylesterase family protein(AT3G09410)	Y	Y
AT3G15160	AP-5 complex subunit zeta-1(AT3G15160)	Y	N
AT3G16785	phospholipase D P1(PLDP1)	Y	N
AT3G18140	Transducin/WD40 repeat-like superfamily protein(LST8-1)	Y	Y
AT3G18370	C2 domain-containing protein(ATSYTF)	Y	N
AT3G21710	transmembrane protein(AT3G21710)	Y	Y
AT3G23080	Polyketide cyclase/dehydrase and lipid transport superfamily protein(AT3G23080)	Y	Y
AT3G23640	heteroglycan glucosidase 1(HGL1)	Y	Y
AT3G26470	Powdery mildew resistance protein, RPW8 domain-containing protein(AT3G26470)	Y	N
AT3G47730	ATP-binding cassette A2(ABCA2)	Y	N
AT3G49210	O-acyltransferase (WSD1-like) family protein(AT3G49210)	Y	N
AT3G50240	ATP binding microtubule motor family protein(KICP-02)	Y	N
AT3G53570	serine/threonine-protein kinase AFC1(FC1)	Y	Y
AT3G57300	DNA helicase INO80-like protein(INO80)	Y	Y
AT3G57800	basic helix-loop-helix (bHLH) DNA-binding superfamily protein(AT3G57800)	Y	Y
AT3G62190	Chaperone DnaJ-domain superfamily protein(AT3G62190)	Y	Y
AT3G62750	beta glucosidase 8(BGLU8)	Y	N
AT4G00030	Plastid-lipid associated protein PAP / fibrillin family protein(AT4G00030)	Y	N
AT4G01690	Flavin containing amine oxidoreductase family(PPOX)	Y	Y
AT4G08035	ncRNA(AT4G08035)	N	Y
AT4G10060	Beta-glucosidase, GBA2 type family protein(AT4G10060)	Y	N
AT4G11830	phospholipase D gamma 2(PLDGAMMA2)	Y	Y
AT4G13640	Homeodomain-like superfamily protein(UNE16)	Y	Y
AT4G18120	miscRNA(ML3)	Y	Y
AT4G19040	ENHANCED DISEASE RESISTANCE 2(EDR2)	Y	Y
AT4G24550	Clathrin adaptor complexes medium subunit family protein(AT4G24550)	Y	Y
AT4G26860	Putative pyridoxal phosphate-dependent enzyme, YBL036C type(AT4G26860)	Y	Y
AT4G28470	26S proteasome regulatory subunit S2 1B(RPN1B)	Y	N
AT4G32840	phosphofructokinase 6(PFK6)	Y	N
AT4G33150	lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional enzyme(AT4G33150)	Y	Y
AT4G34240	aldehyde dehydrogenase 3I1(ALDH3I1)	Y	Y
AT4G35230	BR-signaling kinase 1(BSK1)	Y	N
AT4G35785	RNA-binding (RRM/RBD/RNP motifs) family protein(AT4G35785)	Y	Y
AT4G36190	Serine carboxypeptidase S28 family protein(AT4G36190)	Y	N
AT4G38350	Patched family protein(AT4G38350)	Y	Y
AT4G39990	RAB GTPase homolog A4B(RABA4B)	Y	N
AT5G06700	trichome birefringence-like protein (DUF828)(AT5G06700)	Y	N
AT5G09410	ethylene induced calmodulin binding protein(EICBP.B)	Y	Y
AT5G09890	Protein kinase family protein(AT5G09890)	Y	Y
AT5G11640	Thioredoxin superfamily protein(AT5G11640)	Y	N
AT5G12170	CRT (chloroquine-resistance transporter)-like transporter 3(CLT3)	Y	Y
AT5G15270	RNA-binding KH domain-containing protein(AT5G15270)	Y	Y
AT5G16290	VALINE-TOLERANT 1(VAT1)	Y	Y
AT5G22770	alpha-adaptin(alpha-ADR)	Y	Y
AT5G23080	SWAP (Suppressor-of-White-APricot)/surp domain-containing protein(TGH)	Y	Y
AT5G26030	ferrochelatase 1(FC1)	Y	Y
AT5G26240	chloride channel D(CLC-D)	Y	N
AT5G27600	long-chain acyl-CoA synthetase 7(LACS7)	Y	N
AT5G44090	Calcium-binding EF-hand family protein(AT5G44090)	Y	N
AT5G46470	disease resistance protein (TIR-NBS-LRR class) family(RPS6)	Y	N
AT5G47010	RNA helicase(LBA1)	Y	N
AT5G49640	hypothetical protein(AT5G49640)	N	N
AT5G51230	VEFS-Box of polycomb protein(EMF2)	Y	Y
AT5G53090	NAD(P)-binding Rossmann-fold superfamily protein(AT5G53090)	Y	N
AT5G60620	glycerol-3-phosphate acyltransferase 9(GPAT9)	Y	N
AT5G65470	O-fucosyltransferase family protein(AT5G65470)	Y	N
	Total of "Y":	97	51
	Total of "N":	2	48

