DISSERTATION

FUNCTIONAL ANALYSIS OF THREE ARABIDOPSIS SR PROTEINS (SCL33, SC35, SCL30A) IN PLANT DEVELOPMENT AND SPLICING

Submitted by

Julie Thomas

Department of Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2012

Doctoral Committee:

Advisor: A.S.N. Reddy

Pat Bedinger
Marinus Pilon
Jeff Wilusz

ABSTRACT

FUNCTIONAL ANALYSIS OF THREE ARABIDOPSIS SR PROTEINS (SCL33, SC35, SCL30A) IN PLANT DEVELOPMENT AND SPLICING

Precursor-mRNA (Pre-mRNA) splicing is dependent on many RNA binding proteins that recognize sequence signals in RNA and regulate splicing. The serine/arginine (SR)-rich proteins are a family of RNA binding spliceosomal proteins that perform essential functions during spliceosome assembly by interacting with splicing regulatory sequences in pre-mRNA as well as with other spliceosomal proteins. These RNA-protein and protein-protein interactions of SRs play a crucial role in constitutive splicing as well as in alternative splicing (AS). Since SR genes regulate their own splicing and subsequently affect AS of other *SR* mRNAs as well as that of many other coding genes, elucidation of functions of the SRs is critical for understanding gene regulation at the pre-mRNA splicing level.

The roles of plant SR proteins in growth and development are poorly understood and no loss-of-function mutants have been characterized. In *Arabidopsis* there are 18 SRs that are grouped into six subfamilies. Since the SR family is expanded in plants with many paralogous genes with potential redundant functions, I used a gene knockout strategy to generate mutants lacking two or more *SR* genes. To address the role of SRs in plant development we generated loss-of-function mutants of *SC35*, the sole member of the SC35 family with a counterpart in humans, and of *SCl33* and *SCl30a* that belong to the plant-specific SC35-like (SCL) family. To address potential functional redundancy and/or synthetic phenotypes among these genes we generated three double mutants (*sc35 scl30a, sc35 scl33, scl33 scl30a*) representing all combinations, and a triple mutant. All mutants are viable but displayed complex and opposing flowering phenotypes in different mutants. Among the single mutants *sc35* and *scl30a* showed

early flowering whereas *scl33* showed delayed flowering under both long days (LD) and short days (SD). In double mutant combinations, *sc35 scl30a* flowered early as in single mutants and no additive effect was observed. In contrast *scl33* was epistatic to *scl30a* in the double mutant and to *sc35* and *scl30a* in the triple mutant and these exhibited an even more pronounced late flowering phenotype as compared to *scl33*. The late flowering phenotype of *scl33*, *scl33 scl30a* and *scl33 sc35 scl30a* under both LD and SD and rescue of this phenotype by vernalization, suggest that they regulate the autonomous flowering pathway. In the late flowering mutants expression of *Flowering Locus C* (*FLC*), a key negative regulator of flowering, and *FRIGIDA* (*FRI*), a positive regulator of *FLC* expression are upregulated. In contrast, early flowering mutants (*sc35*, *scl30a* and *sc35 scl30a*) showed increased expression of *FLOWERING LOCUS T* (*FT*), a positive regulator of flowering. These results indicate that members of the SR gene family perform opposing roles in regulating flowering time.

In Arabidopsis, pre-mRNAs of serine/arginine-rich (SR) proteins undergo extensive alternative splicing. However, little is known about the *cis*-elements and *trans*-acting proteins involved in regulating AS. To study the role of SR proteins in AS, a splicing reporter (GFP–intron–GFP) was constructed, consisting of the GFP coding sequence interrupted by an alternatively spliced intron of SCL33. We investigated whether *cis*-elements within this intron are sufficient for AS, and which SR proteins are necessary for regulated AS. Expression of the splicing reporter in protoplasts faithfully produced all splice variants from the intron, suggesting that *cis*-elements required for AS reside within the intron. To determine which SR proteins are responsible for AS, the splicing pattern of the GFP–intron–GFP reporter was investigated in protoplasts of three single and three double mutants of SR genes. These analyses revealed that SCL33 and its closely related paralog, SCL30a, are functionally redundant in generating specific splice variants from this intron. Furthermore, SCL33 protein bound to a conserved sequence in this intron, indicating auto-regulation of AS. Mutations in four GAAG repeats within the conserved region impaired generation of the same splice variants that are affected in the scl33 scl30a double mutant. Thus, have identified the first intronic *cis*-element involved in AS of a plant SR gene, and elucidated a mechanism for auto-regulation of AS of this intron.

Global changes in gene expression and splicing in loss-of-function mutants of plant SR proteins have not been carried out. Here, we performed a transcriptome analysis in a triple mutant lacking three SR genes (*SC35, SCL30a* and *SCL33*) using next generation sequencing to monitor transcriptome-wide changes. About 80 million reads (40 M from WT and 40 M from mutant) from two-week old seedlings of wild type and the triple mutant were obtained and analyzed for changes in gene expression and pre-mRNA splicing. Analyses of this RNA-Seq data show that loss of SC35, SCL30a and SCL33 results in significant changes in expression of protein- and non-protein coding genes (miRNAs and other non-coding RNAs) and splicing patterns of many genes. Expression of 737 genes including 5 miRNAs was altered in the mutant, of which 351 are up-regulated whereas 386 are down-regulated by atleast 3 fold. Our analysis also identified 13 novel transcriptional units. In addition, splicing patterns of several genes is altered in the mutant. There is also considerable overlap between differentially expressed and differentially spliced genes in the mutant. Among differentially spliced genes both qualitative and quantitative changes in splicing were observed in the mutant. We validated over 30 genes that are either differentially expressed or spliced using RT-PCR. Analysis of all differentially expressed genes using gene ontology (GO) terms has revealed that genes involved in iron and phosphorous homeostasis and stress responses especially in plant immunity are overrepresented. The set of differentially expressed/spliced genes that we identified here represents direct and

indirect targets of these SR proteins. Identification of direct targets of each of the SRs using methods such as PAR-CLIP will help us not only to identify which of these are indirect targets but also pave the way for using computational tools to identify potential splicing regulatory elements in direct targets.

ACKNOWLEDGEMENTS

What I have today is because of the help of a number of people. This is my thanks to all of them. They have been my friends, advisors, colleagues and family, many who have indirectly, yet importantly shown me the way and walked with me along the path to my PhD. I come from Pantnager, a small town in Northern India known for having the first agricultural university in India. My journey started when I got accepted as a graduate student in the Program of Plant Molecular Biology. This program provided me with funding for a year so that I could rotate in different labs and select a lab of my own research interest. Many long years later I am able to stand proud in Fort Collins, with gratitude for times past and friends made, all of which made this possible. To paraphrase Issac Newton, " We are able to see further, not so much that we are better, but because we stand on the shoulders of people who have walked before us".

Thank you Pallu- for the days when you asked "Mom, you OK? Those were times when I had a computer on my lap (and not her), a plate of dry Pizza by my side, a glass of cold coffee, bundles of papers and books all around, a clock which said 3 AM (3 AM!) and a brain which had stopped working. A little girl's voice kept me going. Sometimes, unexpected people are fond parts of our lives, Savitri and Chandran Mayandi were two such people during my studies, for my daughter Pallu, they were her first grandparents and I will always have them in my heart. Also, tons of admiration for my husband for believing in me and encouraging me and most of the time being patient for taking a long vacation from home.

Dr. Reddy, Dr. Ueli Grossniklaus, Dr. Marja Timmerman and Dr. Andy Pereira- you four pushed me to the furthest of my potential and helped me find, through adversity, all that I could achieve, but never dared to dream. That is something for which I can never thank you enough.

"How do you thank someone who has taken you from Crayons to perfumes…it isn't easy…but I will try" is a line from the movie "To Sir with Love". Each one of you- in different ways- have been the Sidney Poitier person to me. Also special thanks to my advisory committee who have talked to me, guided me and encouraged me. I thank Dr. Asa Ben-Hur and lab members Mike Hamilton and Mark Rogers for help in the Bioinformatics analysis of RNA-Seq data.

Maria- a friend from Pakistan who is family now. You made me "chill" whenever my brain stopped on me. Love you. Alicia- you made work fun. When work pushed me to the uptown green house, you pulled me to the downtown dream house (movies). Yes- we wined and dined so we could smile whenever Reddy gave us the "Are you working?" look. My friends here in Fort Collins, Ashley, Farrah, Vince and Jared for the immense support you have given me, for being there when I needed someone.

Thanks to all the support staff here in the biology department and also to the department for an opportunity to have a teaching assistantship whenever it was required. The teaching breaks from research were quite good because the interaction with students gave me experience in formulating part of my knowledge into practical words of understanding.

I must thank Reddys's lab members, especially those who have left (Helena, Gulshad) and the people who are still working, for giving me a nice interactive atmosphere to work and share ideas; Dr. Irene Day for showing me everyday that we as women can do great in professional life even though, we have home-front responsibilities; Dr. Saiprasad Palusa, Dr. Salah Abdel-Ghani and Dr. David Xing for supporting me during my research and providing immense contribution for the submission of papers. Of course, not to forget some undergraduate champions: Iman, Katherine, Bridget, Annie and Sarah. They worked with enthusiasm and have had a major role in finishing some of the tedious parts of the project, especially outrivaling the challenges needed for optimal plant growth. Dr. Reddy, I am very grateful for you leaving your office door open at all time and giving me needed assistance in thoughts and experimental help for even the tallest hurdle.

This country is a wonderful experiment. Where else would people from other nations be given opportunities to learn and progress professionally, to go for the moon and reach it still? To the people who make up this nation, and the social fabric and systems they build to make stories such as mine possible- my heart-felt appreciation.

Life is often the journeys we make and the chapters we write. Some rainbows. Some pot holes. But yet, so many wonderful lessons to learn along the way. Even if its with scraped knees and knocked heads. But then what would it be if we stayed put wherever we were? As they say "Ships are safe in the harbor, but that is not what ships are for". This journey has been a most improbable one. But again that's what has made it so much more sweet to savor. To all of you especially my parents who cultivated in me the ability to have immense patience to finish tasks, however long it takes- thank you. You are part of me and I'm all the more blessed for what that means.

Julie Thomas, PhD.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Gene expression is a process by which the information in DNA is decoded in the form of proteins and regulatory RNA molecules. Proper expression of genes in response to intrinsic as well as external signals is crucial for an organism’s development and its adaptation to the external environment. Transcription, the first step in decoding information from genes, makes copies of RNA from a gene. In eukaryotes, the primary transcript or precursor mRNA (pre-mRNA) is then processed to generate a functional mRNA. The functional mRNA is transported into cytoplasm where it is translated into a protein, which in some cases is further processed or modified to form a functional protein. Hence, the process of gene expression can be regulated at the transcriptional, post-transcriptional, translational and post-translational level. As my research is focused on the roles of the serine/arginine-rich (SR) family of proteins in pre-mRNA splicing, I will discuss different steps of gene regulation briefly first and then elaborate on various aspects of pre-mRNA splicing.

Transcriptional Regulation

Transcriptional regulation is primarily controlled by DNA sequences upstream of the transcription start site that bind to the core transcriptional machinery proteins, namely: RNA polymerase, TATA-box binding proteins, transcription factors, activators and repressors. In many instances DNA sequences, called enhancers, that are far way from the gene and located either upstream or downstream of a gene can also regulate transcription (Ong and Corces, 2012). Genomic DNA in eukaryotic cells is tightly packaged into chromatin, which consists of histone and non-histone proteins. At the lowest level of chromatin organization, DNA is wrapped around nucleosomes that are made of histones. This architecture of the genome not only allows packaging of DNA in such a way that it is contained within the nucleus but also plays an important role in gene regulation (Felsenfeld, 1992; Onder et al., 2012). Recent studies show that many localized chromatin modifications such as methylation, acetylation, phosphorylation and ubiquitination impart positional information and regulate whether a particular DNA sequence in the genome will be transcribed or not (Wolffe, 1997; Wu, 1997; Wolffe, 2001; Onder et al., 2012). Such modifications either expose or hide regions of DNA for transcriptional regulation. Hence, eukaryotic transcription is quite complex because it is not only dependent on the DNA sequence but also the accessibility of the condensed genomic DNA to the transcriptional machinery, which involves chromatin remodeling (Segal et al., 2003; Suganuma and Workman, 2011). The regulation of transcription determines which pre-mRNAs are transcribed.

Pre-mRNA Processing

In eukaryotes, processing of precursor-mRNAs comprises a number of successive steps after initiation of transcription to generate functional mRNA. The three major processing events are 1) addition of a cap at the 5'end, 2) addition of a poly (A) tail at the 3) 3' end and pre-mRNA splicing, the removal of non-coding sequences (introns) and joining of coding sequences (exons) (Moore and Proudfoot, 2009). Pre-mRNAs from some genes produce a single mRNA (constitutive splicing, CS) or multiple mRNAs (alternative splicing, AS).

The 5'end is modified by adding a cap, which consists of a modified guanine nucleotide connected to the mRNA molecule by an uncommon 5' to 5' triphosphate linkage. This guanosine is methylated at the 7-position by a 7-methyl transferase and is denoted as 7-methylguanylate cap (m^7G). The cap-binding complex (CBC) that binds the cap has numerous functions in mRNA biogenesis including efficient splicing, mRNA export, protection of the transcripts from nuclease degradation, 3′-end formation and translation by stabilizing the interaction of the 3′-end processing machinery and translation (Lewis and Izaurralde, 1997). In Arabidopsis the single and double knockout mutants of the cap-binding protein (CBP) genes *CBP20* and *CBP80* are viable, but they displayed slow growth and late flowering with serrated leaf margins. These mutants are hypersensitive to abscisic acid, show increased drought tolerance, and exhibit changes in alternative splicing (AS) of a number of genes (Hugouvieux et al., 2001; Papp et al., 2004; Kuhn et al., 2008).

The 3’ end of most pre-mRNAs are polyadenylated by polyadenylate polymerase. Polyadenylation is necessary for stability of the mRNA, nuclear export and translation (Garneau et al., 2007). A multi-protein complex consisting of cleavage/polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), polyadenylate polymerase (PAP), polyadenylate binding protein 2 (PAB2), cleavage factor I (CFI), and cleavage factor II (CFII) cleaves the 3' end of the nascent RNA and polyadenylates it after cleavage by CPSF, 10–30 nucleotides downstream of its binding site (Moore and Proudfoot, 2009). The use of alternative polyadenylation (APA) sites generates different transcripts with altered coding capacity. In animals APA in some cases causes cancer and animal embryo abnormalities (Syed et al., 2012). Genome-wide studies have revealed that plants use APA extensively to generate diversity in their transcriptomes. Although each transcript produced by RNA polymerase II has a poly (A) tail, over 50% of plant genes studied possess multiple APA sites in their transcripts (Wu et al., 2011). The best-studied differential polyadenylated transcripts in plants are related to flowering time control pathways and stress responses (Simpson et al., 2003; Quesada et al., 2005).

Pre-mRNA splicing is an important aspect of gene regulation, which is discussed below in more detail.

Translational Regulation

Post-transcriptional processing and regulatory mechanisms control how efficiently an mRNA is translated to protein. Translational initiation is the main regulatory and rate-limiting step in translation, requiring over 25 proteins compared to a few for the subsequent steps: elongation and termination. The rate of initiation of translation is influenced by *cis*-elements in the mRNA, often in 5’ and 3’ UTRs (Holcik and Pestova, 2007). Following splicing, the pre-mRNA remains bound to a multi-protein exon junction complex (EJC) involved in mRNA export and nonsense mediated decay (NMD) (Le Hir et al., 2000; Le Hir and Andersen, 2008). NMD targets mRNA with nonsense mutations that introduce premature termination codons (PTCs) for degradation (Chiu et al., 2004), thus causing the ribosome to terminate prematurely and trigger mRNA degradation.

Coupling of different steps in gene expression

Different steps in gene expression (transcription, capping, splicing, and polyadenylation) are coupled and the transcriptional process in the nucleus significantly impacts downstream events of gene expression including translation in the cytoplasm (Maniatis and Reed, 2002; Reed, 2003 Tilgner, 2012 #14002; Lareau et al., 2007a; Tilgner et al., 2009). Recent studies have clearly established that transcription and posttranscriptional processes are coupled (Maniatis and Reed, 2002; Reed, 2003; Tilgner et al., 2009). For efficient coordination between splicing and transcription, the phosphorylated C-terminal domain (CTD) of RNA polymerase II (RNAPII) is required. A strong link between RNAPII elongation rate and alternative splicing outcome has been shown (Syed et al., 2012). The kinetic model of splicing proposed that slowing or pausing of RNAP II increases the window of time for a weak splice site to recruit the splicing machinery. Also the discovery that the majority of splicing events occur co-transcriptionally has led to the possibility that transcription and the state of chromatin play a role in alternative splicing regulation *in vivo*. Interestingly, trimethylated histone (H3K36me3) is more enriched in constitutive exons than in alternatively spliced ones, suggesting that distinctive subsets of histone modifications may regulate splicing patterns. Evidence for such epigenetic regulation of alternative splicing is accumulating over time (Luco et al., 2011). Studies also suggest a link between nuclear pre-mRNA processing and mRNA export and subsequent translation and mRNA degradation via NMD. Extensive coupling of AS with NMD functions is required to maintain optimal cellular protein concentrations, by eliminating erroneous mRNAs (Isken et al., 2008). During splicing, the EJC is deposited upstream of exon junctions. The EJC complex contains splicing-associated factors SRm160 and RNPS1, Y14 and binding partner Magoh, eIF4AIIIa (a DEAD-box RNA helicase), and Upf3 (Tange et al., 2004). The EJC serves as a binding platform for factors involved in mRNA export and nonsense mediated decay (NMD) (Le Hir and Andersen, 2008). SRs (e.g., ASF/SF2) bound to mRNAs enhance translation, thus also providing a link between splicing, nuclear export and translation (Gudikote et al., 2005). The spliced mRNA is recruited at a higher rate to the polysome compared to unspliced mRNA (Beilharz and Preiss, 2004).

PRE-mRNA SPLICING

In 1977 Phillip Sharp and Richard Roberts groups independently discovered the presence of non-coding intervening sequences (introns) in protein coding genes by performing RNA-DNA hybridization studies with adenovirus mRNA and its genomic DNA. This discovery led to a new paradigm that genes can be split with coding and noncoding regions. The Nobel Prize in Physiology or Medicine in 1993 was awarded to Richard Roberts and Phillip Sharp for this seminal discovery. With the completed genomes of many organisms we now know that a vast majority of eukaryotic protein coding genes (e.g., 80-90% of genes in photosynthetic eukaryotes (Labadorf et al., 2010)) contain one or more introns. The primary transcript (also called precursor mRNA) contains both coding regions (exons) and introns, which are excised precisely while the exons are joined to generate functional mRNAs by pre-mRNA splicing.

Constitutive versus Alternative Splicing

Multi-exon genes can produce a single transcript from a gene, which is referred to as constitutive splicing (CS) or they can produce more than one transcript from a gene by a process called alternative splicing (AS). During AS pre-mRNAs from a gene are spliced differently so that multiple transcripts generated from a single gene have different combinations of whole or part of exons and introns. Five major classes of alternative splicing occur in eukaryotes (Figure 1.1). These include exon skipping (an exon is either include or excluded, also called cassette exon), mutually exclusive exons (the inclusion of an exon precludes inclusion of adjacent exon), alternative 5' splice site selection, and alternative 3' splice site selection and intron retention. Two or more of these basic AS events can occur simultaneously in a pre-mRNA to generate other types of AS (e.g., exon skipping and intron retention or occurrence of both alternative 5' and 3' splice sites). Alternative splicing thus makes it possible for a single gene to produce more than one mRNA, and changes the one-gene one-enzyme paradigm as AS allows a single gene to code for multiple proteins. A majority of the AS events occur within the translated regions of mRNAs (Gupta et al., 2004; Stamm et al., 2005), dramatically affecting the properties of proteins in terms of enzymatic properties, protein interaction, localization and stability, as well as signaling or regulatory activities (for review, see (Stamm et al., 2005).

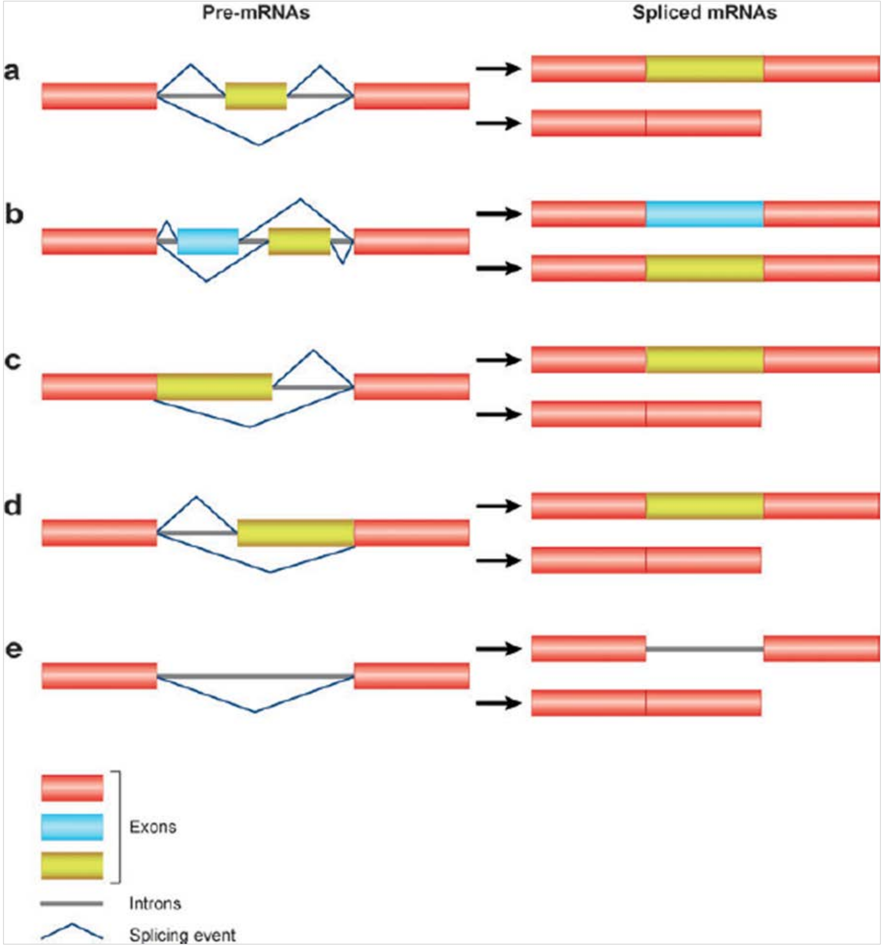


Figure 1.1: Different products of alternative splicing (AS) of pre-mRNA.
Different size transcripts can be formed by a variety of AS events that can occur singly or in combinations. a) Cassette exon, resulting from either inclusion or exclusion of exon. b) mutually exclusive exons. Different sized mRNAs formed by c) alternative 5’ splice site and d) alternative 3’ splice site usage. e) Intron retention, (adapted from (Reddy, 2007).

Spliceosome – a complex machinery that performs pre-mRNA splicing

Pre-mRNA splicing is carried out by the spliceosome, a large ribonucleoprotein complex that recognizes the sequences at the exon/intron and intron/exon boundaries called the 5’ donor site (GT) and 3’ acceptor site (AG) with invariant dinucleotides, as well as the branch point and the polypyrimidine tract in the intron. The 5’ and 3’ splice sites (ss) are quite conserved between plants whereas the polypyrimidine tract that is enriched in both U and C in animals is enriched mostly with U in plants (Figure 1.2) (Reddy, 2007).

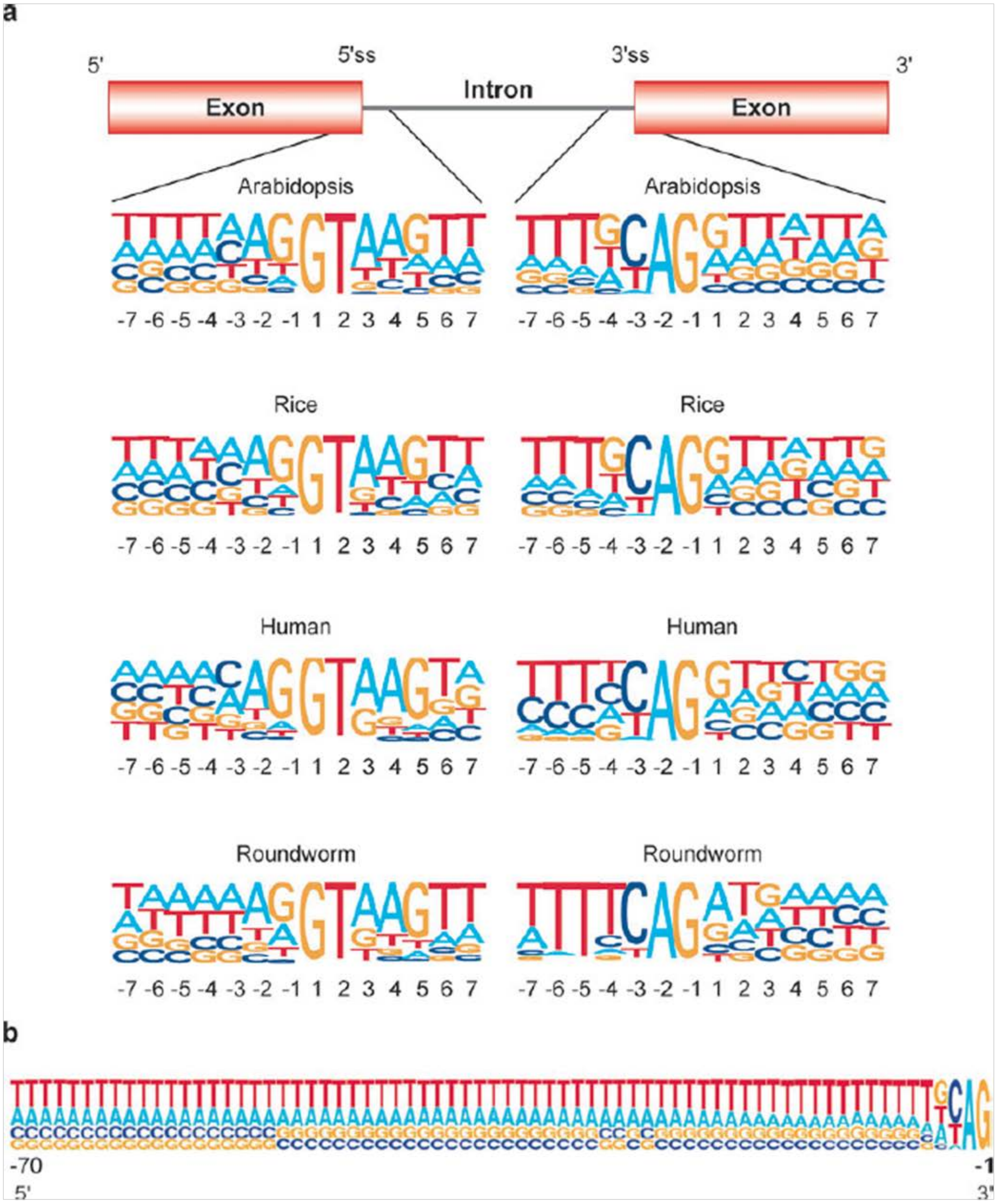


Figure 1.2: Splice site (SS) signals at introns.

(a) Sequence logo of the introns in plants and animals. Schematic representation of intron flanked by exons and the nucleotide frequencies given by the height at each position. b) Sequence logo at the 70 nucleotide 3’ end of Arabidopsis introns. Data was taken from annotated sequences of the organisms shown [adapted from (Reddy, 2007)].

In plants, as in animals, there are two types of spliceosomes. The major type is called the U2 type, which performs splicing of U2-dependent introns, whereas the minor U12 type is involved in splicing rare U12-dependent introns (Shukla and Padgett, 1999; Simpson and Brown, 2008)}. Both spliceosomes consist of five snRNAs (U1, U2, U4, U5, U6 in the major spliceosome and U11, U12, U4atac, U5, and U6atac in the minor spliceosome). Each snRNA binds to proteins to form snRNPs, which together with many non-snRNP proteins regulate splicing (Valadkhan and Jaladat, 2010). The spliceosome, which contains about 300 proteins, one of the most complex cellular machines, performs to transesterification reactions to excise an intron and join two exons (Rappsilber et al., 2002; Jurica and Moore, 2003; Wahl et al., 2009). A schematic diagram showing the roles of different snRNPs in spliceosome assembly is presented in Figure 1.3 (Koncz et al., 2012). Although plant spliceosomes have not been isolated so far, extensive bioinformatics analyses for spliceosomal components have been done using the predicted proteome of completely sequenced genomes of plants (Wang and Brendel, 2004; Ru et al., 2008; Koncz et al., 2012)}. Based on the conservation of most of the RNA and protein components in the spliceosome between mammalian and plant systems, it is thought that plant spliceosomes are likely to be similar to their animal counterparts. The spliceosome is assembled in an ordered, stepwise manner, during which complexes E, A, B, and C are formed on the pre-mRNA to catalyze the removal of introns and ligation of exons (Stark et al., 2001; Jurica et al., 2002). During the first step of spliceosome assembly, U1snRNP recognizes the 5’ ss and the U2snRNP auxiliary factor (U2AF), a dimer consisting of U2AF35 and U2AF65, recognizes the 3’ ss along with the polypyrimidine tract to form the early (E) complex. U2AF then recruits U2snRNP to the branch point to form prespliceosome complex A. The U4/U5/U6 tri-snRNP then joins complex A to form precatalytic complex B. The association of NineTeen Complex

(NTC), which was isolated in animal and plant systems, with complex B converts this into activated B complex (Koncz et al., 2012) and U1 and U4 snRNP are released. The activated B complex is then converted into complex C in which two trans-esterificaiton reactions take place, which involves conformational rearrangements. The assembly and disassembly of spliceosomes, involving a series of conformational rearrangements of its components at each step, requires ATP-dependent DExH/D-box RNA helicases, which are not shown in Figure 1.3.

Extent of Alternative splicing

During the last decade the extent of AS has been studied extensively in both plants and animals. Before the advent of all the high-throughput technologies that can be used to understand AS events, there had been reports of AS in both plant and animals but AS was considered rare. The first reports of AS were on pre-mRNAs of globin and adenovirus genes (Felber et al., 1982; Mariman et al., 1983). AS was observed and studied in plants soon after, initially in the study of transposons and their effect on expression of adjacent genes. Generation of splice variants was observed in maize Adh1 with an insertion of a Ds2 transposable element (Simon and Starlinger, 1987). The insertion of a 1.3kb Ds2 transposon in the maize Adh1 gene was shown to produce two transcripts of 3 and 1.6 kb, the large one with the Ds2 and smaller one without Ds2 (Simon and Starlinger, 1987). Similarly, a 409 bp Ds1 transposon in the maize waxy (Wx) gene was shown to generate several Wx transcripts due to AS of Ds1 sequences from Wx pre-mRNA. It was also suggested that these features might enhance the ability of Ds1 to function as a mobile intron (Wessler, 1991). Among plant protein-coding genes without a transposable element, ribulose bisphosphate carboxylase/oxygenase (rubisco) activase of Arabidopsis and spinach were the first genes shown to undergo AS and produce two distinct polypeptide that differ in 37 amino acids (Werneke et al., 1989).

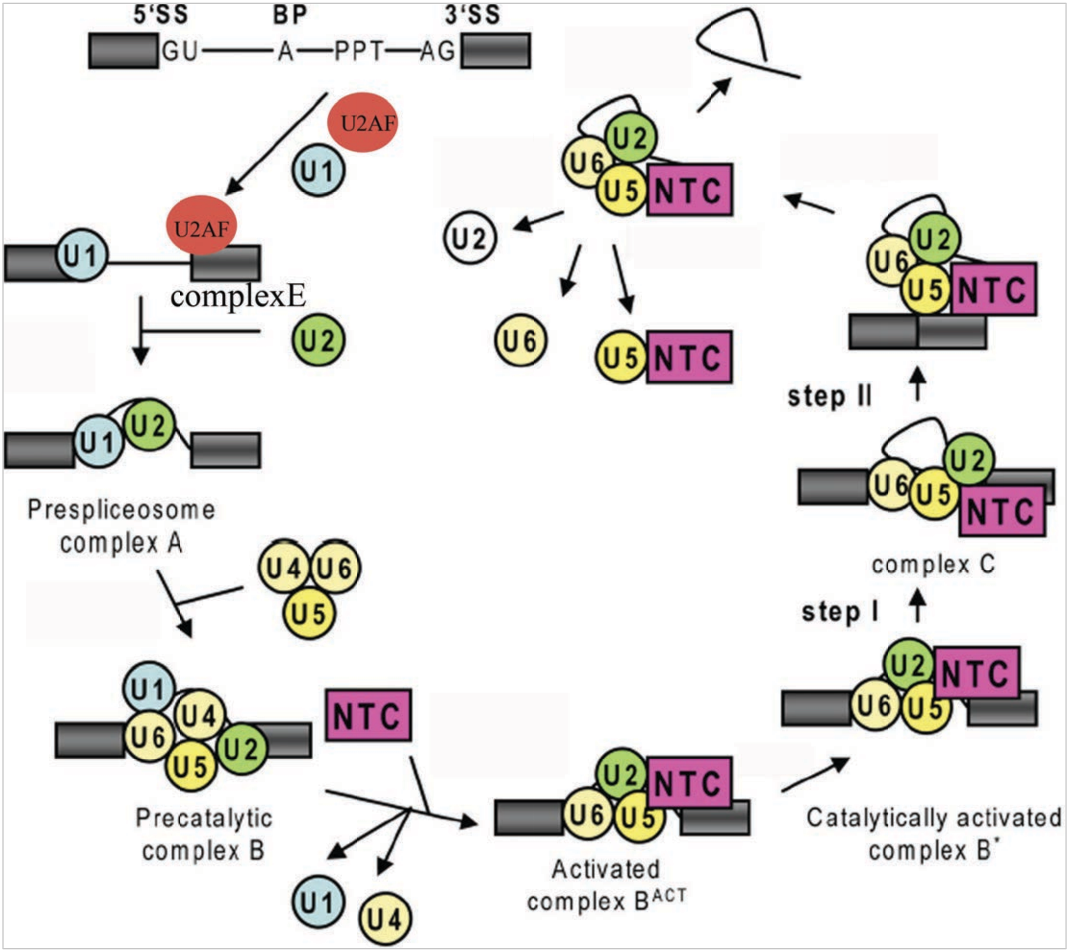


Figure 1.3: **Schematic representation of spliceosomal assembly.**

The dynamics of spliceosomal assembly based on mammalian systems forms the basis of the model for pre-mRNA splicing in Arabidopsis. Splicing is catalyzed by the spliceosome, a large RNA-protein complex comprised of 5 small nuclear ribonucleoprotein particles (snRNPs): U1, U2, U4, U5, and U6. Shown above is the canonical splice site (SS) of an intron containing the signals GU at 5' SS, AG at 3' SS, branch point (BP) and polypyrimidine (PPT) tract. U1 binds to the 5' SS with accessory proteins and U2AF (U2snRNP auxillary factor), and forms the E-complex. U2 snRNP is recruited to the branch point through interactions with the E-complex components and possibly U1 snRNP to form the A-complex. The U4/U5/U6 tri-snRNP is recruited to the assembling spliceosome to form complex B. NTC (Arabidopsis 19 complex) is bound to the B-complex to form an activated complex B'. Following several rearrangements, complex C (the spliceosome) is activated for catalysis and later the U2/U5/U6 remains bound to the lariat, the 3' site is cleaved and the exons ligated. Exons are indicated by gray boxes, thin black lines show intron and intron lariat (adapted from (Koncz et al., 2012)).

Identification of AS using different technologies

During the last decade a large number of tools have been developed to predict alternative splicing globally. Prior to high throughput next generation sequencing (NGS) technologies, three main techniques were used to analyze transcriptome data for splicing patterns - expressed sequence tags (ESTs), splice-junction arrays and genome tiling microarrays (Johnson et al., 2003). During the last few years RNA sequencing (RNA-Seq) using NGS has been used increasingly (Pan et al., 2008; Wang et al., 2008; Graveley et al., 2011). The EST centered, Sanger method of low throughput first generation sequencing was the first method used to sequence cDNA and ESTs (300-800bp in length). The ESTs/cDNAs are aligned onto the gene sequence to predict AS. The alignment of independent ESTs/cDNAs arising from the same gene allowed the identification of AS, as shown in humans where the first studies showed 133 out of 392 known genes underwent AS (Mironov et al., 1999), which was the first glimpse of the extent of AS in human genes. Soon after, an early draft of the human genome was used to align 2.1 million ESTs and mRNA sequences, identifying 6,201 alternative splices in 2,272 genes (Modrek et al., 2001), establishing the prevalence of AS in the human genome. Moreover, 70–88% of AS events were predicted to change the protein product due to replacement of the amino or carboxy terminus, in-frame addition or removal of a functional module. Only 19% caused a frameshift and truncation of the protein (Modrek et al., 2001). For a systematic analysis of AS in a sequenced genome, splice-junction microarrays have been used to quantitatively analyze AS events (Johnson et al., 2003). Oligonucleotides designed across the splice junctions connecting exons are used to query the presence of transcripts across the splice junctions and identify AS by hybridization. However, this method is limited by the extent of prior knowledge of the AS variants with gene structures, but is a useful validation of known AS discovered by other methods.

One of the surprising findings of the human genome project is that the gene content in humans (about 21000 genes) is very similar to much simpler and less complex organisms such as *C. elegans* (Modrek and Lee, 2002; Hayden, 2010). The lack of an association between gene number and organismal complexity has resulted in an increased interest in alternative ways for an organism to evolve proteome diversity. Alternative splicing (AS) has been proposed to be a major factor in expanding the transcriptome and proteome diversity, adding to regulatory and functional complexity, and thereby organismal complexity (Xing and Lee, 2007; Power et al., 2009).

The advent of NGS (Niedringhaus et al., 2011) using diverse methods - pyrosequencing (Roche 454), sequencing by synthesis (Illumina GAII) or sequencing by ligation (ABI SOLiD) - have provided high-throughput platforms for generation of large amounts of DNA sequence information quickly at a low cost. The sequencing of cDNA fragments termed RNA-Seq (RNA sequencing) is done by preparing cDNA from the target tissue/organism and using short cDNA fragments for sequencing by one of the NGS methods (Figure 1.4). This can provide millions of transcript sequences that could be used to analyze AS (Wang et al., 2009). Subsequently, the RNA-Seq data is processed by appropriate programs such as TopHat that first aligns the sequence reads to a reference genome and then assembles the clusters into transcriptional units by a suite of tools called Cufflinks that identify and quantify AS products (Trapnell et al., 2010). RNA-Seq has been used successfully to precisely quantify transcript levels, confirm or revise previously annotated 5' and 3' ends of genes, and map exon/intron boundaries (Trapnell et al., 2010; Roberts et al., 2011). This method has emerged as a powerful technology for transcriptome analysis and mining for in-depth transcriptional landscape by producing millions of short reads and also in identifying rare splice variants.