	COMMON TARGET GENES BETWEEN HAL AND HAS		
GENE ID	**GENE DESCRIPTION**	**INTRONS (Y/N)**	**ALT. SPLICED (Y/N)**
AT1G02890	AAA-type ATPase family protein(AT1G02890)	Y	Y
AT1G07705	NOT2 / NOT3 / NOT5 family(AT1G07705)	Y	Y
AT1G16650	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(AT1G16650)	Y	N
AT1G18450	actin-related protein 4(ARP4)	Y	N
AT1G47550	exocyst complex component sec3A(SEC3A)	Y	Y
AT1G51110	Plastid-lipid associated protein PAP / fibrillin family protein(AT1G51110)	Y	N
AT1G53780	26S proteasome regulatory complex ATPase(AT1G53780)	Y	Y
AT1G60070	Adaptor protein complex AP-1, gamma subunit(AT1G60070)	Y	Y
AT1G71480	Nuclear transport factor 2 (NTF2) family protein(AT1G71480)	Y	N
AT1G75210	HAD-superfamily hydrolase, subfamily IG, 5'-nucleotidase(AT1G75210)	Y	N
AT2G16940	Splicing factor, CC1-like protein(AT2G16940)	Y	Y
AT2G31960	glucan synthase-like 3(GSL03)	Y	Y
AT2G32700	LEUNIG-like protein(LUH)	Y	Y
AT2G41680	NADPH-dependent thioredoxin reductase C(NTRC)	Y	N
AT2G43070	SIGNAL PEPTIDE PEPTIDASE-LIKE 3(SPPL3)	Y	N
AT3G13065	STRUBBELIG-receptor family 4(SRF4)	Y	N
AT3G13445	TATA binding protein 1(TBP1)	Y	Y
AT3G47730	ATP-binding cassette A2(ABCA2)	Y	N
AT4G02570	cullin 1(CUL1)	Y	Y
AT4G03280	photosynthetic electron transfer C(PETC)	Y	Y
AT4G16765	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein(AT4G16765)	Y	Y
AT4G25500	arginine/serine-rich splicing factor 35(RS40)	Y	N
AT4G35785	RNA-binding (RRM/RBD/RNP motifs) family protein(AT4G35785)	Y	Y
AT4G36648	ncRNA(AT4G36648)	Y	N
AT4G38350	Patched family protein(AT4G38350)	Y	Y
AT5G03910	ABC2 homolog 12(ABCB29)	Y	N
AT5G13240	transcription regulator(AT5G13240)	Y	N
AT5G14120	Major facilitator superfamily protein(AT5G14120)	Y	N
AT5G16715	protein EMBRYO DEFECTIVE 2247(EMB2247)	Y	N
AT5G26030	ferrochelatase 1(FC1)	Y	Y
AT5G43710	Glycosyl hydrolase family 47 protein(AT5G43710)	Y	N
AT5G46780	VQ motif-containing protein(AT5G46780)	N	Y
AT5G51290	Diacylglycerol kinase family protein(ACD5)	Y	N
AT5G57940	cyclic nucleotide gated channel 5(CNGC5)	Y	Y
	Total of "Y":	33	17
	Total of "N":	1	17

	UNIQUE TARGET GENES OF AL		
GENE ID	**GENE DESCRIPTION**	**INTRONS (Y/N)**	**ALT. SPLICED (Y/N)**
AT1G12820	auxin signaling F-box 3(AFB3)	Y	N
AT1G25570	Di-glucose binding protein with Leucine-rich repeat domain-containing protein(AT1G25570)	Y	N
AT1G27690	lipase, putative (DUF620)(AT1G27690)	Y	N
AT1G54200	DNA mismatch repair Msh6-like protein(AT1G54200)	N	N
AT3G16785	phospholipase D P1(PLDP1)	Y	N
AT3G50240	ATP binding microtubule motor family protein(KICP-02)	Y	N
AT3G62700	multidrug resistance-associated protein 10(ABCC14)	Y	N
AT4G29780	nuclease(AT4G29780)	N	N
AT4G30340	diacylglycerol kinase 7(DGK7)	Y	Y
AT4G33150	lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional enzyme(AT4G33150)	Y	Y
AT4G37280	MRG family protein(AT4G37280)	Y	N
AT5G11700	ephrin type-B receptor(AT5G11700)	Y	Y
AT5G22450	spectrin beta chain, brain(AT5G22450)	Y	Y
	Total of "Y":	15	4
	Total of "N":	2	13