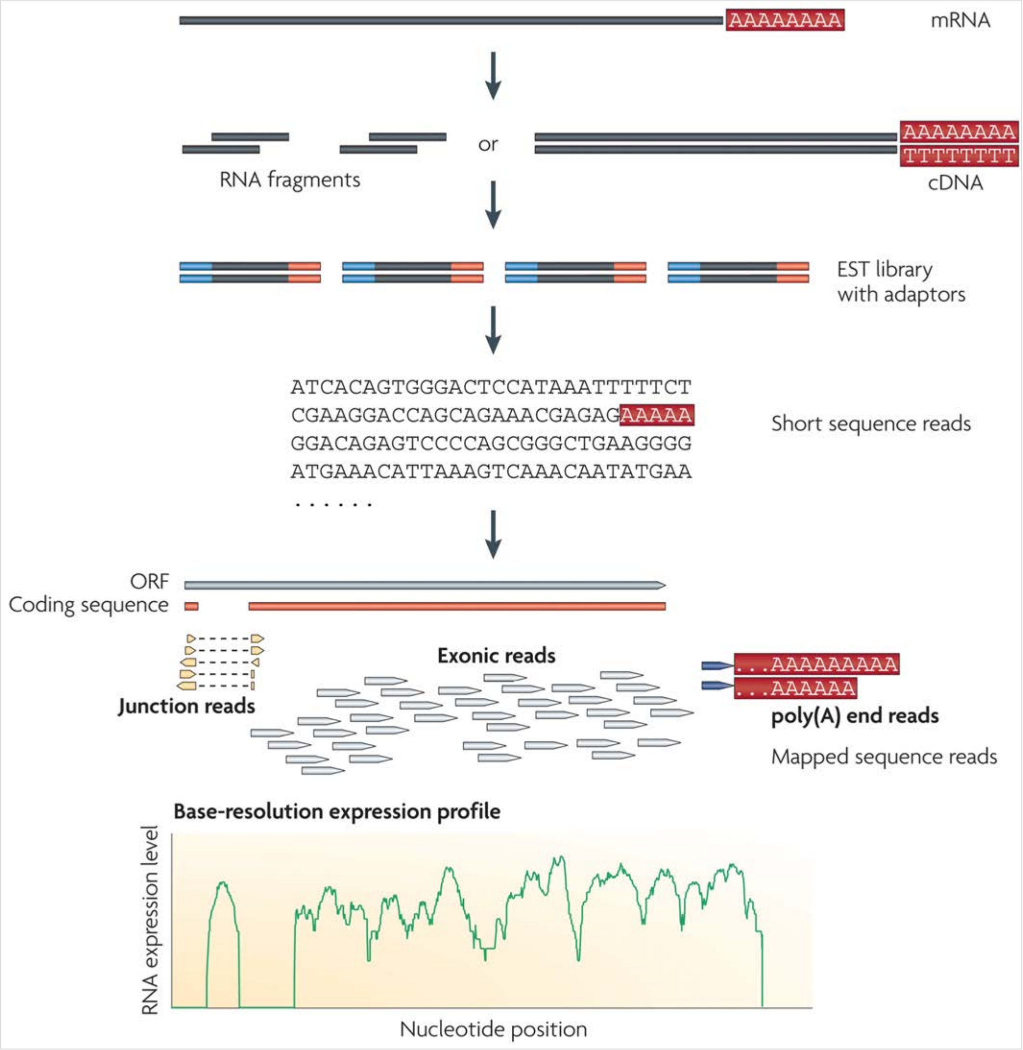


Figure 1.4: RNA-Seq for transcriptome analysis.
In the high-throughput method of transcriptome analysis, cellular mRNA is converted into a library of cDNA fragments, through RNA or cDNA fragmentation. Sequencing adaptors (blue) are added to the cDNA fragments that are then sequenced using high-throughput next generation sequencing technology. The resulting sequence reads are aligned with the reference genome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. The splice-junction reads shows the intronic region with a very low RNA expression level pattern, while the exonic regions on the profile mapping indicate high RNA expression level. These three types are used to generate a base-resolution expression profile count for each gene, as shown by the example at the bottom; a yeast ORF with one intron [adapted from (Wang et al., 2009)].

In depth analysis of AS in plants has greatly increased in recent years as more RNA-Seq analyses are performed (Figure 1.5). The most recent deep transcriptome analysis in Arabidopsis has revealed that over 60% of intron-containing genes produce two or more transcripts (Marquez et al., 2012). Similar studies in rice also predicted AS to be about 50% in the rice transcriptome and also a significant level of trans-splicing was also discovered (Lu et al., 2010; Zhang et al., 2010). In a comprehensive analysis of the Arabidopsis genome, 6,772 introns that exhibit tandem acceptor sites (NAGNAG) were identified. These sequences are quite widely distributed in SRs and show that out of 36 identified introns there are in 30 SR and SR-related protein-coding genes with a NAGNAG acceptor. After experimental analysis, eight out of the 15 candidates studied showed differences in AS under several different seedling growth stages and tissue types, with the most pronounced effect seen under cold stress conditions (Schindler et al., 2008).

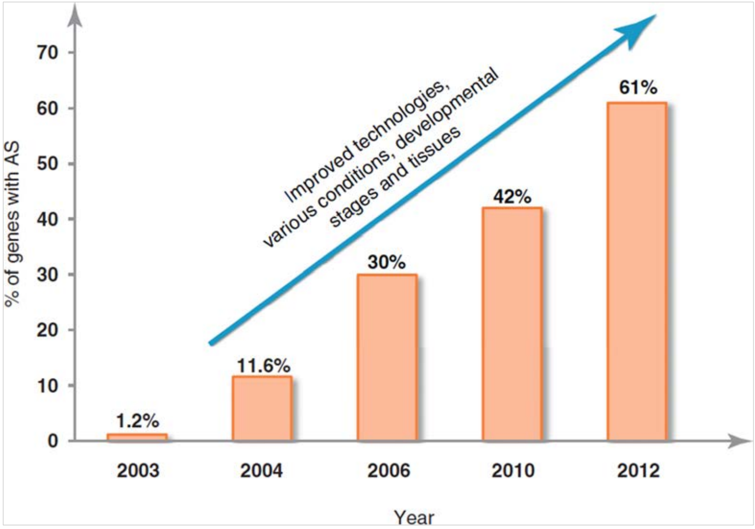


Figure 1.5: Estimates in alternative splicing (AS) in Arabidopsis over time.
Early studies in 2003 with expressed sequence tags (ESTs) estimated 1.2% AS, which increased with more sequence information available from deep EST and cDNA sequencing. The advent of high-throughput next generation sequencing (NGS) has contributed to much greater coverage of the transcriptome to provide an overall 50-fold increase in estimates of AS during the last decade [adapted from (Syed et al., 2012)].

Recent global transcriptome studies are not only providing the extent of AS in eukaryotes but also indicate that plant and animals differ considerably in the most prevalent types of AS. For example, in plants a vast majority of splice variants (up to 56%) are due to intron retention, whereas it is not that prevalent in metazoans (5% in humans) (Iida et al., 2004; Ner-Gaon et al., 2004; Wang and Brendel, 2006a; Baek et al., 2008; Filichkin et al., 2010; Labadorf et al., 2010). In contrast, exon skipping is the most common form of AS in animals (58% in humans), which is less prevalent in plants (8% in Arabidopsis). Such differences point to the idea that the mechanisms by which spliceosomal machinery recognizes introns and exons may differ in plants and animals.

Two models – exon definition and intron definition – have been proposed to illustrate the mechanisms by which splice sites are recognized (Figure 1.6). In the exon definition model, splice sites are predominantly recognized across the exon and that involves initial interaction across the exon between factors recognizing the 5'ss and the upstream 3'ss (Sterner et al., 1996). In the alternative intron definition model, interactions occur first across the intron between factors recognizing the 5'ss and the downstream 3'ss (Berget, 1995). Current thinking is that splicing of pre-mRNAs with long introns as in humans use exon definition whereas those with small introns as in plants use the intron definition model. It is proposed that the small size and composition (U or UA richness) of plant introns contribute to the intron recognition model. The U-rich regions are recognized by UBPs (UBP1, RBP45, and RBP47) and are essential for efficient splicing. Mutations in U-rich regions in a number of plant introns alter splicing efficiency and can activate cryptic splice sites. The U or UA code also is a major factor for CS, as introns lacking this undergo AS events more frequently (Brown et al., 2002; Simpson et al., 2004).

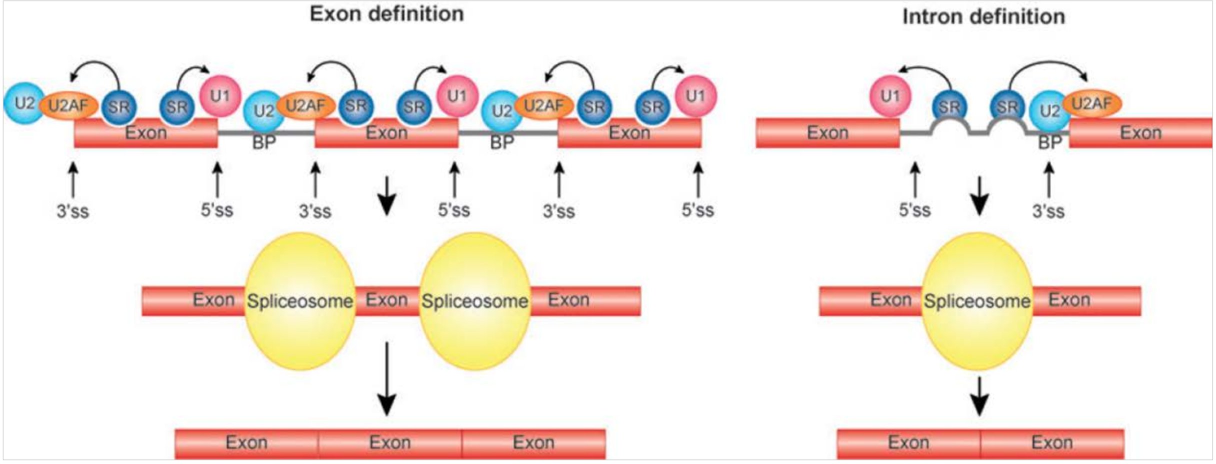


Figure 1.6: Exon and intron definition models of pre-mRNA splicing.
In the exon definition model, the splicing machinery recognizes splicing regulatory elements across the exon. SR proteins and other splicing regulators that bind to exonic splicing enhancers (ESE, shown as half circles) assemble U1 snRNP to the 5’ SS and U2AF to the 3’ SS, and subsequently U2 snRNP to the branch point (BP). In the intron definition model, the splicing regulator proteins recognize intronic splicing regulator (ISR) sequences (half circles) that assemble the U1 and U2AF snRNPs to the splice sites [adapted from (Reddy, 2007)].

Regulation of Splicing

Although most of the core sequence elements involved in spliceosome assembly are conserved across species and are necessary for splicing, they alone are not sufficient as these sequences are short and loosely conserved in higher eukaryotes. Hence, in most cases, there are additional sequences in exons and introns, which are collectively referred to as splicing regulatory elements (SREs) that are important for constitutive splicing as well as alternative splicing. These are auxiliary splice signal sequences (*cis*-elements), such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) that were shown to aid in CS as well as complex AS events

(Chasin, 2007; Wang and Burge, 2008; Barash et al., 2010). The lack of in-vitro splicing techniques in plants has slowed down progress in understanding splicing regulation in plants. Also animal splicing assays are not useful to study plant splicing as plant introns are not accurately spliced in animal splicing extracts. Because of these reasons splicing regulation in plants has to be studied exclusively *in vivo* with the help of the latest high throughput techniques. Below I describe the factors that contribute to CS and AS in animals and also discuss this in relation to plants so that we can eventually have a better understanding of highly complex splicing regulation that is critical for plant growth, differentiation and response to environmental conditions.

Differences in gene architecture between plant and animals

The architecture of plant genes is one reason for differences in the prevalence of types of AS events between plants and mammals. Plant genes are generally shorter, with shorter introns compared to animals. The average size of animal introns is 3000 nucleotide (nt) compared to 173 nt in plants (e.g. Arabidopsis), whereas the size of exons are quite similar with 140nt in animals and 172nt in plants. Consequently, it is thought that the small size of introns contributes to their retention as it is more prevalent in plants. Also, the intron definition model is believed to operate in plants, whereas the exon-skipping prevalence in animals is explained by the exon definition model. There are some exceptions to this as there are some long introns in plants that undergo a higher rate of intron retention. Conversely the puffer fish with small introns of 600bp undergoes exon skipping. In general very little is known about SREs in plants. Therefore, more efforts are needed to understand the putative *cis*-elements defining the differences in splicing between plants and animals.

SPLICING REGULATORY ELEMENTS

Methods to identify AS elements and splicing code

Splicing signals along with combinatorial action of transacting factors, mainly RNA-binding proteins, modulate the assembly and activity of the spliceosomal complex. A major focus now is on understanding the molecular mechanisms of AS with respect to sequence elements and on cracking the splicing code. To identify *cis*-elements i.e., the targets of RNA-binding splicing regulators, many high-througput methods were developed. One method termed selective evolution of ligands by exponential enrichment (SELEX) identifies the optimal binding site for an RNA binding protein by iterative selection from a pool of degenerate motifs (Turek and Gold, 1990; Coulter et al., 1997). More direct *in vivo* strategies have been developed to characterize the interaction of proteins with RNA by isolation of protein-RNA complexes from *in vivo* is the most direct method. A method called CLIP (crosslinking and immunoprecipitation), uses UV-C light to create a covalent bond between the protein and RNA at positions where they are bound, and an antibody to a specific protein is used to immunoprecipitate RNA-protein complexes. RNA from this complex is extracted to make cDNA and sequenced (Ule et al., 2003). In HITS-CLIP (high-throughput sequencing CLIP) RNAs are sequenced using NGS (Ule et al., 2005). A modified version of CLIP is PAR-CLIP that stands for photoactivatable ribonucleoside-enhanced CLIP that incorporates 4-thiouridine into RNA prior to crosslinking using UV-A light (Hafner et al., 2010a, b). In another method iCLIP (individual nucleotide resolution CLIP) RNAs cross-linked to short peptides are isolated to produce cDNAs with a truncation at the crosslink sites, and high-throughput sequencing of the truncated cDNAs positions the protein–RNA crosslink site to high resolution (Konig et al., 2010).

The study of genome-wide protein–RNA interactions provide only a part of the information for splicing regulation, since protein-binding sites are often located far from alternative exons, as shown by the first analyses of Nova–RNA interactions using CLIP (Konig et al., 2010). Often the RNA motifs recognized by RBPs are degenerate and occur frequently in pre-mRNAs, e.g. Nova proteins recognize the motif YCAY (Y represents pyrimidine) in clusters of multiple tetramers (Buckanovich and Darnell, 1997), many of which turn out to be non-functional. The real utility of genome-wide studies is derived by the integration of multiple, independent data sets, such as genome-wide protein–RNA interaction sites to generate 'RNA splicing maps', which determine the position-dependent regulatory effects of protein–RNA interactions. The initial approach was the integration of bioinformatically identified Nova-binding sites with splicing profiles identified by splice-junction microarrays, and later a more specific splicing map was obtained by protein–RNA interaction sites determined HITS-CLIP.

ESEs comprise a variety of sequences present in most exons in animal systems (Schaal and Maniatis, 1999b, a; Fairbrother et al., 2002). SR proteins are involved in regulation of splicing by binding ESEs through their N-terminal RRM domains, mediating protein–protein interactions through C-terminal RS domains and facilitating spliceosome assembly (Graveley et al., 1998). ESSs are a diverse variety of sequences often bound by a class of hnRNP splicing repressors that function in a variety of ways: hnRNP I by blocking interactions between U1 and U2 snRNPs (Izquierdo et al., 2005), hnRNP A1 by inhibition of splicing by binding to and looping out exons or by displacing snRNP binding (Zhu et al., 2001).

In animals, intronic SREs can include the G triplet (GGG) or G run motifs in clusters, and enhance recognition of adjacent 5'ss or 3'ss (McCullough and Berget, 1997), CA repeats in introns can enhance splicing of upstream exons through binding of hnRNP (Hui et al., 2005), UGCAUG hexanucleotide ISEs are bound by splicing factors Fox-1/Fox-2 (Underwood et al., 2005), and YCAY motif pairs can function as either ESSs or ISSs (Hui et al., 2005). A mammalian splicing code has been recently developed that includes new classes of splicing patterns, to identify regulatory programs in different tissues, and mutation-verified regulatory sequences (Barash et al., 2010).

In plants with smaller introns, the landscape for prediction is different and studies to identify ESE motifs in Arabidopsis were initiated (Pertea et al., 2007). They extracted 50bp of the ends of internal exons of Arabidopsis genes with high-quality gene models, and identified potential ESE hexamers based on assumption of having higher frequency in exons than introns, with higher frequency in exons with weak splice sites. Around half of the ESE hexamers identified had experimental evidence including motifs like GAAGAA that is validated in humans. The Arabidopsis glycine-rich RNA-binding proteins AtGRP7 and AtGRP8 were shown to regulate AS of their own pre-mRNA and of each other and *cis*-elements that bind to AtGRPs have been identified (Schoning et al., 2007; Schoning et al., 2008).

Researchers are also developing computational methods to predict SREs around 6-8 nucleotides in size that bind to RNA-binding proteins. A computational method called RESCUE (Relative Enhancer and Silencer Classification by Unanimous Enrichment) (Fairbrother et al., 2002) has been applied to many genomes for finding motifs enriched near weak splice sites of CS introns based on the principle that SREs compensate for poor splice-site recognition. RESCUE combined with the rationale that SREs are likely evolutionarily conserved was used to predict ESEs in Drosophila, of which 58% of these putative SREs were found identical to those in human, mouse, or pufferfish ESE sequences (Brooks et al., 2011). The prediction analysis identified 22 hexamers near the 5’ss and 34 hexamers in exons near the 3’ss of short introns as putative ESEs, with five sequences enriched near both splice sites (CTGGAG, CTGGAT, CTGGAA, CCTGGA, GGAAAC). Similarly, putative ISEs were identified using the RESCUE method in Drosophila, by searching for enriched hexamers within introns relative to exons, and enriched near weak splice sites relative to strong splice sites. A similar fraction of ISEs (59%, 136 out of 231) compared to ESEs (58%: 57 out of 99) were found conserved between Drosophila and one or more vertebrates, but only 2 ISEs (CTCTCT and TTATAA) were identical in all four species.

Regulation of splicing by RNA structure

Some studies suggest that the secondary structure of RNA may also play a role in regulating AS (reviewed in Reddy et al., 2012). Lately, high throughput technologies such as parallel analysis of RNA structure (PARS) (Kertesz et al., 2010) and fragmentation sequencing (Frag-Seq) (Underwood et al., 2010) have been used to unravel the structures of pre-mRNA at a very high resolution and eventually to understand the dynamic interplay between splicing regulatory proteins and RNA structure in controlling both CS and AS (Kertesz et al., 2010) (Isaacs et al., 2006). Since initial pre-mRNA folding occurs co-transcriptionally, for subsequent spliceosomal recognition of the 5’ ss, 3′ ss, and branch point, RNA structure plays an important role (McManus and Graveley, 2011). The *cis*-elements of the EDA exon of the fibronectin gene are sequestered by folding of the RNA molecule and this affects the availability of ESS and ESE to SR proteins (Muro et al., 1999). The most striking example of RNA structure in alternative splicing comes from the Drosophila *Dscam* gene by maintaining an incredibly large number (about 38,000) of Dscam RNA isoforms. Competing RNA secondary structures play a role in mutually exclusive splicing of 4 cassette exon clusters (May et al., 2011).

The riboswitch mechanism of small RNAs sensing metabolites in response to binding specific small molecules has gained attention because of the structural changes in RNA and its association with AS (Bocobza et al., 2007; Batey, 2012). Recent studies have identified thiamine pyrophosphate (TPP) binding riboswitches in fungi, algae and plants (Cheah et al., 2007; Croft et al., 2007; Wachter, 2010). In plants a post-transcriptional mechanism that uses a riboswitch to control a metabolic feedback loop through differential processing of the precursor RNA 3' terminus has been reported (Wachter et al., 2007). When cellular thiamin pyrophosphate (TPP) levels rise, metabolite sensing by the riboswitch located in TPP biosynthesis genes directs formation of an unstable splicing product, and consequently TPP levels drop (Cheah et al., 2007).

Alternative splicing and NMD

NMD is a surveillance mechanism that recognizes and removes mRNA containing Premature stop codons (PTCs) (Maquat, 2004), and it can regulate gene expression (Lejeune and Maquat 2005). PTC are found within some splice isoforms that do not get translated but are targeted for nonsense-mediated decay (NMD) (Belgrader et al. 1994), during the cell's RNA surveillance process (Lewis et al., 2003; Maquat, 2004). This regulatory mechanism termed RUST (regulated unproductive splicing and translation) can modulate the expression of proteins. The process was first shown to occur in *Caenorhabditis elegans* (Morrison et al. 1997; Mitrovich and Anderson 2000), and recent analysis of rice and Arabidopsis AS events suggest that more than one-third of splice variants have a PTC and are likely targets of NMD. As much as half of all intron retention events are also candidates for NMD (Palusa and Reddy, 2010; Kalyna et al., 2012). The analysis of 270 AS genes (950 transcripts) showed that 102 transcripts from 97 genes (32%) were identified as NMD targets. Recently the GRP7/8 and SOC1 plant genes, involved in the circadian clock and flowering control respectively, were shown to be regulated or putatively regulated by AS/NMD.

AS and fate of isoforms

Although the extent AS is very high we are yet to understand the significance of thousands of newly discovered splice variants. Research in this area should address whether these splice variants make functional protein, act as activators or suppressors to regulate other genes, or whether these are produced to balance the production of excess functional products, and lastly if they are just by products of sloppy splicing. AS cannot only aid in proteome diversity but also generate truncated proteins that may play other regulatory roles. AS, therefore, is likely to be involved in many plant processes such as seed germination, disease resistance, flowering time and the circadian clock; as well as physiology, metabolism, and responses to environmental conditions, all of which have important consequences on adaptation of plants to their environment and can lead to improvement of crop plant traits (Wang and Brendel, 2006a; Chen et al., 2007; Zhu et al., 2007; Ali and Reddy, 2008b; Lorkovic, 2009; Xu et al., 2011).

Researchers have been using several different approaches to understand the functions of these AS variants. The first approach is a tedious process of experimentally testing the association of each isoform with function. There are few good examples of genes associated with isoform functions and these are: Rubisco activase (the first gene in plants that was discovered to be alternatively spliced), different FCA isoforms, some defense related genes (MLA gene in barley & Arabidopsis R gene), waxy gene isoform in rice, and SR45 isoforms.

However, splice variants may not code for proteins but play a role regulating the level of functional proteins. In the second approach, a high-resolution AS RT–PCR was used to identify endogenous AS isoforms, which increase in abundance when NMD is impaired in the

Arabidopsis NMD factor mutants *upf1-5* and *upf3-1*. Alternative splicing is a major determinant in the production of variant mRNA transcripts some of which contain PTCs and might be targeted by NMD. One way to ensure mRNA quality control mechanisms is NMD, which degrades mRNAs that possess a premature termination codon (PTC+). This process of regulation is probably more efficient than either turning off overall transcriptional machinery or synthesizing a truncated protein and later processing it for degradation. Also, this post-transcriptional mechanism can fine-tune the relative levels of mRNA isoforms from a gene, which are either productive (protein-coding) or unproductive AS variants and thus regulates the levels of functional proteins. About 13–18% of Arabidopsis intron-containing genes are potentially regulated by AS/NMD. This compares well to the 14% and 20% reported for Drosophila and *Caenorhabiditis elegans*. Recent examples of plant genes regulated or putatively regulated by AS/NMD splice-variants are GRP7/8 (cross/auto-regulating components of circadian clock genes) and SOC1 (flowering control), SRs and PTB protein splicing factors (involved in a range of developmental and stress response processes) and HSF2a (a heat shock).

Interestingly, the majority of intron retention transcripts that were analyzed were not turned over by NMD despite containing PTCs with a downstream splice junction or the ones with long 3′-UTRs. The third approach is to uncover the fate of the transcripts that do not undergo NMD and explore whether they enter the translational machinery and form functional truncated proteins or eventually get degraded. PTC-containing transcripts have the potential to be translated into truncated proteins or peptides. Intron-containing mRNA transcripts with a PTC were shown to be associated with ribosomes in plants and are coded into peptides lacking certain domains so that they can act as act as both positive and negative regulators and affect regulatory feedback loops (Ingolia et al., 2011). Such regulation exists in both animals and plants and are called small interfering peptides (siPEPs) or micro-proteins (miPs) named after their analogy with siRNAs and miRNAs (Seo et al., 2011a; Seo et al., 2011b; Staudt and Wenkel, 2011). One of the examples of AS/miP- dependent strategy was studied in the transcription factor gene IDD14 (Seo et al., 2011b). An alternatively spliced IDD14 form (IDD14β), which is produced predominantly under cold conditions, lacks the functional DNA-binding domain but is able to form heterodimers with the functional IDD14 form (IDD14α). IDD14α/β heterodimers have reduced binding activity to the promoter of *Qua-Quine Starch* (*QQS*) gene. A similar regulation was also found in cases of cold and circadian associated (CCA) gene in Arabidopsis (Seo et al., 2012).

AS in generating variation

Acquisition of new functions (neofunctionalization) of duplicated genes is thought to be crucial in driving the evolution of developmental and morphological complexity in vertebrates. Likewise, it has been proposed AS could also play a role in evolution of eukaryotes by increasing the protein diversity (Kopelman et al., 2005). There are examples that AS is lost as the gene duplicates: the common ancestor of mangrove and popular had the gene encoding the chloroplast ribosomal protein RPL32 transferred to the nuclear genome and inserted into the last exon of a Cu-Zn superoxide dismutase (SOD). This chimeric gene undergoes AS to produce these two separate gene products. After divergence from mangrove the chimeric gene duplicated and lost its AS ability and sub-functionalization; the daughter gene encoding either RPL3 or SOD (Ueda et al., 2007).

Analysis of the sequence diversity between 18 different ecotypes/accessions in Arabidopsis has shown variation in the protein coding region of genes or potentially generated protein isoforms in different accessions (Gan et al., 2011). The sequence variation in these 18 accessions was also observed in disruption of 2572 splice sites and the RNA-binding motifs for splicing factors, which can impact protein expression and activity, and is proposed to be a basis for selection for adaptation of different ecotypes to their environments (Gan et al., 2011).

SR PROTEINS - SPLICING REGULATORS

There are many of RNA binding proteins (RBPs) in the dynamic structure of the spliceosome that help in splicing (Matlin and Moore, 2007). An important class of RBP involved in splicing in animals and plants are the serine/arginine (SR)-rich SR proteins. In animals, the SR proteins in combination with other splicing factors play a major role not only in CS and AS but numerous other processes like mRNA export, RNA stability, NMD, mRNA surveillance, and also as a carbohydrate binding protein on the cell surface (Bourgeois et al., 2004; Hatakeyama et al., 2009; Twyffels et al., 2011). A classical SR protein family member is defined by four main criteria: i) structural similarity, ii) dual function in CS and AS through complementation of splicing deficient S100 HeLa cytoplasmic extracts or in an alternative splicing assay, iii) the presence of phospho-epitope recognition in the RS domain by mAb104; and iv) purification using magnesium chloride (Long and Caceres, 2009). Recently, the SRs have been redefined and a standardized nomenclature has been adopted for both plant and animal SRs. All SR proteins have a modular structure consisting of one or two N-terminal RNA recognition motifs (RRMs) and a variable length C-terminal domain rich in serine and arginine residues (the RS domain) of at least 50 amino acids with > 40% RS content (Barta et al., 2010; Manley and Krainer, 2010).

Because SR proteins have been functionally associated with the regulation of alternative splicing, it could be expected that the number of SR protein family members would increase with increased prevalence of alternative splicing. There are only two SR proteins in the fungus *Schizosaccharomyces pombe* and multiple SR proteins are expressed in plants and metazoans. Among multicellular organisms, humans are considered to be the most complex, but they have only 11 SRs (Figure 1.7). However, plants possess the most SR proteins amongst any organisms studied, with Arabidopsis encoding 18 SRs, rice 22 and soybean with 25.

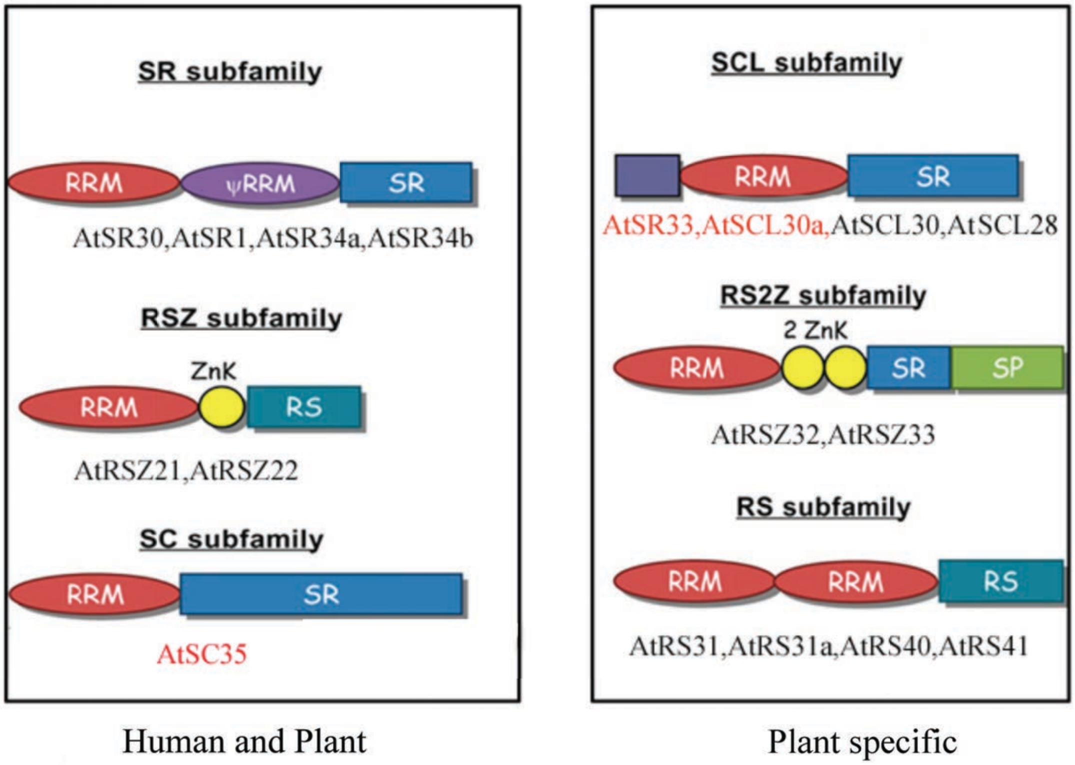


Figure 1.7: Domain architecture of the Arabidopsis SR proteins.
Left, SR protein subfamilies SR, RSZ and SC have orthologs in humans (SR ortholog SRSF1), (RSZ ortholog SRSF7) & (SC ortholog SRSF2). Right, the plant specific SCL family (SC-35 like) is quite similar to the human (SRSF2/SC35) RRM domain (in red) but differs in an N-terminal charged extension of arginine, proline, serine, glycine and tyrosine amino-acids (purple). The proteins of the plant specific RSZ subfamily possess two Zn-knuckles and have an additional SP-rich domain. The plant specific RS family has two RRMs and is quite similar to the SR subfamily in humans. Unlike one of the RRM in SRs, the RRM of the RS family lacks the SWQDLKD motif.

All eukaryotic SR genes were classified into subfamilies using the RRM domain structure. There are five major SR groups, which can be further divided into at least 11 sub-families (Richardson et al., 2011). Out of the 11 sub-families, five of these sub-families are extensively represented by photosynthetic eukaryotes (RS, RSZ, RS2Z, SCL and SR), six sub-families by metazoans 9G8/SRp20 (SRSF7), SRp38 (SRSF10), SRp40 (SRSF5), SRp55/75 (SRSF6/SRSF4), SF2 (SRSF1) and SRp54 (SRSF11), and a single sub-family SC35/SRSF2 shares members from both metazoans and plants. Different plant species show different rates of expansion within each subfamily. For example the SCL subfamily in Arabidopsis has 4 genes (*atSCL28*, *atSCL30*, *atSCL30a*, *atSCL33*), maize has 3 (*ZmSCL25*, *ZmSCL25a*, *ZmSCL30a*), and rice has 4 (*OsSCL30a*, *OsSCL25*, *OsSCL26*, *OsSCL30*, *OsSCL57*) (Richardson et al., 2011). Within an Arabidopsis SR subfamily there are closely related protein pairs (atSR34/SR1 and atSR34b, atRS31 and atRS31a, atRS40 and at RS41, atRSZ32 and atRSZ33, atSCL33 and atSCl30a, at RSZ22 and atRSZ22a). These observations raise the question of whether these have distinct functions or are redundant in their functions. Looking at the overall SR duplication in plants, the high number of SRs probably arose because plants are sessile and have to respond to many environmental changes. Since SRs are critical splicing factors for control of gene expression at various levels, they might be performing unique or overlapping functions.

The RRM domain in SR proteins can recognize and bind to various loosely conserved *cis*-regulatory elements in RNA sequences on the pre-mRNA (Shen and Green, 2004). In fact, in several cases, sequences identified as binding sites for one SR protein can also be recognized by other SR proteins, and the lack of stringent RNA binding specificity of SR proteins may partially account for their apparent redundancy in function. The carboxyl terminal RS domain is characterized as being disordered, nevertheless this domain participates in spliceosomal assembly by promoting either protein-protein or protein-RNA interactions. The RS domain undergoes substantial phosphorylation and dephosphorylation to modulate its interaction with other RNA or proteins and contains signals to target the polypeptide to the nucleus (Blencowe et al., 1998; Blencowe and Ouzounis, 1999; Manley and Krainer, 2010).

SR proteins are involved in pre-mRNA splicing by recruiting the splicing machinery to splice sites. There are several examples where SRs bind to the spliceosomal complex and recruit U1 snRNP to the 5'splice site, U2AF to the 3 'splice site, or the U2snRNP to branch point, and bind SREs (ESS, ESE, ISS or ISE) for both CS and AS; (Golovkin and Reddy, 1996; Reed, 1996; Golovkin and Reddy, 1999; Lam and Hertel, 2002; Reddy, 2007) (Figure 1.8). SR proteins are concentrated in nuclear speckles, and are recruited from these sites to sites of RNAPII (RNA polymerase II) transcription (Misteli et al., 1997; Ali et al., 2003; Ali and Reddy, 2006; Ali et al., 2008; Ali and Reddy, 2008a; Spector and Lamond, 2011). It has been reported that SC35 promotes RNAP II elongation in some genes, confirming the existence of coupling between transcription and splicing and its effect in maintenance of genome stability (Qian et al., 2011). Although in plants a direct link between AS with chromatin state or RNAPII has not yet been characterized, the spatial organization of the SRs and other splicing factors in the perichromatin regions of the nucleus have been studied in both plants and animals (Ali and Reddy, 2008a; Niedojadlo et al., 2012). This area needs to be explored further in plants as there could be a correlation between environmental stress and AS variants affected by chromatin modification and epigenetic responses.

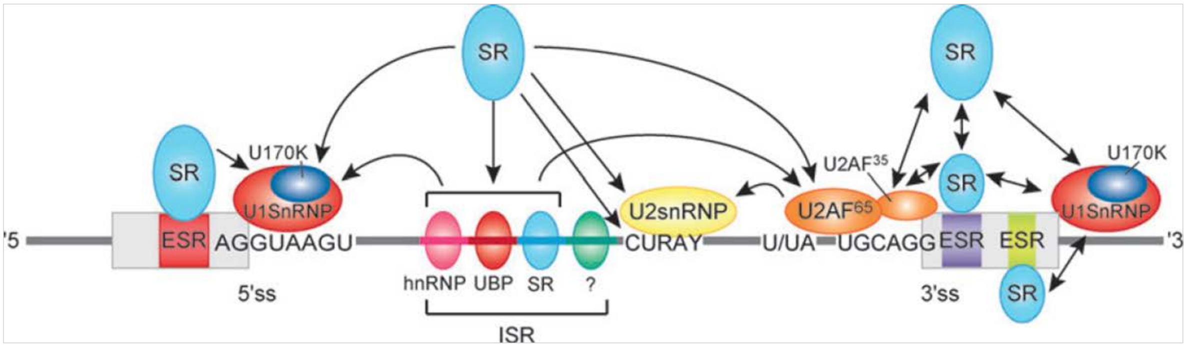


Figure 1.8: Model of spliceosome assembly with plant SR and other RNA binding proteins. Exons (white boxes) shown with exonic splicing regulators (ESRs); and introns shown as a horizontal line in between containing the 5’ and 3’ SS. Colored segments in the intron are intronic splicing regulators (ISRs) and the U/UA region in place of the animal polypyrimidine tract. RNA binding proteins such as SR proteins and other hnRNPs can interact with exonic/intronic sequence elements and promote recognition of 5’ and 3’ sites by recruiting U1 snRNP, U2AF, and other spliceosomal proteins. SR proteins also link the components at the 5’ and 3’ splice sites. Some interactions shown here have been experimentally validated (e.g., interaction among SR proteins, interactions of several SR proteins with U1-70K, U2AF, UBP binding to U-rich sequences, etc.), and certain others are putative interactions. Arrows indicate SR protein-mediated interactions. SR, serine/arginine-rich protein; U2AF65, U2 auxillary factor large subunit; U2AF35, U2 auxillary factor small subunit; hnRNP, heterogeneous nuclear ribonucleoprotein particle proteins [adapted from (Reddy, 2007)].

Although the existence of plant SR proteins has been known for sometime, biochemical analysis has been hampered by the lack of in-vitro splicing extracts (likely because of huge amount of pigments, cell wall components/lipids or vacuolar compounds in the cellular extracts). It has been known that SRs not only function as single gene products in regulating AS but pre-mRNAs from these genes themselves undergo substantial alternative splicing. In Arabidopsis there is a six-fold increase in the SR gene transcriptome (14 SR genes giving rise to 93 distinct AS isoforms) (Palusa et al., 2007a). In a study of AS of SRs in 20 plant species it was found that alternative 3′ splicing is the most common AS event type among SR genes (134 genes), followed by intron retention (111 genes), alternative 5′ splicing (109 genes), skipped exons (106 genes) and finally alternative 3′ and 5′events (http://combi.cs.colostate.edu/as/gmap_SR genes)

(Richardson et al., 2011). Furthermore, AS of most pre-mRNAs is altered by stresses and environmental signals. Extensive AS in SRs and its regulation by environmental signals is likely to add to multilayered control of gene-expression by SRs in their role as master regulators.

In Arabidopsis seedlings, 13 SR genes are alternatively spliced to generate 75 transcripts, of which 53 contain a premature termination codon (PTC) and these are changed in a mutant (*upf3)* in which NMD is impaired, suggesting a strong correlation between NMD and SR splicing (Palusa and Reddy, 2010). The PTC products are mainly generated from the long intron of plant-specific subfamilies RS, RS2Z and SCL. There is PTC in most of these AS variants and so they either undergo degradation via NMD, or if they escape the NMD process, encode extremely truncated proteins containing only a part of RRM that might function as activators/suppressors for gene-expression. Much of the functional analysis of these isoforms has been hindered due to lack of knockout mutants for the SR genes.

SR protein-protein interactions have been studied by immunoprecipitation, yeast two-hybrid and other in-vitro binding assays. The atU1-70K was found to interact with atSR34/SR1, atRS21, atRSZ22 and some plant specific SR proteins such as atSCL33 (Golovkin and Reddy, 1998, 1999). Some SR plant proteins not only interact with U170K but also with U2AF65 and U2AF35, stabilizing the spliceosomal complex at the 5' and 3' splice sites for efficient splicing (Ellis et al., 2008). There are several phenotypes associated with SRs and most of them until now have been studied in overexpressor lines. Overexpression of atSRp30, a member of Arabidopsis SF2/ASF subfamily, resulted in the morphological and developmental phenotype of late flowering (Lopato et al., 1999b). An increase in levels of atRSZ33 protein levels caused severe pleiotropic changes in plant development resulting from increased cell expansion and alterations in cellular polarity (Kalyna et al., 2003). As in animals, several SRs autoregulate AS of their own pre-mRNAs (Lopato et al., 1999b; Kalyna et al., 2003; Isshiki et al., 2006; Thomas et al., 2012).

SR45, an SR-like protein, is the most intensively studied SR like protein. The *sr45* mutant showed a late flowering phenotype by influencing the autonomous pathway and has altered leaves and root morphology. There were also changes in the AS pattern of other SR genes (*atRSp31*, *atRSp31a*, *atSRp34* & *atSRp34b*) in the *sr45* mutant (Ali et al., 2007). One of two alternatively spliced SR45 isoforms was found to complement exclusively the mutant's flower phenotype, while the other rescues only the root defect thus assigning distinct functions to each isoform (Zhang and Mount, 2009). Additionally, the *sr45-1* mutant displays defects in the maintenance of DNA methylation and shows epigenetic regulation of late flowering phenotype (Austin et al., 2012). In addition to developmental phenotypes, *sr45* shows hypersensitivity to abscisic acid (ABA) and sensitivity to 3% glucose (Carvalho et al., 2010) and this phenotype is complemented by either one of the splice variants. Stress or changes in environmental conditions are also major factors in alternative splicing pattern and according to a GO (Gene Ontology) analysis of alternatively spliced genes, the majority of genes associated with biotic or abiotic stress are represented in the AS category (Filichkin et al., 2010; Gan et al., 2011). The alternative splicing patterns of *SR34/SR1, SR34b, RS40, RS31* and *SR33* were altered by cold (4^0C), and under heat (37^0C) the isoforms of *RS30, SR1, SR34b, RS31a, RS40, RSZ32, RSZ33, SR33* and *SCL30a* were affected. Interestingly the two isoforms of *SCL33* showed opposite affects under high and low temperature, as these isoforms were increased under heat and reduced by cold (Palusa et al., 2007a).