	UNIQUE TARGET GENES OF AS		
GENE ID	**GENE DESCRIPTION**	**INTRONS (Y/N)**	**ALT. SPLICED (Y/N)**
AT1G08520	ALBINA 1(ALB1)	Y	N
AT1G22930	T-complex protein 11(AT1G22930)	Y	Y
AT1G23900	gamma-adaptin 1(GAMMA-ADAPTIN 1)	Y	Y
AT1G25540	phytochrome and flowering time regulatory protein (PFT1)(PFT1)	Y	Y
AT1G29940	nuclear RNA polymerase A2(NRPA2)	Y	N
AT1G77680	Ribonuclease II/R family protein(AT1G77680)	Y	N
AT2G01340	plastid movement impaired protein(At17.1)	Y	N
AT2G36340	DNA-binding storekeeper protein-related transcriptional regulator(AT2G36340)	Y	N
AT2G47600	magnesium/proton exchanger(MHX)	Y	Y
AT3G33530	Transducin family protein / WD-40 repeat family protein(AT3G33530)	Y	Y
AT4G11830	phospholipase D gamma 2(PLDGAMMA2)	Y	Y
AT4G12030	bile acid transporter 5(BAT5)	Y	Y
AT4G19710	aspartate kinase-homoserine dehydrogenase ii(AK-HSDH II)	Y	Y
AT5G13950	nuclear factor kappa-B-binding protein(AT5G13950)	Y	Y
AT5G16030	mental retardation GTPase activating protein(AT5G16030)	Y	Y
	Total of "Y":	15	10
	Total of "N":	0	5

	COMMON TARGET GENES BETWEEN AL AND AS		
GENE ID	**GENE DESCRIPTION**	**INTRONS (Y/N)**	**ALT. SPLICED (Y/N)**
AT1G02305	Cysteine proteinases superfamily protein(AT1G02305)	Y	N

APPENDIX

Appendix 1. Script for HyperTRIBE analysis protocol

#Analysis of RNA-seq data

#Go to JupyterLab, a web-based user interface that allows users to easily access a Linux terminal

enter this web address: https://jupyter2.rc.colorado.edu/hub

#log into supercomputer SUMMIT

ssh -l username@colostate.edu login.rc.colorado.edu

#enter CSUpassword,push #Verify login on DUO key two-factor verification

#loading container (which contains downloaded tools including STAR, trimmomatic, etc)

ssh scompile

#go to working directory where you will be working your RNA-sequencing analyses

cd /projects/username@colostate.edu

#Download the software contents of Nature Protocol GitHub which includes prewritten shell and Perl scripts

#It should download a file called, "HyperTRIBE", but for my own interests, I changed "HyperTRIBE" to "ARAB_HT". I did this manually by right-clicking the name on JupyterLab

git clone https://github.com/rosbashlab/HyperTRIBE

#Download annotations for the transcriptome and create STAR/indices for the reference genome

#HyperTRIBE requires transcriptome annotation in two formats (RefSeq annotation in refFlat format from the UCSC Genome Browser and gene transfer format (GTF) and genome sequence in FASTA format.

#Download the genome file and the annotation (GTF) file into the directory

#The genome sequence and TAIR10 genome annotations were found in Arabidopsis.org(more specifically:https://www.arabidopsis.org/download/index-auto.jsp%3Fdir%3D%252Fdownload_files%252FGenes%252FTAIR10_genome_release)

wget http://ftp.gramene.org/CURRENT_RELEASE/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz

wget https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff

#Unzipping to uncompress the files

gunzip Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz

gunzip TAIR10_GFF3_genes.gff

#There is currently no downloadable annotation file that is in refFLAT format, so we will organize the columns of the gff file to make it into our desirable format.
#The GTF/GFF annotation format is used for building the STAR indices and the refFlat annotation format is used for majority of the HyperTRIBE shell scripts

#Download gff3ToGenePred from UCSC website:
https://genome.ucsc.edu/goldenPath/help/hubQuickStartSearch.html

#Run gff3ToGenePred to convert your file to GenePred file format (= refFlat)

gff3ToGenePred TAIR10_GFF3_genes.gff -T -o TAIR10_GFF3_genes.genepred

#After you've converted your file to the genePred format, you'll then have to use your own scripting method to add in the extra "geneName" column to turn your genePred file into a refFlat file. You'll have to decide what you want to put in this column. In this case, we will repeat the first column since it also represents the gene name.

awk '{print$ 1 1 2 3 4 5 6 7 8 9 10}' TAIR10_GFF3_genes.genepred >
TAIR10_GFF3_genes.refFlat

#Download your samples (fastq.gz file) onto a new directory file

mkdir TRIBEsamples

#On supercomputer Summit, there is an "Upload Files" button that you can easily download your files onto the working directory

#Once uploaded, unzip the compressed fastq.gz file (ex. LCS7704_ASN_AL_1_R1.fq.gz)

gunzip LCS7704_ASN_AL_1_R1.fq.gz

#Continue unzipping all of the compressed downloaded files

#Go back to the home directory and make a new directory called "align_scripts" which is where you will create and run a shell script called, buildarab_index.sh, to create the STAR indices.