In humans, mutations of SRs cause a number of diseases. The metastasis-associated lung adenocarcinoma transcript 1, MALAT1, is a long non-coding RNA (lncRNA) that has been discovered as a marker for lung cancer metastasis and SFRS1 is involved in MALAT1 processing at a transcriptional level via RNA polymerase II (Eissmann et al., 2012; Tripathi et al., 2012). Overexpression studies have also shown that SF2/ASF, SC35 and SRp20 are associated with malignant ovarian cancer (Fischer et al., 2004) and SF2/ASF upregulation also causes various forms of cancer. Interestingly, HIV-1 virus uses several human SRs to produce 40 different mRNAs from its pre-mRNA by using a combination of several alternative 5' and 3'ss (Stoltzfus and Madsen, 2006). SMA (spinal muscular atrophy) is a severe hereditary disorder that results in exon inclusion of the SMN2 gene by SF2/ASF (Wirth et al., 2006). SF2/ASF and Arp40 bind to an ISS and promote exclusion of exon 9 of CFTR (cystic fibrosis transmembrane conductance regulator) (Buratti et al., 2007). There are also high levels of SRp20 in bipolar patients (Watanuki et al., 2008). PfSR1, a novel AS factor in *Plasmodium falciparum* influences the AS activity of three endogenous genes, but most importantly the overexpression of this gene caused inhibition of parasitic proliferation in human RBC (Eshar et al., 2012).

SR and antagonistic partners regulate splicing events

SR proteins generally function as splicing enhancers, but are also known to function as negative regulators in some cases. On the other hand the hnRNP family are negative regulators of splicing. These two main families of splicing factors regulate splicing antagonistically. The exon6 cluster of a highly alternatively spliced gene *Dscam*, is regulated by the specific interaction between the hnRNP protein HRP36 and an SR protein for correct inclusion of a single exon (Olson et al., 2007). In vertebrates, a well-studied member of hnRNP is the polypyrimidine tract-binding protein (PTB) that targets CU rich regions of pre-mRNA and suppresses the inclusion of exons (Oberstrass et al., 2005). In Arabidopsis, there are three paralogues of PTB/hnRNP1 that undergo extensive auto/cross regulation of their expression

(Wachter et al., 2012). A total of 21 glycine rich RNA binding proteins are reported in Arabidopsis and these are homologous to human hnRNP/A or B, out of which only five (three UBA2, AtGRP7 and AtGRP8) of these have been studied. The best-studied plant hnRNP proteins to date are the Arabidopsis orthologs of the animal negative splicing regulator PTB, and the glycine-rich RNA-binding proteins, GRP7 and GRP8, components of a slave oscillator coupled to the circadian clock (Schoning and Staiger, 2009; Staiger and Koster, 2011; Wachter et al., 2012).

Identification of cis elements of SRs

Many of the approaches that have been described above under "identification of SREs" such as Systematic Evolution of Ligands by Exponential Enrichment (SELEX) , CLIP, HITS-CLIP and PAR-CLIP for RNA binding proteins are being used to identify physiological RNA targets of SR proteins. SELEX together with large-scale bioinformatics tools have identified potential sequence targets for animal SF2/ASF, and SC35, SRp40, 9G8 and SRp20 for mainly ESEs (Long and Caceres, 2009). An adaptation of SELEX is genomic SELEX that uses real genomic sequences rather than random pools to identify authentic RNA-protein interaction. Computational tools such as RESCUE (relative enhancer and silencer classification by unanimous enrichment) and PESE's (putative ESEs) are based on the mammalian system of splicing regulation that focus on exon skipping (Fairbrother et al., 2002). These approaches identified candidates for ESEs that occurred in exons with weaker splice sites compared to stronger ones. It has been documented that mutation of these PESE's resulted in almost 82% decreased efficiency in splicing, supporting the authenticity of these sequences (Zhang et al., 2005).

A modification of chromatin immunoprecipitation method (ChIP) has been developed to authenticate protein-binding to RNA sequences. The most recent technique that has paved the way to map specific sequences for RNA-protein interaction is called CLIP (cross-linking and immunoprecipitation) (Ule et al., 2003). The SF2/ASF binding sequence (UGRWG) has been discovered using CLIP (Sanford et al., 2009). A modification of the CLIP protocol called iCLIP (Konig et al., 2010), which allows high-resolution identification of RNA-protein crosslinked sites, was used to investigate the binding specificity and endogenous RNA targets of SRSF3 and SRSF4. SRSF2 and 3 bound mainly to intronless transcripts, implicating their role in histone modification (Anko et al., 2012). The validity of this technique was proven by the fact that a similar consensus sequence (CU-rich) of SRSF3 was found before using the SELEX method (Cavaloc et al., 1999).   Efforts to use more robust techniques, such as HITS-CLIP and PAR-CLIP for the identification of RNA targets and SREs for both animal and plant SRs are underway.

RATIONALE AND SIGNIFICANCE OF MY THESIS RESEARCH

As described above studies in animals show that SR proteins perform roles in CS and AS and many additional roles, which include export of mRNA to cytoplasm, mRNA stability, translation, genome maintenance and microRNA biogenesis (Huang and Steitz, 2005; Long and Caceres, 2009; Wu et al., 2010). SR genes in plants are considerably expanded with paralogs and plant-specific subfamilies.   Although plant SR proteins have been known for over a decade their biochemical analysis has been hampered due to the lack of plant-derived *in-vitro* splicing extracts. Interestingly, as discussed above pre-mRNAs of Arabidopsis SR genes undergo extensive AS giving rise to about 100 transcripts, thereby increasing the transcriptome complexity of SRs by about six fold (Palusa et al., 2007b). The increase in SR gene number and diversity in plants as compared to animals is attributed to the difference in pre-mRNA splicing between plants and animals. Despite the presence of a large number of SR proteins in plants their potential roles in plant growth, development and plant responses to environment, regulated splicing in plants is poorly understood. The reason I became interested in Arabidopsis SR proteins was to address the challenge of uncovering the functions of the SR proteins due to several genome duplication events giving rise to paralogous SR gene pairs. An interesting aspect to look at would be to know whether their functions are redundant or whether they evolved to perform new functions. Some of the fundamental unanswered questions about plant SR proteins are: i) what is the role of individual SR proteins in plant growth and development, ii) to what extent SR proteins play unique and redundant roles, iii) which individual gene's pre-mRNA splicing is regulated by a given SR, iv) what fraction of AS is controlled by a given SR or a combination of SRs and v) what are the global targets of a given SR or a combination of SRs. My thesis addressed some of these questions for three SR proteins (SC35, SCL33 and SCL30a). Two of these are members of the SCL family (SR33 and SCL30a) and are plant specific. The third one (SC35) was selected because its RRM domain is ~50% similar to the RRM in the SCL subfamily. SC35 is the only member in the SC subfamily in Arabidopsis and is considered an ortholog of animal SC35 (Barta et al., 2010). In the first chapter, I studied the role of these three SRs individually or in combination in plant growth and development using loss-of-function mutants. In chapter 2, I investigated the role of the same three SR proteins in splicing of SCL33 pre-mRNA using a novel splicing reporter and mutants that I generated and identified SCL33 binding sequences, and in the 3rd chapter I used NGS to interrogate global changes in gene expression and AS in a triple SR mutants. This revealed genome-wide direct and indirect targets affected by the loss of SR proteins, which will provide novel insight into understanding pre-mRNA splicing regulation by SR proteins. Chapters 2 to 4 are written in a manuscript format. Chapter 3 is in press in Plant Journal and the other two chapters will be submitted for publication soon.

Chapter 2

OPPOSING ROLES OF MEMBERS OF THE SERINE/ARGININE (SR)-RICH PROTEIN FAMILY IN REGULATING FLOWERING TIME

SUMMARY

The potential functional redundancy and/or synthetic phenotypes among single SR proteins (SCL33, SC35, SCL30a) was addressed by generation of three double mutants (*sc35 scl30a, sc35 scl33, scl33 scl30a*) representing all combinations and a triple mutant. All mutants are viable suggesting that loss of these SRs singly or in combination does not lead to lethality. However, complex and opposing flowering phenotypes were observed, in these mutants. Among the single mutants *sc35* and *scl30a* showed early flowering whereas *scl33* showed delayed flowering under both long days (LD) and short days (SD). In double mutant combinations, *sc35 scl30a* flowered early as in single mutants and no additive effect was observed whereas *scl33* was epistatic to *scl30a* in the double mutant and to *sc35* and *scl30a* in the triple mutant and these exhibited an even more pronounced late flowering phenotype as compared to *scl33*. The late flowering phenotype of *scl33, scl33 scl30a* and *scl33 sc35 scl30a* under both LD and SD, and rescue of this phenotype by vernalization, suggest that they regulate the autonomous flowering pathway. In the late flowering mutants expression of *Flowering Locus C* (*FLC*), a key negative regulator of flowering, and *FRIGIDA* (*FRI*), a positive regulator of *FLC* expression are upregulated. In contrast, early flowering mutants (*sc35*, *scl30a* and *sc35 scl30a*) showed increased expression of *FLOWERING LOCUS T* (*FT*), a positive regulator of flowering. These results indicate that members of the SR gene family perform opposing roles in regulating flowering time.

INTRODUCTION

Generation of functional mRNAs from multi-exon genes in eukaryotes is an essential step in gene expression and it involves excision of introns from the nuclear pre-mRNA and joining of exons (Sharp, 1994). This process occurs in the spliceosome, a large ribonucleoprotein machine consisting of five small ribonucleoprotein particles (snRNPs) and many non-snRNP proteins (Black, 2003; Bessonov et al., 2008; Wahl et al., 2009). Pre-mRNAs from over 60% of intron-containing genes in plants and about 95% in animals undergo alternative splicing to generate more than one transcript from a single gene (Pan et al., 2008; Filichkin et al., 2010; Lu et al., 2010; Marquez et al., 2012; Reddy et al., 2012b; Syed et al., 2012). Although the functions of most of the splice variants in plants are unknown at this time, it is clear from global transcriptome studies that alternative splicing is highly pervasive in all multicellular organisms (Reddy, 2007; Reddy et al., 2012b). Multiple roles for alternative splicing have been proposed, which include enhancing the coding capacity of a genome and potential regulation of functional transcript levels by producing splice variants that are targets of nonsense-mediated decay (Reddy, 2007; Kalsotra and Cooper, 2011; Syed et al., 2012). In animals, regulated splicing has been shown to affect diverse biological processes associated with development and disease (Kalsotra and Cooper, 2011).

Serine/arginine-rich (SR) proteins are a family of non-snRNP proteins in the spliceosome that play multiple roles in constitutive splicing as well as in the regulation of alternative splicing (Long and Caceres, 2009; Reddy and Ali, 2011; Busch and Hertel, 2012). In animals, SR proteins function in various other processes including mRNA export, mRNA stability, translation and genome stability (Lemaire et al., 2002; Sanford et al., 2004; Manley and Krainer, 2010). The presence of one or two RNA binding domains in the N-terminus of SRs allows them to interact with specific sequence elements in pre-mRNA, and a serine/arginine rich domain at the C-terminus enables interaction with other snRNP and non-snRNP proteins, and these interactions promote or suppress pre-mRNA splicing (Reddy, 2007; Long and Caceres, 2009; Reddy and Ali, 2011). In plants, the SR protein family is considerably expanded as compared to animals with some plant-specific members (Richardson et al., 2011). Based on sequence similarity and the type and organization of various domains, plant SR proteins are grouped into six subfamilies (SR, RSZ, SC35, SCL, RS2Z and RS) of which the last three subfamilies are plant-specific (Barta et al., 2010). In Arabidopsis, five of the six subfamilies contain two or more SR proteins whereas the SC35 subfamily has a single member. Interestingly, pre-mRNAs from a majority of Arabidopsis *SR* genes (14 out of 18) and other plants' SR genes undergo extensive alternative splicing and increase the number of transcript isoforms by about five-fold (Isshiki et al., 2006; Palusa et al., 2007a). Many *SR* splice variants contain a premature termination codon and are likely targets of NMD (Palusa and Reddy, 2010). Furthermore, environmental signals have been shown to modulate alternative splicing of SR pre-mRNAs (Palusa et al., 2007a).

Despite the critical roles of SR proteins in pre-mRNA splicing and gene regulation, little is known about their roles in plant growth and development. The increased number of SR genes in flowering plants as compared to animals (18 in Arabidopsis, 21 in maize and 11 in humans), and the presence of plant- and animal-specific SR proteins as well as differences in the prevalence of different types of alternative splicing between plants and animals (Reddy, 2007; Syed et al., 2012), raise many questions pertinent to the roles of plant SR proteins in splicing regulation and growth and development. Although animal SR proteins are functionally redundant in splicing complementation assays performed *in vitro*, genetic studies have revealed that some SRs are functionally redundant whereas others are essential (Sanford et al., 2003). For instance, knocking out *SRSF1 (ASF/SF2)* in a chicken cell line and *C. elegans* caused cell lethality and late embryonic lethality, respectively. Similarly, SRSF2 (SC35) and SRSF3 (SRp20) were shown to be essential for embryonic development in mouse (Sanford et al., 2003). Thus far, no loss-of-function mutants of plant *SR* genes have been characterized. However, overexpression of *AtSRp30* or *AtRSZ33* in wild type background showed multiple developmental and morphological changes and altered alternative splicing pattern of several *SRs* and other splicing related genes, which is consistent with the role of SRs as regulators of splicing (Lopato et al., 1999b; Kalyna et al., 2003). It is known in animals that different cellular amounts of an SR protein have different effects on constitutive and alternative splicing (Long and Caceres, 2009; Busch and Hertel, 2012). Hence, the phenotypes observed with overexpression of an SR protein may not be the same as loss-of-function mutants. In addition to these overexpression studies with *SRs*, a knockout mutant of an *SR-like* gene, *sr45*, has been characterized. The *sr45* mutant showed pleiotropic phenotypes including reduced root length, late flowering and hypersensitivity to glucose (Ali et al., 2007l ; Carvalho et al., 2010). Here, we have performed a systematic genetic analysis using gene knockouts to address the functions of three *SR* genes, *SC35*, *SCL33* and *SCL30a*, whose functions in plant development have not been investigated. One of these (*SC35*) is an ortholog of animal *SC35 (SRSF2),* whereas the other two are specific to plants. Since the SR family is expanded in plants and several SRs are represented by paralogs, generation of higher order mutants (double and triple) is needed to address functional redundancy among different SRs, especially among paralogs. To address potential functional redundancy/synthetic phenotypes among these three SRs and to evaluate the phenotypic effects of loss of any two combinations or all three of these genes we generated double and triple mutants. Phenotypic analysis of all seven (three single, three double and one triple) mutants revealed that loss of these genes, either singly or in combination, does not lead to lethality but they have opposing roles in regulating flowering time. Detailed characterization of flowering phenotypes under different conditions revealed mutants causing late flowering phenotypes affected the autonomous pathway by upregulating a key flowering repressor, and early flowering mutants showed increased expression of a flowering promoter in the photoperiod pathway. This analysis represents the first phenotypic characterization of plant *SR* loss-of-function mutants.

RESULTS

Loss of SC35, SCL33 and SCL30a individually or in combination of two or three is not lethal to plants

In plants, the roles of SR proteins in plant development are not well understood. In fact, thus far no loss-of-function mutants of plant SR proteins have been characterized. Here we have generated knockout mutants of three *SR* genes, *sc35, scl33* and *scl30a* in Arabidopsis that belong to two different families of plant SR proteins. SC35 is the sole member of the SC35 family and is a homolog of human SC35. SCL33 and SCL30a are members of the SCL family, which consists of four closely related members. Furthermore, members of the SCL subfamily are closely related to the SC35 subfamily. Genotyping by genomic PCR and RT-PCR confirmed the presence of T-DNA in these genes and the absence of transcripts corresponding to these genes (Thomas et al., 2012). To investigate potential redundancy and/or synthetic phenotypes we generated three double mutants (*sc35 scl30a, sc35 scl33, scl33 scl30a*) that represent all combination of these three genes by crossing single mutants. In addition, a triple mutant (*scl33 sc35 scl30a*) was generated. The genotypes of all double and triple mutants were also confirmed by genomic PCR and gene expression analysis by RT-PCR (Thomas et al., 2012). Mutant seedlings grown on MS plates did not show any significant differences in seedling size and root growth (Figure 2.1a). All homozygous mutants were grown in the greenhouse and phenotypic analysis of silique size, seed number and embryo development of these mutants did not show significant differences to wild-type (WT) (Figure 2.1b). These results indicate that loss of function of these three *SR* genes either singly or in any combination of two or three is not lethal to the plant.

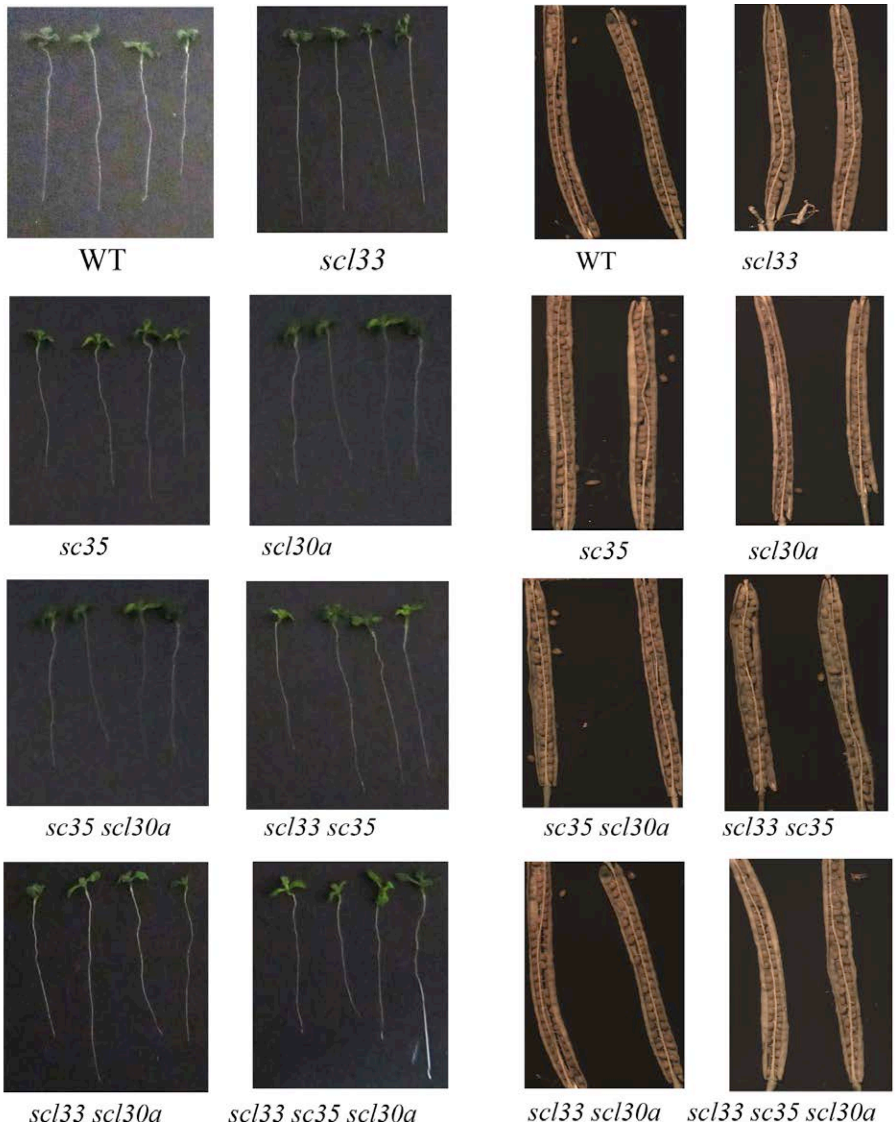


Figure 2.1. Root growth (a) and silique development (b) in wild type and all *sr* mutants.

Flowering time is altered in the SR mutants

While growing these plants in the greenhouse we observed that several of these mutants have altered (early or late) flowering phenotypes. Arabidopsis is a facultative long day (LD) species that flowers later in short days (SD) than in LD. Extensive genetic and molecular studies on flowering have defined at least four major flowering pathways that converge on a few flowering promoters and regulate transition from vegetative to reproductive phase (Boss et al., 2004; Amasino, 2010; Srikanth and Schmid, 2011). These include the photoperiodic, the vernalization, the autonomous and hormonal pathways (Figure 2.2) (Boss et al., 2004; Amasino, 2010; Srikanth and Schmid, 2011).

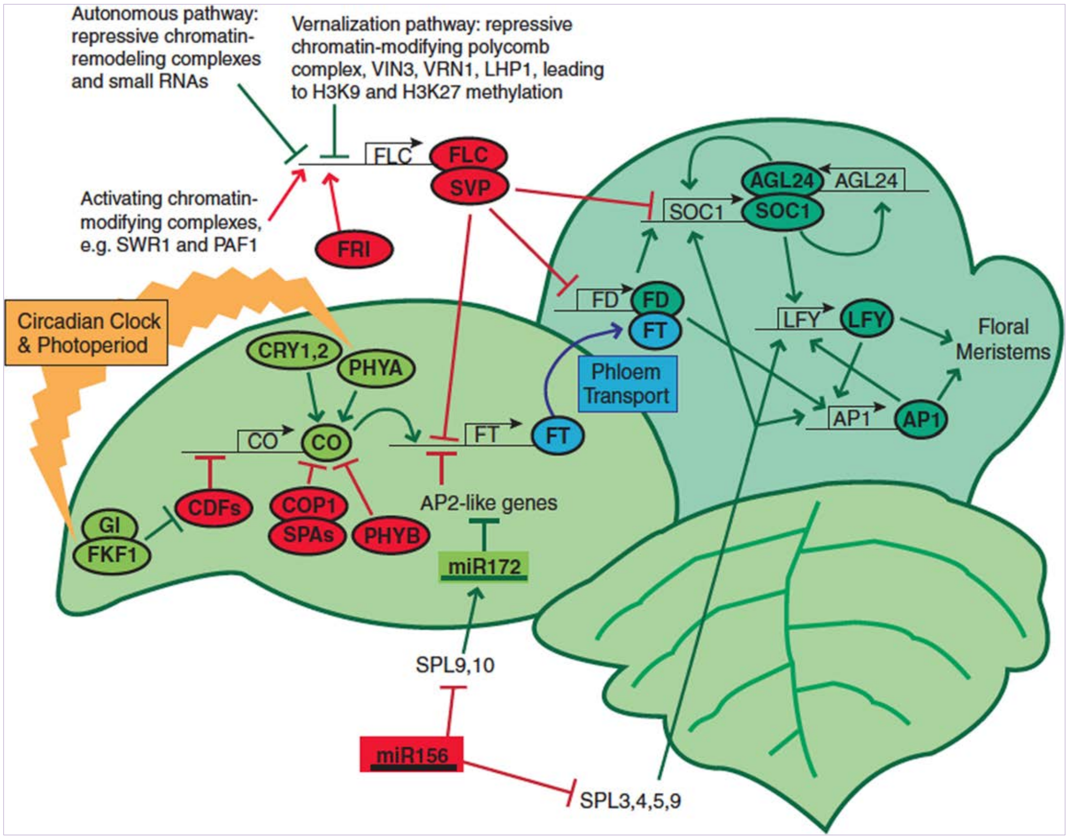

Autonomous pathway: repressive chromatin-remodeling complexes and small RNAs
Vernalization pathway: repressive chromatin-modifying polycomb complex, VIN3, VRN1, LHP1, leading to H3K9 and H3K27 methylation
Activating chromatin-modifying complexes, e.g. SWR1 and PAF1
FLC
FLC
SVP
FRI
AGL24
AGL24
SOC1
SOC1
Circadian Clock & Photoperiod
CRY1,2
PHYA
FD
FD
FT
Phloem Transport
LFY
LFY
Floral Meristems
CO
CO
FT
FT
AP1
AP1
CDFs
COP1
SPAs
PHYB
GI
FKF1
AP2-like genes
miR172
SPL9,10
miR156
SPL3,4,5,9

Figure 2.2. Flowering time regulation in Arabidopsis. Pathways of flowering time showing interaction of genes, proteins (ovals) and microRNAs. Solid green or red lines with an arrow represent promotion, and those with a perpendicular bar represent repression. Components that promote flowering are shown in green, and those that repress flowering are shown in red (Amasino, 2010).

Flowering time under different conditions is quantified by counting the rosette leaves at the time of appearance of the first flower, or the number of days a plant takes to produce the first flower (Koornneef et al., 1991; Simpson and Dean, 2002). By growing mutants with altered flowering time under different photoperiods with or without vernalization one can place a mutant in a flowering pathway. Here, we preformed a detailed characterization of the flowering phenotype of all seven mutants under long day (LD, 16 h:8 h light:dark), short day (SD, 8 h:16 h light:dark), vernalized long day (VLD) and vernalized short day (VSD) conditions to gain insights into the roles of these SRs in regulating flowering time. All mutants were grown simultaneously under identical conditions. In all cases, flowering time was quantified by counting i) the number of days a plant takes to produce the first flower (Boyes et al., 2001) and ii) the number of rosette leaves at the time of appearance of first flower.

SR mutants showed both late and early flowering under long day

Under long day (LD) conditions, wild-type plants flowered at an average of 30.42±0.98 days and the rosette leaf count (RLC) average at the time of flowering was 13.2±0.24 leaves (Table 2.1). Of the three single mutants, *scl33* showed delayed flowering with an average flowering time of 34.71±1.06 days, while the other two single mutants, *sc35* and *scl30a*, were early flowering compared to the wild-type with an average flowering time of 26.7±0.46 and 26.6±0.58 days, respectively (Figure 2.3a, 2.3b, right and Table 2.1). Of the three double mutants, only *sc35 scl30a* flowered earlier than WT with an average flowering time of 26.8±0.6 days, while *sc35 scl33* flowered at 31.5±0.3 days resembling WT. However, the double mutant *scl33 scl30a* was significantly late flowering (43.07±0.5 days), as was the triple mutant *scl33 sc35 scl30a*, which took 44.8±0.2 days to flower (Figure 2.3b, right and Table 2.1).

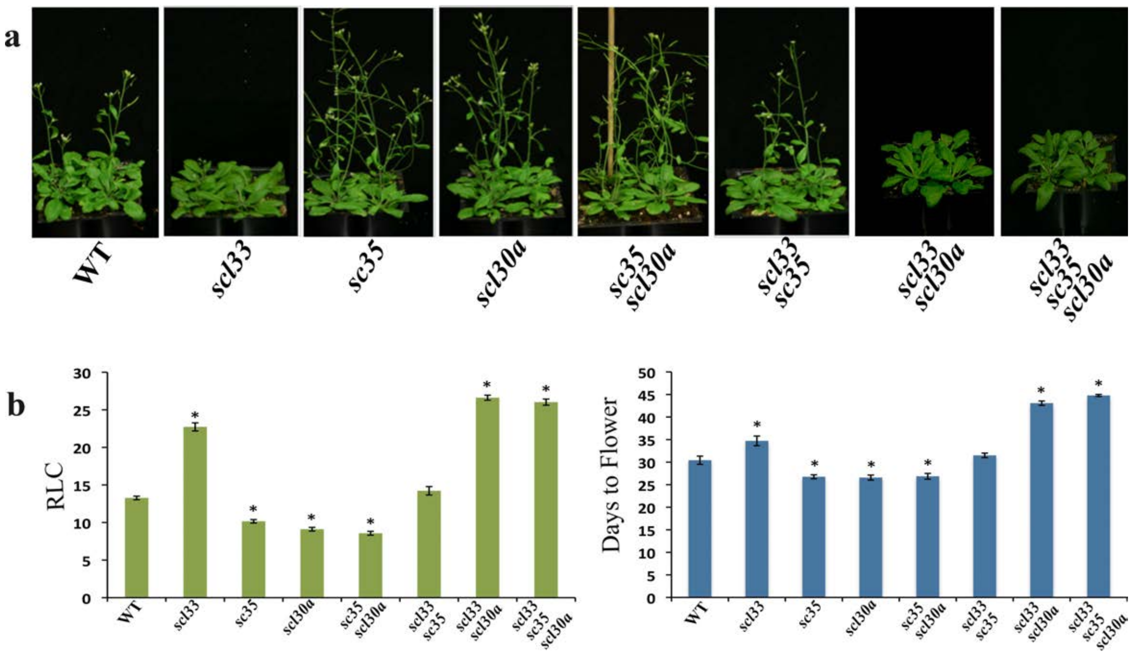


Figure 2.3. SR mutants show either early or late flowering phenotype under long day (LD). a) Pictures of 4-weeks old wild type, single (*scl33*, *sc35*, *scl30a*), double (*sc35 scl30a*, *scl33 sc35*, *scl33 scl30a*), and triple (*sc35 scl30a scl33*) mutant plants. b) Quantification of flowering time in wild type and *sr* mutants under LD (16 hr light: 8 hr dark) Left: Flowering time as measured by the number of days to produce the first flower. Right: Quantification of flowering time based on the number of rosette leaves at the time the first flower appeared. The error bars represent standard error of the mean. Asterisks indicate significant differences (p<0.05) between wild type and *sr* mutants. RLC, Rosette Leaf Count.

Table 2.1. Flowering time in *sr* mutants under different conditions. Flowering time is given in RLC±SE and days to flower in parenthesis. The flowering phenotype and affected flowering pathway are denoted in the last two columns.

Genotype	LD	LDV	SD	SDV	Flowering phenotype
WT	13.26 ± 0.25 (30.42)	9.41 ± 0.21 (26.79)	53.76 ± 0.46 (105.35)	46.76 ± 0.46 (62.38)	
scl33	22.16 ± 0.69 (34.71)	8.78 ± 0.15 (23.92)	60.47 ± 1.05 (114.41)	48.84 ± 1.05 (67.23)	Late
sc35	10.19 ± 0.38 (26.75)	8.80 ± 0.19 (25.19)	50.11 ± 0.74 (102.11)	42.92 ± 0.74 (58.15)	Early
scl30a	9.55 ± 0.36 (26.57)	9.47 ± 0.24 (26.38)	50.23 ± 1.06 (101.47)	41.30 ± 1.06 (61.38)	Early
sc35 scl30a	8.57 ± 0.25 (26.85)	9.20 ± 0.15 (25.41)	51.11 ± 1.26 (102.11)	40.23 ± 1.26 (61.61)	Early
sc35 scl33	14.22 ± 0.56 (31.5)	9.59 ± 0.65 (26.46)	63.52 ± 1.48 (116.88)	52.76 ± 1.48 (74.46)	Late under SD
scl33 scl30a	26.61 ± 0.35 (43.07)	10.31 ± 0.38 (27.42)	81.64 ± 1.17 (137.52)	53.61 ± 1.17 (76.34)	Late
scl33 sc35 scl30a	26.02 ± 0.60 (44.78)	9.76 ± 0.31 (27.23)	77.58 ± 0.72 (126.17)	54.30 ± 0.72 (78.69)	Late

The RLC of mutants compared to WT (Figure 2.3b, left) also supported the early flowering phenotype of *sc35*, *scl30a* and *sc35 scl30a*, and the late flowering phenotype of *scl33*, *scl33 scl30a*, *scl33 sc35 scl30a* (Table 2.1, Figure 2.4). The RLC (14.22±0.56) of *scl35 scl33* and the number of days it took to flower, is similar to wild type (Table 2.1).

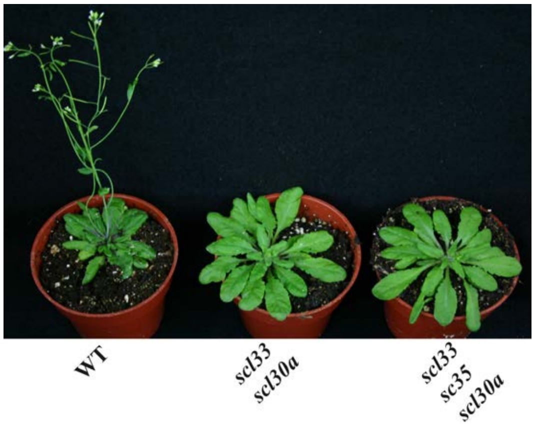


Figure 2.4. Wild type and two late flowering mutants (*scl33 scl30a*, *scl33 sc35 scl30a*) grown under LD. The WT plant that has already flowered has 14 rosette leaves whereas the mutants, which are yet to flower have higher number of rosette leaves (~40 - 44) after 35 days of planting.

SR mutants that are late flowering under LD showed late flowering in SD also

Mutants that show late flowering under LD but not in SD fall in the photoperiodic pathway, whereas autonomous pathway mutants flower late in both conditions (Koornneef et al., 1991; Boss et al., 2004; Amasino, 2010). To test if the observed late and early flowering phenotypes of *SR* mutants also occur in short day and to determine if any of the late flowering mutants fall under the photoperiodic pathway, all mutants were grown under SD and flowering time was quantified as above. Under SD conditions WT flowered in 105.3±0.4 days and all the late flowering mutants still flowered late as in LD condition (Figure 2.5a & 2.5b, right, Table 2.1). The difference in flowering time of the mutants to WT are: *scl33* (9 days), *scl33 scl30a* (32 days), *scl33 sc35 scl30a* (21 days), while the *scl33 sc35* mutant, which was close to WT under LD, flowered 12 days later under SD (Table 2.1). The two single mutants *sc35* and *scl30a* as well as the double mutant *scl35 scl30a* remained slightly early flowering compared to the wild-type flowering in the 101-102 day range, although these numbers were not statistically significant.

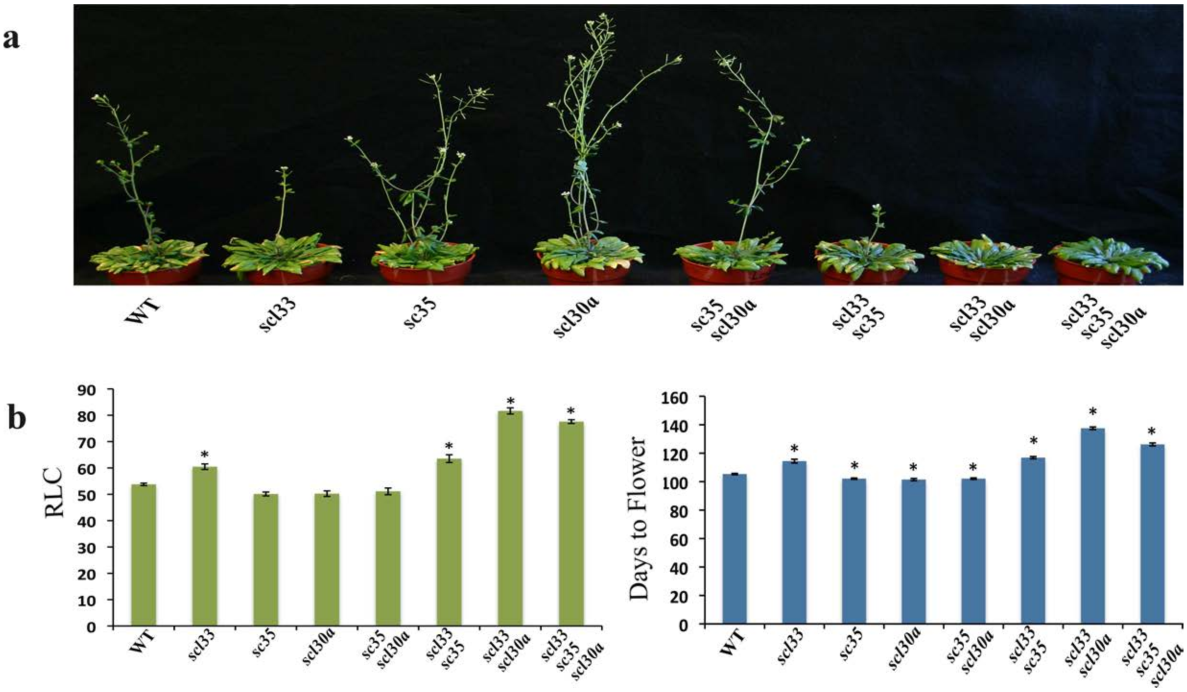


Figure 2.5. SR mutants showed early or late flowering phenotype under short day (SD). a) Eight-week old wild type and mutant plants. b) Quantification of flowering time in wild type and *sr* mutants. Left: Rosette leaf number at the time of appearance of the first flower. Right: Flowering time measured as the number of days to produce the first flower. The error bars represent standard error of the mean and significant differences (*p*<0.05) between wild type and mutant are denoted by asterisks.

The RLC of the late flowering mutants was also high, with the difference to WT (given in parenthesis) for the mutants: *scl33* (7), *scl33 sc35* (10) *scl33 scl30a* (28), *scl33 sc35 scl30a* (24) (Figure 2.5b, left, Table 2.1). This analysis of flowering in SD confirms that the late flowering mutants are not in the photoperiodic pathway as they are late flowering in both LD and SD conditions. These results place the late flowering mutants (*scl33*, *scl33 scl30a*, *scl33 sc35 scl30a*) in the autonomous pathway (Koornneef et al., 1991). However, the double mutant *scl33 sc35* that flowered like WT under LD showed late flowering under SD. Mutants that showed early flowering phenotype under LD showed early flowering phenotype under SD also. However, it is not as pronounced as in LD. Therefore, these mutants, like *early flowering* 1 (*elf1)* and *elf2*, are considered early flowering but photoperiod sensitive (Zagotta et al., 1992).

Late flowering phenotype of mutants under LD and SD is rescued by vernalization

Late flowering mutants in the autonomous pathway are rescued by vernalization. To test if the late flowering phenotype of *SR* mutants is rescued by vernalization, we vernalized seeds of *SR* mutants, quantified flowering time and compared it to their flowering time for non-vernalized seeds. Under vernalized LD conditions, a decrease was observed in flowering time for WT plants (Figure 2.6a & 2.6b right), with an average of 26.8±0.3 days.

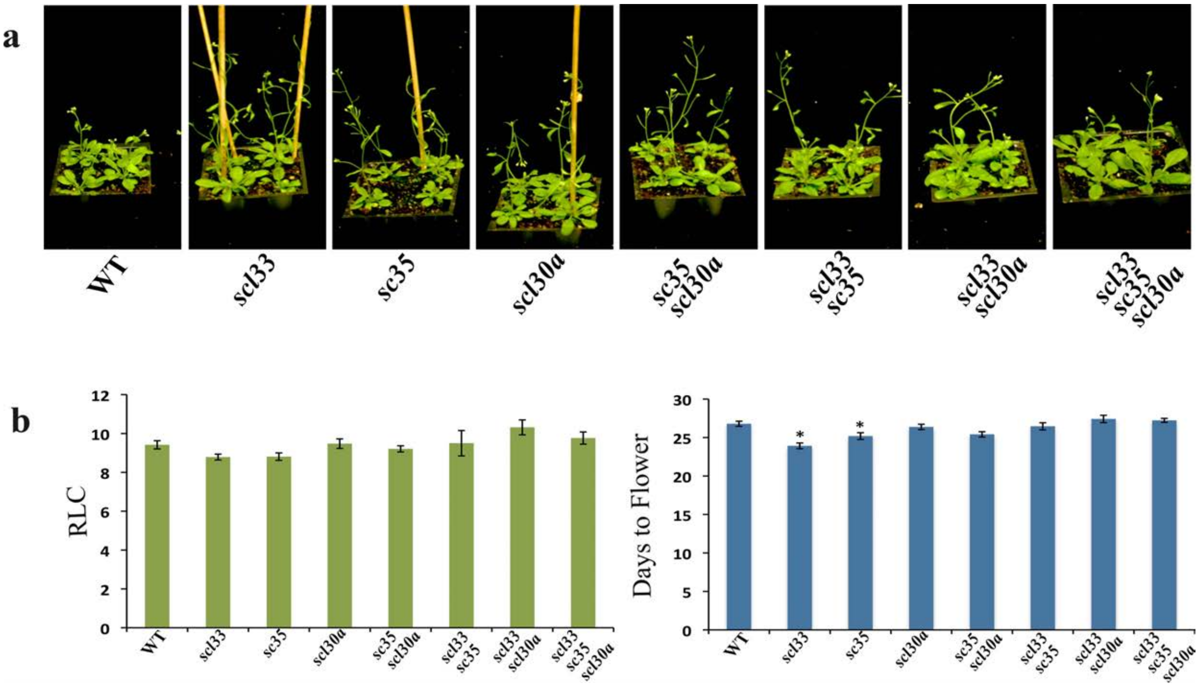


Figure 2.6. Vernalization (49 days) rescued late flowering phenotype of *sr* mutants under long day. a) Pictures of 4-weeks old wild type, and all mutant plants grown from vernalized seeds.
b) Quantification of flowering time in wild type and *sr* mutants under long day conditions
Left: Rosette leaf number at the time of appearance of first flower. Right: quantification is based on the number of days to produce the first flower. The error bars represent standard error of the mean. Asterisks indicate significant differences (p<0.05) between wild type and *sr* mutants. RLC, Rosette leaves count.