#You will be outputting these generated indices into a new folder: star_index or your desired output name.

mkdir align_scripts

nano buildarab_index.sh

#Edit buildarab_index.sh with these codes:

```
#!/usr/bin/bash

#SBATCH --job-name=execute_star-build
#SBATCH --nodes=1
#SBATCH --ntasks=8 # modify this number to reflect how many cores you want to use (up to 24)
#SBATCH --partition=shas-testing
#SBATCH --qos=testing     # modify this to reflect which queue you want to use. Options are 'normal' and 'testing'
#SBATCH --time=0:7:25   # modify this to reflect how long to let the job go. This indicates 4 hours.
#SBATCH --output=log_star-build_%J.txt

# Load singularity

source /projects/dcking@colostate.edu/paths.bashrc

# Build star indexes for arabidopsis

STAR --runThreadN $SLURM_NTASKS \
--runMode genomeGenerate \
--genomeDir /projects/username@colostate.edu/ARAB_HT/star_index \
--genomeFastaFiles Arabidopsis_thaliana.TAIR10.dna.toplevel.fa \
```

--sjdbGTFfile TAIR10_genes.gff \

--sjdbGTFtagExonParentTranscript Parent \

--sjdbOverhang 140 \

#This will generate the reference genome STAR indices

#Now we have to trim and align our samples onto the reference genome STAR indices

#Move to the "CODE" directory which has all of the pre-written scripts that were downloaded from GitHub and open the "trim_and_align.sh" shell script.

nano trim_and_align.sh

#Update this script to reflect the location of the softwares (Trimmomatics and Picard) and the STAR indices.

star_indices="/projects/username@colostate.edu/ARAB_HT/star_index"

TRIMMOMATIC_JAR="/projects/dcking@colostate.edu/src/Trimmomatic-0.36/trimmomatic-0.36.jar"

PICARD_JAR="/projects/dcking@colostate.edu/jar/picard.jar"

#This shell script will also be executed when you are ready to trim low quality bases/reads and align your reads.

#The following code utilizes trimmomatics to remove the first 6 nucleotides of the read for potential error from hexamer mispriming and filtering out reads with an average quality score less than 25.

#You can change these variables into your desirable parameters.

java -jar $TRIMMOMATIC_JAR SE -phred33 $trim_input $trim_outfile HEADCROP:6 LEADING:25 TRAILING:25 AVGQUAL:$avgquality MINLEN:19

#Align library with STAR

input=$trim_outfile

STAR --runThreadN 8 --outFilterMismatchNoverLmax 0.07 --outFileNamePrefix $prefix"_" --outFilterMatchNmin 16 --outFilterMultimapNmax 1 --genomeDir $star_indices --readFilesIn $input

--outFilterMismatchNoverLmax 0.07: number of mismatches is <= 7% of mapped read length (maximum of 5 mismatches in 75 nucleotide reads to adjust for any multiple editing marks by HyperTRIBE.

--outFilterMatchNmin 16: min number of bases mapped genome per read

--outFilterMultimapNmax 1: output reads that only map to one loci

#It also includes codes that utilize samtools to remove low quality alignment (quality score >10), convert sam to bam, sort the bam file before using Picard to remove duplicates, and create a SAM file from the sorted bam file.

```
samtools view -@ 4 -Sh -q 10 $output > $prefix"_highquality.sam"
mv $prefix"_highquality.sam" $output
bam_out=$prefix".bam"

#convert sam to bam

samtools view -@ 4 -bhS $output > $bam_out
rm $output

#sort the bam file before using picard

sort_out=$prefix".sort.bam"
samtools sort -@ 6 $bam_out -o $sort_out
rm $bam_out

#run Picard to remove duplicates

input_for_picard=$sort_out
dupremove_bam=$prefix"_nodup.bam"
```

```
java -Xmx4g -jar $PICARD_JAR MarkDuplicates INPUT=$input_for_picard
OUTPUT=$dupremove_bam METRICS_FILE=dup.txt
VALIDATION_STRINGENCY=LENIENT REMOVE_DUPLICATES=true TMP_DIR=tmp
ASSUME_SORTED=true
rm $input_for_picard

#sort the output bam file from picard

sort_out=$prefix".sort.bam"
samtools sort -@ 6 $dupremove_bam -o $sort_out
rm $dupremove_bam

#The next step of HyperTRIBE requires the sam file to be sorted
#Create a SAM file from this sorted bam file

samtools view -@ 4 -h $sort_out > $prefix".sort.sam"
samtools index $sort_out

echo "Done with STAR mapping and PCR duplicate removal with PICARD"
echo "created sam file: $prefix.sam"
-----------------------------------------
```

#Go back to the align_scripts directory and make a new shell script called trim_wtRNA.sh. When running this script, it will execute the trim_and_align.sh with the given sample file. (ex. LCS7704_ASN_M_3_R1.fq)

```
nano trim_wtRNA.sh
```

#Edit trim_wtRNA.sh with these codes:

```
---------------------------------------------
#!/usr/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks=9
#SBATCH --time=1:00:00
#SBATCH --qos=normal
#SBATCH --partition=shas
#SBATCH --output=M_3_R1_RNA.%j.out

source /projects/dcking@colostate.edu/paths.bashrc

cmd='/projects/username@colostate.edu/ARAB_HT/CODE.copy/trim_and_align.sh
/projects/username@colostate.edu/ARAB_HT/TRIBEsamples/LCS7704_ASN_M_3_R1.fq'

echo $cmd
```

time eval $cmd

#To run this shell script, write this line on the command line:

sbatch trim_wtRNA.sh

#Executing this job will STAR map and perform PCR duplicate removal with PICARD the samples. It will output a summary of the quality control and these output files with these endings : .sort.bam, .sort.bam.bai, .sort.sam, .trim.fastq files.

#Exit from the Summit computer and enter to your Linux terminal

#Access the MySQL database by logging in using your root password

mysql -h localhost -u root -p

#Now you have to create a MySQL database that you will download your sample SAM files into

CREATE DATABASE arabidopsis;

#Create a new directory into your local computer to continue working on the samples

mkdir username_arabidopsis

cd username_arabidopsis

#Redownload the HyperTRIBE package

git clone https://github.com/rosbashlab/HyperTRIBE

#Download the genome (refFlat) annotation files

#Go into the annotations directory where you can download the genome annotation files into:

cd HyperTRIBE

cd annotations

rsync -auvz -e 'ssh -p 22'

username\@colostate.edu@login.rc.colorado.edu:/projects/username@colostate.edu/ARAB_HT/

Arab_build/TAIR10_GFF3_genes.refFlat .

#Make a sample directory to input all of the samples

mkdir samples

cd samples

#Use this command to sync each samples onto the local computer

rsync -auvz -e 'ssh -p 22'

username\@colostate.edu@login.rc.colorado.edu:/projects/username@colostate.edu/ARAB_HT/samples/SAMPLE_NAME .

#Do this command until you have received all of the SAM/BAM/BAM.BAI files from Summit

#Go to the directory where load_table.sh script is located and go ahead and edit the file to change HyperTRIBE_DIR to the correct absolute path of where all the HyperTRIBE scripts are found

nano load_table.sh

HyperTRIBE_DIR= "Users/reddylab/username_arabidopsis/HyperTRIBE/CODE"

#This file will execute sam_to_matrix.pl file, which creates the matrix file for each sample, and load_matrix_data.pl, which will load the matrix file to the mysql database

#Go to the directory where load_matrix_data.pl script is located and edit the file with the correct database, user, and password to connect to the mysql database

nano load_matrix_data.pl

my $database = "arabidopsis"

my $user = "root"

my $password = "PASSWORD"

#This file will connect to the mysql database and create a table with the sample information/statistics (chromosome name, A count, T count, etc.)

#Now it's time to download each of your sample SAM file into the arabidopsis database by using this following command:

/Users/reddylab/username_arabidopsis/HyperTRIBE/CODE/load_table.sh /Users/reddylab/username_arabidopsis/samples/SAMPLE_NAME table_name exp_name replicate_number .