In general, vernalization affected all late flowering mutants, restoring the flowering time similar to WT. The late flowering mutant *scl33* that flowered 4 days later without vernalization, flowered 2 days earlier than WT after vernalization treatment (Table 2.1). Likewise the other late flowering mutants *scl33 scl30a* and *scl33 sc35 scl30a* that showed a difference of about 13 and 14 days to WT, showed less than a day difference to WT after vernalization (Table 2.1). The early flowering mutant genotypes also showed a similar flowering time as WT after vernalization. The RLC of mutant genotypes after vernalization showed no significant difference between late flowering *SR* mutants and WT, supporting the rescue of the late flowering phenotype in mutants by vernalization (Figure 2.6b, left, Table 2.1).

Vernalization dramatically reduced the flowering time (about 47 to 61 days depending on the mutant) of all late flowering mutants grown under SD (Table 2.1). However, the flowering time in these mutants is not rescued to wild type level (Table 2.1, Figure 2.7). There is also a decrease in the RLC compared to WT under SD after vernalization (shown as difference in number of RLC to WT): *scl33* (2), *scl33 sc35* (6), *scl33 scl30a* (7), *scl33 sc35 scl30a* (8) (Table 2.1, Figure 2.7). The exception is *scl33 sc35*, which shows a similar response under SD and SD after vernalization with the difference in days to WT being similar (12 days) although the RLC drops from 11 to 6. This rescue by vernalization can probably be improved by vernalization treatment of more than 49 d as shown for other mutants (Martinez-Zapater and Somerville, 1990).

***Flowering locus C* (*FLC*) expression and FLC splice variants are up-regulated in late flowering mutants.**

FLC, which encodes a MADS-box containing transcription factor, is a potent repressor of flowering (Michaels and Amasino, 1999, 2001). The autonomous and vernalization pathways converge through *FLC*, a repressor of the floral pathway integrators (Searle et al., 2006). FLC inhibits flowering by repressing the key flowering time integrators: FLOWERING LOCUS T (*FT), SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 (SOC1)* and a bZIP transcription factor, *FD* (Amasino, 2010). Several genes involved in both autonomous and vernalization pathways are known to regulate *FLC* expression (Michaels et al., 2005; Amasino, 2010).

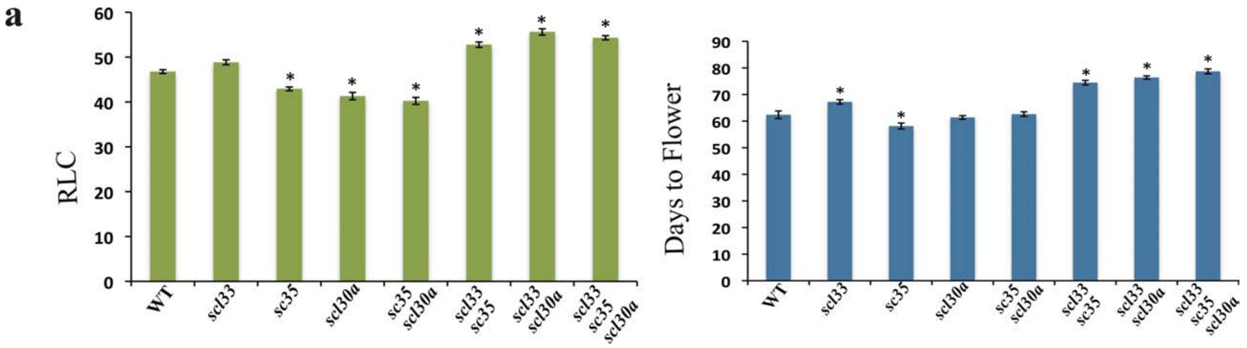


Figure 2.7. The late flowering phenotype of mutants in SD was also rescued by vernalization. Quantification of flowering time in wild type and *sr* mutants grown from vernalized seeds under SD. Quantification of days to rosette (left) and rosette leaves (left) was done as described Figure 2.3. The error bars represent standard error of the mean and significant differences between wild-type and mutant are denoted by asterisks.

To determine the role of FLC in late flowering mutants, we quantified the expression of FLC transcripts by quantitative reverse transcriptase PCR (qRT-PCR) using a primer set that amplifies all FLC transcripts. RNA from 12 and 29 day-old seedlings was used for this analysis. The expression levels of FLC in mutants relative to WT are shown in (Figure 2.8). High level of *FLC* transcripts was observed in two late flowering mutants (*scl33 scl30a* and *scl33 sc35 scl30a*), whereas two of the early flowering single mutants (*sc35, scl30a*) showed reduced expression of FLC (Figure 2.8). These results suggest that the observed flowering phenotypes of these mutants is due to altered expression of *FLC*. However, no significant difference in expression of *FLC* between the late flowering mutant *scl33* and wild type was observed.

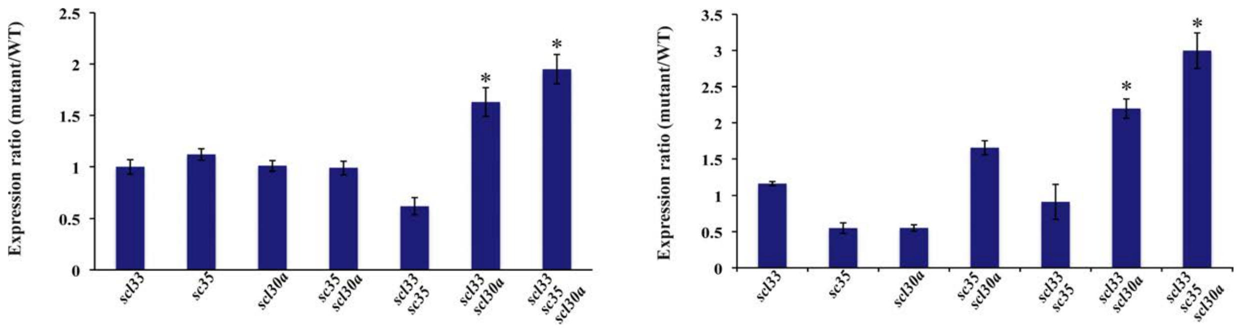


Figure 2.8. Changes in expression of *FLC* in *sr* mutants at different stages of plant development. qRT-PCR analysis using *FLC* specific primers that amplify all isoforms was done using leaf tissue at different times of plant growth: i) 12 days after sowing, ii) 29 days after sowing. For each time point three biological replications were used for qRT-PCR as described in Methods. The expression ratio above 1 indicates increased expression in the mutant and below one indicates decreased expression as compared to wt type. The error bars represent standard error of the mean. Asterisks indicate significant differences (*p*<0.05) between wild type and *sr* mutants.

The FLC pre-mRNA undergoes AS and produces four different splice variants according to TAIR annotation (http://www.arabidopsis.org/servlets/TairObject?id=136002&type=locus). Furthermore, variations in flowering time in *Brassica* and *Capsella* natural populations were attributed to an altered splicing pattern of *FLC* (Slotte et al., 2009; Yuan et al., 2009). Since SRs are key regulators of alternative splicing, we analyzed the levels of three of the four predicted splice variants in wild type and mutants using isoform-specific primers (Figure 2.9a) in 29-day old seedlings. All four *FLC* isoforms encode proteins that contain MADS box and KH domain but differ in the length of the C-terminal extension (Figure 2.9a) and none of them are candidates of NMD. Differential expression of the *FLC* isoforms was observed in *SR* late flowering mutants. Isoform 1 showed increase in *scl33 scl30a* and *scl33 sc35 scl30a* mutants with slight increase in *scl33* mutant. Isoform 2 and 3 increased in late flowering mutants *scl33 scl30a* and *scl33 sc35 scl30a*, whereas only isoform 2 was more in *scl33* (Figure 2.9b). The increased level of total transcripts and individual isoforms of *FLC* in the late flowering mutants support its role in late flowering in these mutants.

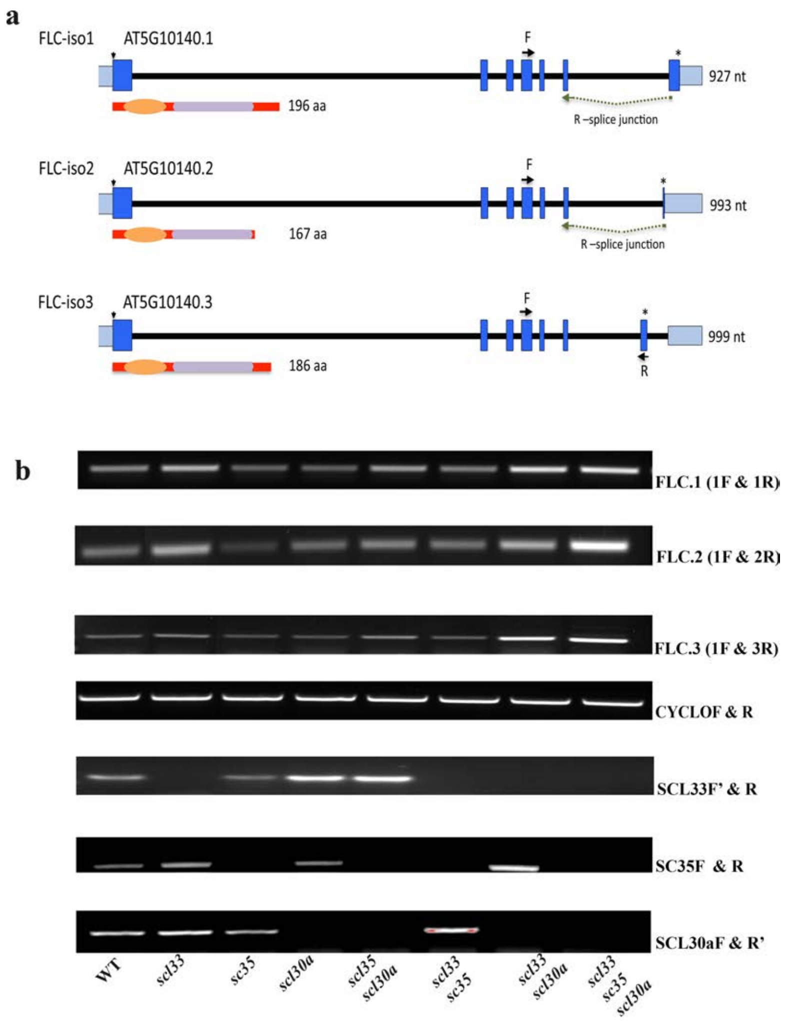


Figure 2.9. Expression of *FLC* isoforms in wild type and *sr* mutants in leaves from 29 day-old plants. a) Schematic diagram showing the gene structure and known splice variants of *FLC*. Introns are indicated by black lines, exons shown in blue boxes and UTRs (5' and 3') in light-blue boxes, predicted proteins are shown below. DNA binding domains are indicated in orange and K-box domain in light purple. The size of the transcript (nt) and the predicted protein (aa) of each isoform are shown at the right, the position of primers are indicated as forward (F) primer (black arrow) and reverse (R) or R-splice junction primers (in purple), the R-splice junction primers are indicated by dashed lines for isoforms 1 and 2. The position of start and stop codons are indicated by an arrowhead and an asterisk, respectively. b) Analysis of FLC splice variants (top three panels) in wild type and *sr* mutants by RT-PCR using isoform specific primers shown in Figure 2.9A and Material & Methods. The fourth panel shows cyclophilin control. The expression of *SCL33*, *SC35* and *SCL30a* in wild type and all *sr* mutants using gene-specific primers is shown in the bottom three panels.

FRIGIDA (*FRI*) is one of the key positive regulators of *FLC* expression (Amasino, 2010). *FRI* induces *FLC* expression and delays flowering (Michaels and Amasino, 1999, 2001; Choi et al., 2009). To determine if the increased expression of *FLC* in some of the mutants is due to increased expression of *FRI*, we performed qRT-PCR with the same RNA samples used to quantify *FLC* expression. Mutants that showed an increase in *FLC* expression also showed a higher level of *FRI* transcript (Figure 2.10), suggesting that increased expression of FRI has induced *FLC* expression.

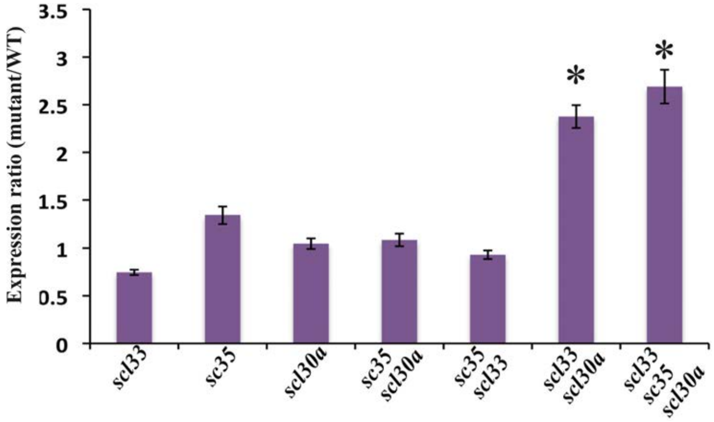


Figure 2.10. Quantitative analysis of expression of *FRI* in leaves from 29 day-old plants grown in LD. The sequences of primers used in qRT-PCR are provided in Material & Methods.

FLOWERING LOCUS T (*FT*) is up-regulated in early flowering mutants

FT is a key positive regulator of flowering, and three flowering time pathways (autonomous, vernalization and photoperiod) converge to regulate *FT* expression. We analyzed expression of *FT* by qRT-PCR in all mutants using the same RNA that was used to analyze *FLC* and *FRI* expression above. Two single mutants (*sc35*, *sc30a*) and the double mutant (*sc35 scl30a*) that are early flowering showed the highest increase in expression of *FT* (Figure 2.11). Thus, the expression of *FT* correlates with early flowering phenotype.

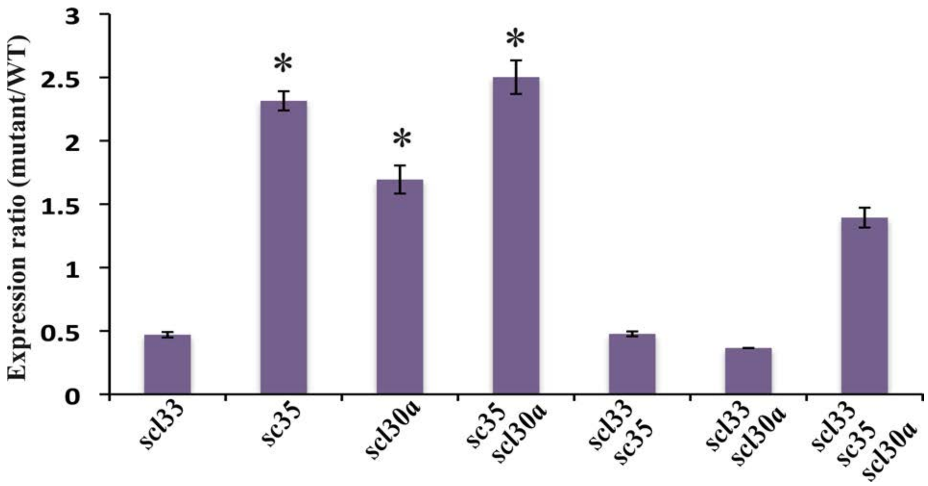


Figure 2.11. Levels of *FT* in wild type and *sr* mutants. qRT-PCR was performed with RNA extracted from leaves of 29 day-old plants grown in LD. Primer sequences are presented in Material & Methods.

DISCUSSION

To study the functions of SR proteins, we conducted a systematic molecular genetic analysis using knockout mutants of three SR proteins (SC35, SCL30a and SCL33) individually or in combination with one or two other SRs, and analyzed the phenotypic changes in plant growth and development. As SR proteins are key regulators of splicing, loss of one or more SRs could be lethal to plants. In mouse, deletion of *SC35* caused embryonic lethality whereas in *C. elegans* depletion of expression of SC35 using RNAi was not lethal (Longman et al., 2000; Sanford et al., 2003; Ding et al., 2004). Remarkably, plants lacking three SR genes either singly or in combination of two or three did not cause lethality or sterility problems. Analysis of single, double and triple mutants of these *SR* genes suggest that they are either not essential or there is some functional overlap among these and other SCL subfamily members or other SRs. Studies with higher-order mutants in which *SC35* and all members of the SCL subfamily (four) are mutated should address this possibility.

The only significant phenotype that we observed in these mutants is that they are either early or later flowering. The *sc35* and *scl30a* single mutants and the *sc35 scl30a* double mutant are early flowering in both LD and SD conditions with a photoperiod sensitive phenotype as has been observed with *elf1* and *elf2* mutants (Zagotta et al., 1992). The double mutant *sc35 scl30a* showed no additive effect, suggesting that they both affect the same pathway. All three early flowering mutants have increased expression of the flowering promoter *FT*. Furthermore, the expression level of *FLC* is down-regulated in the early flowering mutants except for the *sc35 scl30a* double mutant. Although the *sc35 scl30a* mutant did not show reduced *FLC* levels, expression of FT is high, suggesting that some other mechanism regulates *FT* expression in this double mutant. The *SCL33* belonging to the same subfamily as the *SCL30a* gene, surprisingly displays a late flowering mutant phenotype. Since *scl33* is late flowering under both photoperiods we place it in the autonomous pathway.

However, in combinations with other mutants a range of phenotypes were obtained, highlighting complex interactions of SRs in regulating flowering. Under LD conditions, *scl33 sc35* shows wild type phenotype even though RLC is more (14.22) suggesting a weak late flowering phenotype. However, under SD conditions *scl33 sc35* showed significantly increased flowering time (Table 2.1). Under LD the antagonistic effect of *sc35* on *scl33* is more pronounced than under SD signifying an epistatic effect of *sc35* on *scl33* under LD. Suppression of *AtGRP7,* an RNA binding protein whose expression is regulated by circadian clock, also leads to a strong late flowering phenotype in SD but not LD (Streitner et al., 2008). Pronounced delay in late flowering under SD but not in LD was also reported in mutants of GA biosynthesis and signaling mutants (Wilson et al., 1992; Peng et al., 1997). A similar flowering phenotype was also observed in a gain-of-function mutation in indole-3-acetic acid 7 (IAA7)/auxin resistant 2

(AXR2), a component in auxin signaling (Mai et al., 2011). It is therefore possible that some aspect of GA biosynthesis and signaling pathways may have been impaired in the *scl33 sc35* mutant.

Although *SCL33* and *SCL30a* are closely related they seem to perform non-redundant functions as loss of either one of these leads to a change in flowering time (late and early respectively). The mutant *scl30a* is early flowering, while the combination of *scl33 scl30a* in the double mutant enhanced the late flowering phenotype beyond that of *scl33* alone under both LD (43d and 26 RLC) and SD (137d and 81 RLC). This shows an enhanced epistatic action of the *scl33* mutant on *scl30a*, in which the absence of the two paralogous gene products belonging to the SCL subfamily drives the pathway into a pronounced late flowering program. The triple mutant *scl33 sc35 scl30a* also showed a strong late flowering phenotype under LD and SD conditions as compared to *scl33*. Since the late flowering mutant phenotypes in *scl33 scl30a* and the triple mutant are more enhanced than in *scl33*, an enhanced epistatic effect of the *scl33* in mutant combinations with its paralogous partner *scl30a*. Thus, the flowering phenotypes of *SR* mutants indicate complex interactions among SRs. For instance, *scl33* was epistatic to *sc35* and *scl30a* in the triple mutant but it was not epistatic to *sc35* in the double mutant combination with *sc35*.