#Do this command until you have downloaded all samples into the designated database

/Users/reddylab/username_arabidopsis/HyperTRIBE/CODE/load_table.sh /Users/reddylab/username_arabidopsis/samples/SAMPLE_NAME.sort.sam MYSQL_TABLE EXP_NAME REPLICATE/TIMEPOINT

#This command will execute load_table.sh script with the sample given
#Go to the directory where find_rnaeditsites.pl and edit edit the file with the correct database, user, and password to connect to the mysql database

nano load_matrix_data.pl
my $database = "arabidopsis"
my $user = "root"
my $password = "PASSWORD"

#Go to the directory where rnaedit_wtRNA_RNA.sh script is located and edit the file to change HyperTRIBE_DIR to the correct absolute path of where all the HyperTRIBE scripts are found

nano rnaedit_wtRNA_RNA.sh

#Change the ANNOTATION to the correct absolute path of where the genome annotation file will be located

HyperTRIBE_DIR= "Users/reddylab/username_arabidopsis/HyperTRIBE/CODE"

annotationfile="Users/reddylab/username_arabidopsis/

#Change the the wtRNA variables, timepoint, and the desired edit_threshold/read_threshold

wtRNAtablename= "RNA"

wtRNAexp= "rnalibs"

wtRNAtp= "1"

RNAtablename= "RNA"

RNAexp= "rnalibs"

timepoint= (2 3 4 5 6)

#This file will take that combination of table name, experiment name, and replicate/timepoint variable and use it to extract the base composition between the rna library and wtRNA library to call the edit sites

#Inputting multiple timepoints allow the script to run through multiple RNA libraries in a loop

#Once this script is executed, it calls out the edit site, apply the edit threshold and read threshold, and organize the output into a bedgraph format (a list of editing sites found)

#The script is able to generate these bedgraph track by executing the find_rnaeditsites.pl script which does a pairwise comparison of RNA against wtRNA for each nucleotide in the transcriptome to detect a set of editing sites

#Then it runs Threshold_editsites_20reads.py to ensure that the editing sites are required to have at least 11% editing and a coverage of at least the number of reads desired

#Go to the directory where find_rnaeditsites.pl script is located and edit the file to input the mysql variables

nano find_rnaeditsites.pl

my $database = "arabidopsis"

my $user = "root"

my $password = "PASSWORD"

#All of the scripts are edited with the correct input/variables at this point

#To execute all of the scripts; use this command:

./rnaedit_wtRNA_RNA.sh

#Once the scripts are runned, there should be a bedgraph for each pairwise comparison of the wtRNA vs. RNA library

#It also executes the summarize_results.pl which will organize the edit sites into a more concise order

#The final files generated from these scripts: (**# means number)

#a. “rnalibs_wtRNA#_RNAlib#.txt”: This document lists all editing sites found in that RNA library

#b. “rnalibs_wtRNA#_RNAlib#_#%.bedgraph”: This document lists editing sites that meets the given editing and read threshold

#c. “rnalibs_wtRNA#_RNAlib_#%_results.xls”: This document is a summary list of the edit sites within the editing and read threshold (a cleaner version of the bedgraph file)

#To overlap the triplicates to identify matching edit coordinates, go onto Python

#Write this into a new script:

```
import re

def find_editing_sites(editing_site_files, alignment_files):
    editing_sites_file_names = extract_files_from_list(editing_site_files)
```

```
    alignment_file_names = extract_files_from_list(alignment_files)

    editing_coordinates = extract_editing_coordinates(editing_sites_file_names)
    find_same_editing_sites(editing_coordinates, alignment_file_names)


def extract_files_from_list(file_name):
    if isinstance(file_name, list):
        return [x for x in file_name]
    return [file_name]


def extract_editing_coordinates(alignment_files):
    all_editing_coordinates = {}

    for alignment_file in alignment_files:
        editing_coordinates = []
        with open(alignment_file) as file:
            file_header = file.readline()
            for line in file:
                editing_coordinates.append([x for x in line.split()][1])

        all_editing_coordinates[alignment_file] = editing_coordinates

    return all_editing_coordinates
```

```
def find_same_editing_sites(editing_coordinates,  alignment_file_names):
    for editing_site_file in editing_coordinates.keys():
        for alignment_file in alignment_file_names:
            matches = {}
            coordinates = []
            if alignment_file == editing_site_file:
                continue
            else:
                print('Now searching file {} for the editing coordinates found in file
{}'.format(alignment_file,
                                                                    editing_site_file))
                with open(alignment_file) as file:
                    file_header = file.readline()
                    for line in file:
                        coordinates.append([x for x in line.split()][1])
                    matches = set(x for x in coordinates if x in editing_coordinates[editing_site_file])

                if len(matches):
                    print('--> The file {} has {} matching editing coordinates found in {} the matching
coordinates are:'.format(alignment_file, len(matches), editing_site_file))
                    for match in matches:
                        print('    {}'.format(match))
```

```
        else:
            print('{} has no matching editing coordinates found in {}'.format(alignment_file,
editing_site_file))


find_editing_sites("rnalibs_wtRNA#_RNAlib#_#%.bedgraph",
"rnalibs_wtRNA#_RNAlib#_#%.bedgraph", "rnalibs_wtRNA#_RNAlib#_#%.bedgraph")


---------------------------------------------------------------
#On the final line of the script, substitute "rnalibs_wtRNA#_RNAlib#_#%.bedgraph" with your
desired three bedgraph files to overlap
#Executing this script will find the final matching edit coordinates that are found in all triplicate
samples which is the total list of the high-confidence targets of your protein of interest


#Another option is you could use https://www.biovenn.nl/ and fill in the output data from each
triplicate to each corresponding box to find the overall overlapped results
```

Appendix 2. Script for the visualization of SR45 differential gene expression

```
#Have to convert gff3 genome annotation format into gtf file to use featureCounts


gff3toGenePred
GenePredtoGtf
```

#log in to supercomputer SUMMIT

ssh -l username@colostate.edu login.rc.colorado.edu

#enter CSUpassword,push #answer DUO key on phone app

#enter into scompile, a specific compute node on SUMMIT

ssh scompile

#Use this line into the command line to use featureCounts on all 24 files

featureCounts -p -T 8 -s 2 -a
/scratch/summit/username@colostate.edu/HyperTRIBE/CODE/TAIR10_GFF3.gtf -o
HyperTRIBE_TRIBE.txt
/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_A_1_R
1.sam
/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_A_2_R
1.sam
/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_A_3_R
1.sam
/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AL_1_
R1.sam
/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AL_2_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AL_3_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AS_1_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AS_2_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_AS_3_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HA_1_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HA_2_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HA_3_

R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAL_1

_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAL_2

_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAL_3

_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAS_1

_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAS_2_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_HAS_3_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_M_1_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_M_2_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_M_3_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_WT_1_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_WT_2_R1.sam

/scratch/summit/username@colostate.edu/HyperTRIBE/TRIBEsamples/LCS7704_ASN_WT_3_R1.sam

#outputs of featureCounts

#a count matrix and a summary file that tabulates how many the reads were “assigned” or counted and the reason they remained “unassigned”

import the counts data into Rstudio

```
getwd()


#Set this to your working directory:
# You may need to set this to your own working directory to your scripts directory:
setwd("/set/to/your/scripts/directory")


getwd()
countsData <- read.table(file = "HyperTRIBE_TRIBE.txt", header = FALSE, row.names = 1,
skip = 2)


#Make a metadata:
> id <- c("LCS7704_ASN_A_1_R1.sam", "LCS7704_ASN_A_2_R1.sam",
"LCS7704_ASN_A_3_R1.sam", "LCS7704_ASN_AL_1_R1.sam",
"LCS7704_ASN_AL_2_R1.sam", "LCS7704_ASN_AL_3_R1.sam",
"LCS7704_ASN_AS_1_R1.sam", "LCS7704_ASN_AS_2_R1.sam",
"LCS7704_ASN_AS_3_R1.sam", "LCS7704_ASN_HA_1_R1.sam",
"LCS7704_ASN_HA_2_R1.sam", "LCS7704_ASN_HA_3_R1.sam",
"LCS7704_ASN_HAL_1_R1.sam", "LCS7704_ASN_HAL_2_R1.sam",
"LCS7704_ASN_HAL_3_R1.sam", "LCS7704_ASN_HAS_1_R1.sam",
"LCS7704_ASN_HAS_2_R1.sam",
"LCS7704_ASN_HAS_3_R1.sam","LCS7704_ASN_M_1_R1.sam",
"LCS7704_ASN_M_2_R1.sam", "LCS7704_ASN_M_3_R1.sam",
```

"LCS7704_ASN_WT_1_R1.sam", "LCS7704_ASN_WT_2_R1.sam",

"LCS7704_ASN_WT_3_R1.sam")

> sample <- c("A_1", "A_2", "A_3", "AL_1", "AL_2", "AL_3", "AS_1", "AS_2", "AS_3",

"HA_1", "HA_2", "HA_3", "HAL_1", "HAL_2", "HAL_3", "HAS_1", "HAS_2", "HAS_3",

"M_1", "M_2", "M_3", "WT_1", "WT_2", "WT_3")

> type <- c("Arab","Arab", "Arab", "Arab", "Arab", "Arab", "Arab","Arab", "Arab", "Arab",

"Arab", "Arab", "Arab", "Arab", "Arab", "Arab", "Arab", "Arab", "Arab","Arab", "Arab",

"Arab", "Arab", "Arab")

>isoform <- c("A", "A", "A", "AL", "AL", "AL", "AS", "AS", "AS", "HA", "HA", "HA",

"HAL", "HAL", "HAL", "HAS", "HAS", "HAS", "M", "M", "M", "WT", "WT", "WT")

>rep <- c("1", "2", "3", "1", "2", "3","1", "2", "3","1", "2", "3", "1", "2", "3", "1", "2", "3","1",

"2", "3","1", "2", "3")

> View(metadata)

> View(metadata)

> metadata <- data.frame(sample, name, type, isoform, rep)

> metadata

#Let's give column names onto the metadata

colnames(metadata) <- c("id", "sample", "type", "isoform", "rep")

Let's give countsData some columns names. The first names will be... chr', 'start', etc...

The last names will be names for each sample. We can pull those names from metadata:

```
as.vector(metadata$sample)

# Name countsData columns headers:
colnames(countsData) <- c("chr", "start", "stop", "strand", "length", as.vector(metadata$sample))

# In this section we will prepare our input data for analysis.
# In the instruction for DESeq2, it states: "We read in a count matrix, which we will name cts,
and the sample information table, which we will name coldata."

# OK, our task will be to generate a table called "cts" out of the countsData table.
# Subset the countsData

head(countsData)
dim(countsData)

head(countsData[,6:29])

# Save just the subset as an object called cts:

cts <- as.matrix(countsData[,6:29])

# Next we need to make an information called coltable. We can make this out of the metadata
table.
```

```
class(metadata)

# Reorganize the metadata table so the sample column are now row headers

metadata
rownames(metadata)<-metadata$sample
metadata

coldata <- metadata[,c("type", "isoform", "rep")]
coldata$isoform <- as.factor(coldata$isoform)
coldata$rep <- as.factor(coldata$rep)
rownames(coldata)
colnames(cts)

#Reorder it so you can take control which is your reference level for your pairwise-comparison

coldata$isoform <- relevel(coldata$isoform, "M")

# Yay! Now we have coldata! This is a new metadata object where we have just selected the type
of information that is critical for deseq2 to use.
# One thing we need to explicitly check. The rownames of coldata need to exactly match the
colnames of cts.
```

```
#Check that the names match  --> Should be TRUE

all(rownames(coldata) == colnames(cts))

# Next we will create an ddsHTSeq object out of cts and coldata:
# This will set a base design for your experiment:
# Load all the _counts.txt files and to attach them to the metadata.

dds <- DESeqDataSetFromMatrix(countData = cts, colData = coldata, design = ~ isoform)

############## PRE-FILTERING -- FILTER FOR PRESENT GENES: #################
# Not necessary, but helps keep things fast.
# Exclude all samples that have 0 reads:

keep <- rowSums(counts(dds)) >= 1
dds <- dds[keep,]

#################### NOTE ON FACTOR LEVELS ###################
# Organize the categories based on what makes sense:

dds$isoform <- factor(dds$type, levels = c("A", "AL", "AS", "HA", "HAL", "HAS", "M",
"WT"))
```

```
# PERFORM DESEQ2 analysis:
# This will transform dds into a specialized object with many more fields filled in.

dds <- DESeq(dds)

# Here is a demonstration of the size Factor scaling that was calculated (sizeFactor):
dds$sizeFactor

############## DIFFERENTIAL EXPRESSION ANALYSIS ####################

# calculate the statistically significant differences between different samples

res_MvsWT <- results(dds,lfc=1,contrast=c("isoform", "M", "WT"))

# to save this file as a document

write.table(res_MvsWT, file="Counts_MTvsWT.txt", sep="\t", quote=F, col.names=NA)

############## PERFORM LOG FOLD CHANGE SHRINKAGE FOR VISUALIZATION
####################

# An input requirement of the lfcShrink function is a coef term. This is pulled from the
resultsNames of dds:
```

```
resultsNames(dds)


resLFC_ALvsA <- lfcShrink(dds, coef="isoform_AL_vs_A", res = res_ALvsA, type='apeglm')
resLFC_ASvsA <- lfcShrink(dds, coef="isoform_AS_vs_A", res = res_ASvsA, type='apeglm')
resLFC_HAvsA <- lfcShrink(dds, coef="isoform_HA_vs_A", res = res_HAvsA, type='apeglm')
resLFC_HALvsA <- lfcShrink(dds, coef="isoform_HAL_vs_A", res = res_HALvsA,
type='apeglm')
resLFC_HASvsA <- lfcShrink(dds, coef="isoform_HAS_vs_A", res = res_HASvsA,
type='apeglm')
resLFC_MvsA <- lfcShrink(dds, coef="isoform_M_vs_A", res = res_MvsA, type='apeglm')
resLFC_WTvsA <- lfcShrink(dds, coef="isoform_WT_vs_A", res = res_WTvsA, type='apeglm')


summary(resLFC_ALvsA)


##################  Exploring and exporting results ##################


##### KNOWN GENES:


# Check known genes to make sure everything is working as predicted.


plotCounts(dds1, gene=which(rownames(resLFC_ALvsA) == "AT1G16610")
```