Flowering in plants is regulated by a complex network of signaling pathways that monitor both external cues such as light and temperature and endogenous signals such as hormones (Srikanth and Schmid, 2011). Integration of these signals ensures that plants flower at the right time in its life cycle leading to their reproductive success (Boss et al., 2004; Amasino, 2010; Srikanth and Schmid, 2011). The optimal balance of the splice variants of several genes during plant development play a major role in regulating flowering time and other stress responses (Reddy, 2007; Zhang and Mount, 2009; Carvalho et al., 2010). Alteration of flowering time in *SR* mutants suggests that pre-mRNA splicing plays a role in flowering, and that SRs can either delay or promote splicing. The role of SR genes in thermal induction of flowering is supported by increased expression of SR genes (e.g., RSZ22a, SR30 and SCL33) along with changes in alternative splicing of *FCA*, *MAF2* and *FLM* (Balasubramanian et al., 2006).

Post-transcriptional RNA processing and transcript stability have been shown to regulate the expression of the flowering pathway genes. More than 25 RNA processing factors are involved in the control of flowering, signifying their central role as regulators of post-transcriptional regulators of flowering (Terzi and Simpson, 2008). A number of flowering genes show alternative splicing or alternative polyadenylation of their pre-mRNA (Terzi and Simpson, 2008). For example, the transcript level of plant-specific RNA-binding protein FCA that functions in the autonomous flowering pathway is regulated by alternative splicing and polyadenylation resulting in four FCA (α, β γ and δ) transcripts. FCAγ isoform, one of the four *FCA* transcripts, is crucial for flowering time and the level of the isofrom is controlled by pre-mRNA processing (Quesada et al., 2003).

The RNA recognition motif (RRM) containing proteins comprise one class of spliceosomal proteins. The loss of some spliceosomal proteins has been shown to affect flowering time. A mutant of one of the RRM-containing proteins *SR45*, was shown to display late flowering by influencing the autonomous flowering pathway (Ali et al., 2007). Loss of AtPRP39-1, a protein similar to yeast PRP39 that associates with U1 snRNP and functions in 5' splice recognition, similarly results in a late flowering phenotype (Lockhart and Rymond, 1994; Wang et al., 2007a). Recognition of branch point by U2 snRNP is facilitated by a heterodimeric complex: U2AF35 and U2AF65 (Chusainow et al., 2005). In Arabidopsis, the U2AF35 is encoded by the U2AF35a and U2AF35b genes and suppression of expression of these by RNAi or anti-sense also results in a late flowering phenotype and altered levels of *FLC* and FCA transcripts (Wang and Brendel, 2006b).

Interestingly, in plants, components of U170K and U2AF are known to interact directly with SR and SR-like proteins (Golovkin and Reddy, 1998, 1999; Reddy and Ali, 2011). Hence, SRs are likely to regulate splice site choices by interacting with proteins in U1 snRNP and U2AF. Some evidence in support of this is emerging (Day et al., 2012; Thomas et al., 2012). The SR genes tested here affect flowering time by either directly or indirectly regulating the expression of key flowering genes. This could be due to either direct changes in gene expression of the genes we observed, or indirectly through alternative splicing of other target genes affecting the flowering time genes. It is known that SR genes affect alternative splicing of other SR genes, which therefore might be indirectly regulating the expression of the flowering time genes. It is also possible that altered splicing of transcription factors that regulate expression of flowering genes contributes to observed changes in expression of key flowering genes.

In the studies presented here, the *SCL33* and *SCL30a* both display individual mutant phenotypes and are not fully redundant, whereas in other studies the *SCL33* and *SCL30a* proteins were found to play a redundant role regulating alternative splicing of the long intron of SCL33 (Thomas et al., 2012). This suggests that the paralogs *SCL33* and *SCL30a* may have redundant and non-redundant functions depending on the context of biological function. Although these mutants showed only altered flowering phenotypes under normal growth conditions, it is possible that they may show altered responses to other environmental conditions such as biotic and abiotic stresses. It has been shown that these stresses alter the splicing patterns of several *SR* genes including the ones that are analyzed here (Reddy and Ali, 2011). Therefore, the study of stress responsive phenotypes is another avenue to explore the phenotypes of the *SR* mutants. The availability of these mutants will also be useful to investigate the role of SRs in regulating splicing. This can be accomplished by transiently and/or stably expressing splicing reporters in mutants or protoplasts from mutants or whole plants and analyzing splicing (Thomas et al., 2012). *SR* mutants will also be useful in analyzing global changes in gene expression in plants that lack one, two or three SRs using RNA-Seq approaches. Such studies, coupled with an analysis of global RNA targets for individual SRs, will provide insights into direct and indirect regulation of splicing by individual SRs.

MATERIALS AND METHODS

Generation of Arabidopsis *SR* mutants

The Arabidopsis T-DNA insertion lines for *SCL33* (Salk_058566), *SC35* (Salk_033824), *SCl30a* (Salk_041849) in Columbia background were obtained from the Arabidopsis Biological Resource Center. These lines were grown on MS plates and homozygous mutants were identified by genomic PCR using gene-specific primers and T-DNA primer (LBb1) since none of these mutants could be selected on kanamycin plates because of the silencing of the NPTII gene. Expression of *SR* genes in the mutants was analyzed by RT-PCR using gene-specific primers (Table 2.2). DNase-treated RNA from two-week-old seedlings of wild type (WT) and mutant lines was used for RT-PCR analysis as described earlier (Palusa et al., 2007). The following PCR conditions were used for genotyping: initial denaturation at 94°C for 2 min, followed by 29 cycles at 94°C for 30 sec, 56°C for 30 sec and 72°C for 1 min. The final extension cycle was run at 72°C for 10 min.

Three double mutants were generated by crossing different combinations of homozygous single mutants (*scl33, sc35* and *scl30a*). Two F1 plants from each of three crosses were selfed and seeds were collected. F2 progeny from one of the crosses was genotyped to identify homozygous double mutants. About 60 to 75 soil grown F2 plants for each cross were genotyped. Three homozygous double mutants (*sc35 sc30a, sc35 scl33, scl33 sc30a*) were identified by both genomic PCR and RT-PCR. Triple mutant (*scl33 sc35 scl30a*) was generated by making a cross between two double mutants (*scl33 sc35* and *sc35 scl30a).* Seeds from three selfed F1 plants (*scl33*/*SCL33*, *scl30a*/*SCL30a*, *sc35/sc35)* were collected, but only one line was used for further F2 population analysis. The genomic PCR analysis from 72 F2 progeny resulted in identification of homozygous triple mutants lines, which were further verified by RT-PCR

Flowering time analyses

Seeds utilized in the experiments were collected from fully mature siliques of WT and mutant lines of *Arabidopsis thaliana* ecotype Columbia grown in the University greenhouse. Seeds from WT and all mutants were immersed in water and stratified at 4^0C for 3 days to break the dormancy and later spread onto individual potted soil (PRO-MIX BX mycorise with sphagnum peat moss content almost 75-85% vol.). After 10 days, seedlings from each of the 8 lines were transplanted into individual pots.

For long day (LD) conditions, 36 plants for each line were grown in a walk-in growth chamber under identical conditions with 16 hours light at 100µmol m $^{-2}$ s $^{-1}$ and 8 hours dark at 70% relative humidity and 22˚C. The LD experiments were performed three times and the data were analyzed statistically. Flowering time was scored by counting i) the number of days each plant took to produce the first flower, and ii) the number of rosette leaves at the time of appearance of the first flower.

For short day (SD) conditions, plants were grown under 8 hours of light and 16 hours of darkness and all other condition similar to LD. For the SD experiment, statistical analysis was done on 17 plants. Since these plants are bigger and needed more space to grow, therefore only 17 plants could be grown at a time per genotype.

To perform vernalization experiments, seeds were stratified and then incubated in the dark between layers of filter paper (to remove any moisture) at 4°C. On the 50th day, both vernalized and non-vernalized stratified seeds were potted in soil for 10 days and then the seedlings were subsequently transferred into single pots and grown under either LD or SD conditions. The vernalized and non-vernalized wild type and mutant seeds were grown side by side to compare the flowering time of these two sets of seeds under the same experimental conditions. The number of plants analyzed for LDV was 42 plants and for SDV 13 plants. Analysis of the data, including averages for days to flower, and average number of rosette leaves, the standard deviation and standard error was calculated using Microsoft Excel. Statistical analysis for significance was done using Tukey's paired test using the JMP PRO (http://www.jmp.com/) program with alpha value 0.05.

Analysis of expression of *FLC*

Leaf material from 12 and 29 day-old plants were collected from three plants separately for each line and RNA was extracted using RNAeasy plant mini kit (Qiagen, USA). DNase treated RNA (1.5 µg) was used to synthesize first-strand cDNA using Superscript II reverse transcriptase (Invitrogen, USA), and 100ng RNA in total was taken to run each 20 µl qRT-PCR reaction with FLC primers using SYBR green (TakaRa SYBR® Premix Ex Taq ™ II. The qPCR reaction was run on a Light Cycler ® 480 (Roche) using the following program: Initial denaturation at 95°C for 5 min with ramp rate (rr) as 4.8 (°C/sec) followed by three stages of amplification being [95°C for 10 sec with rr as 4.8 (°C/sec), 60°C for 10 sec with rr as 2.5 (°C/sec) and 72°C for 30 sec with rr as 4.8 (°C/sec)] and three stages of melting curve [95°C for 0.5 sec with rr as 4.8 (°C/sec), 65°C for 1min with rr as 2.5 (°C/sec) and 97°C for continuous cycle with rr as 0.11 (°C/sec) and finally cooling at 40°C for 30 sec with rr as 2 (°C/sec)]. The ct value for actin control in these experiments was standardized to be around 18-20 for each line. The gene expression was calculated with the formula 2^[actin(ct)-flc(ct)] and later the gene expression values of each of the three biological replicates for individual cDNA was used to calculate the average, standard deviation, and standard error in Microsoft excel spread sheet. The ratio as depicted in the bar graphs was calculated by dividing the *FLC* expression of each mutant line to *FLC* expression in WT. Statistical analysis of significance of mutants compared to wt was calculated using Tukey's paired test using JMP PRO program with alpha value being 0.05.

Analysis of FRI and FT expression was done with RNA isolated from 21 days seedlings and used for qRT-PCR analysis as described above using gene specific primers (Table 2.2).

Analysis of *FLC* isoforms

The same cDNA used above for qRT-PCR was used to amplify three FLC isoforms using isoform specific primers (Table 2.2). From the same stock of cDNA as above, 3µl (225ng RNA) is taken for a 20 µl reaction. The PCR conditions used are: initial denaturation at 94°C for 2 min, followed by 29 cycles at 94°C for 30 sec, 56°C for 30 sec and 72°C for 1 min; with a final extension cycle at 72°C for 10 min.

Table 2.2: List of primers for genotyping and analysis of flowering pathway.

Name	**Sequence**
RT-PCR- genotyoping	
SCL33F'	5'-GGTAGATCTCGGTCACGGAG-3'
SCL33R	5'-GTTCCCCACATGTTCC-3'
SC35F	5'-ATCGCTGCTGAACCGATACGAAC-3'
SC35R	5'-CTCCTACGAGGACTGCGGCTTC-3'
SCL30aF	5'-CATGATTGCAGGCAAGAAGA-3'
SCL30aR'	5'-CCAGTAGTAATCCCTAGGA-3'
FLC-isoform specific	
FLC1F	5'-TGTGAGTATCGATGCTCTTGTTCAA-3’
FLC1R	5'- GATGATTATTCTCCATCTGGCTAGCC-3’
FLC2R	5'- CTATCCAAGGAATATCTGGCTAGCC -3'
FLC3F	5’- CACCTTGAGACTGCCCTCTCCG-3’
FLC3R	5’- CACTACTTCTAGACACTTGGAGTTGG-3’
Q-PCR- flowering genes	
qFLCF	5’-AGCCAAGAAGACCGAACTCA-3’
qFLCR	5’-TTTGTCCAGGTGACATC-3’
qFTF	5’-TTGTTGGACACGTTCTTGATC-3’
qFTR	5’-CATCTGGATCCACCATAACCA-3’
qFRIF	5’-CGGCGTTGTCCTCGCCGCGC-3’
qFRIR	5’-GCAACAGCAGCCGTCTCCGCC -3’
qCOF	5’-ATTCTGCAAACCCACTTGCT-3’
qCOR	5'-CCTCCTTGGCATCCTTATCA-3'
qACT2F	5'-GTCGTACAACCGGTATTGTGCTG-3’
qACT2R	5’-CCTCTCTCTGTCCGCATCTTTCATGAG-3’

CHAPTER 2

IDENTIFICATION OF AN INTRONIC SPLiCING REGULATORY ELEMENT INVOLVED IN AUTOREGULATION OF ALTERNATIVE SPLICING OF THE *SCL33* PRE-mRNA

SUMMARY

In Arabidopsis, pre-mRNAs of serine/arginine-rich (SR) proteins undergo extensive alternative splicing (AS). However, little is known about the *cis*-elements and *trans*-acting proteins involved in regulating AS. Using a splicing reporter (*GFP-intron-GFP*), consisting of the GFP coding sequence interrupted by an alternatively spliced intron of *SCL33*, we investigated if *cis*-elements within this intron are sufficient for AS and which SR proteins are necessary for regulated AS. Expression of the splicing reporter in protoplasts faithfully produced all splice variants from the intron, suggesting that *cis*-elements required for AS reside within the intron. To determine which SR proteins are responsible for AS, the splicing pattern of *GFP-intron-GFP* was investigated in protoplasts of three single and three double mutants of SR genes. These analyses revealed that *SCL33* and a closely related paralog, *SCL30a*, are functionally redundant in generating specific splice variants from this intron. Furthermore, SCL33 protein bound to a conserved sequence in this intron, indicating autoregulation of AS. Mutations in four GAAG repeats within the conserved region impaired generation of the same splice variants that are affected in the *scl33 scl30a* double mutant. In conclusion, we identified the first intronic *cis*-element involved in AS of a plant *SR* gene and elucidated a mechanism for autoregulation of AS of this intron.

INTRODUCTION

Alternative splicing (AS), a mechanism for generating multiple transcripts from a single gene, contributes to transcriptome and proteome diversity (Kalsotra and Cooper, 2011). Splice variants from a gene may encode structurally and functionally different proteins that can play important roles in an organism's growth, development and diseases (Reddy, 2007; Kalsotra and Cooper, 2011). Recent genome-wide transcriptome sequencing (RNA-Seq) studies using next generation sequencing indicate that AS is widespread in both animals and plants. Pre-mRNAs from ~60% of multiexon genes in Arabidopsis (Filichkin et al., 2010) and ~48% in rice (Lu et al., 2010) undergo AS. In humans, pre-mRNAs from ~ 95% of intron-containing genes undergo AS with specific isoforms in different tissue types (Pan et al., 2008; Wang and Burge, 2008). Mutations in *cis*-acting elements in RNA or *trans*-acting splicing factors cause misregulation in splicing leading to numerous diseases in humans (Garcia-Blanco et al., 2004; Kalsotra and Cooper, 2011). In plants AS plays important roles in regulating several developmental processes and biotic and abiotic stress responses [reviewed in (Reddy, 2007; Ali and Reddy, 2008b; Gassmann, 2008; Duque, 2011; Reddy and Ali, 2011)].

Although AS is ubiquitous in all multicellular organisms, the frequency of the different types of AS events differs between plants and animals. In plants, intron retention is the most prevalent AS event whereas exon skipping is the most common in vertebrates (Reddy et al., 2012b). The core splicing signals present at the exon-intron (5' splice site [5' SS]), intron-exon boundaries (3' splice site [3' SS]), polypyrimidine tract and the branch point sequence [BPS]) are important for spliceosome assembly. Although there is significant conservation of these core elements across organisms, they alone are not sufficient for constitutive splicing (CS) and AS. Other sequence elements in the pre-mRNA called splicing regulatory elements (SREs) bind to *trans*-acting splicing regulatory proteins. SREs, found in exons (ESEs/ESSs; exonic splicing enhancers/silencers) or in introns (ISE/ISS; intronic splicing enhancers/silencers) play critical roles in both CS and AS (Chasin, 2007; Reddy, 2007). The SREs, which are often short stretches of nucleotides (6-10 nt) (Xiao et al., 2007) function by recruiting *trans*-acting splicing factors that activate or suppress splice site recognition or spliceosome assembly (Chasin, 2007; Barash et al., 2010). These sequence elements, which are generally found in clusters or spaced in regular intervals, influence splice site choice through the specific binding of splicing regulatory proteins such as serine/arginine-rich (SR) proteins or heterogeneous nuclear ribonucleoproteins (hnRNPs) (Long and Caceres, 2009).

In vertebrates, where introns are long, exon and intron specification is thought to occur through ‘exon definition’ that involves interaction of spliceosomal components with the downstream 5’ SS and upstream 3’ SS across the exon (Berget, 1995). In the alternative ‘intron definition’ model, which is thought to occur in organisms with short introns, interactions of spliceosomal proteins occur across the intron between factors recognizing the upstream 5’ SS and the downstream 3’ SS (Berget, 1995). Based on the presence of short introns and the high frequency of intron retention in plants (56% in Arabidopsis and 53.5% in rice as compared to 5% in humans), it is proposed that splice site recognition occurs predominantly by intron definition (Reddy et al., 2012b). Early research on pre-mRNA splicing in plants has shown that AU-rich or U-rich sequences, which are enriched in plant introns, are required for splicing (Filipowicz et al., 1995; Reddy, 2001a; Schuler, 2008). A highly conserved putative AU-rich splicing regulatory *cis*-acting element identified in the gene encoding chloroplast-specific ascorbate peroxidase (chlAPX) isoenzymes represents a plant *cis*-acting element that modulates tissue-specific AS (Yoshimura et al., 2002). The importance of the GC content in exons for efficient splicing

(Carle-Urioste et al., 1997), and an AG-rich exonic element capable of promoting downstream 5' splice site selection have also been reported (McCullough and Schuler, 1997). Using computational tools, a number of putative hexameric exonic splicing enhancers were identified in Arabidopsis (Pertea et al., 2007) . In animal systems many SREs have been experimentally identified that bind splicing factors (Le Guiner et al., 2001; Oberstrass et al., 2005; Chasin, 2007; Fukumura et al., 2007; Jelen et al., 2007). A splicing code was assembled based on hundreds of known features involved in AS to predict exon skipping events in animals (Barash et al., 2010; Rose et al., 2011). However, in plants, aside from a global analysis of gene structure and composition and mutational analysis of splice sites, there has been little experimental or computational analysis done to uncover SREs (Isshiki et al., 2006; Reddy et al., 2012b).

Serine-arginine-rich (SR) proteins are master regulators of CS and AS, each probably regulating the splicing of hundreds to thousands of pre-mRNAs (Long and Caceres, 2009; Reddy and Ali, 2011). SR proteins regulate splicing by binding SREs with their N-terminal RRM domains that mediate RNA–protein interactions and facilitating spliceosome assembly through the C-terminal RS domains, which participate in protein-protein interactions and in some cases interact with RNA (Zahler et al., 1992; Caceres and Krainer, 1993; Le Guiner et al., 2001; Long and Caceres, 2009; Reddy and Ali, 2011). Analysis of pre-mRNA splicing of 18 Arabidopsis *SR* genes revealed extensive AS (Palusa et al., 2007a). Remarkably, over 90 transcripts are produced from pre-mRNAs of 14 *SR* genes, representing more than a five-fold increase in the *SR* transcriptome, and many splice variants are the targets of nonsense-mediated decay (Palusa et al., 2007a; Palusa and Reddy, 2010). Most of the AS variants produced from *SR* pre-mRNAs are generated by AS of introns, with the frequency of occurrence of AS events highest in the longest introns (Palusa et al., 2007a). Most significantly, the AS of some SR genes is controlled in a developmental and tissue-specific manner and altered in response to diverse stresses, suggesting that regulation of AS is likely important for development and stress responses (Reddy and Ali, 2011). Ectopic expression of *RS2Z33* was shown to alter AS of its own pre-mRNA and of other SRs (*atSRp30, atSRp34*), with pleiotropic changes in plant development (Kalyna et al., 2003). Overexpression of SR30 has also been shown to alter alternative splicing of its own pre-mRNA and that of other *SR* pre-mRNAs (Lopato *et al.*, 1999). A loss-of-function mutant *(sr45-1)* of SR45, an SR-like gene, exhibited multiple developmental abnormalities (e.g, delayed flowering, reduced root growth, narrow leaves and altered number of petals and stamens), increased sensitivity to glucose and abscisic acid and altered splicing patterns of several SR genes (Ali et al., 2007; Zhang and Mount, 2009; Carvalho et al., 2010). Interestingly, the long splice variant of *SR45* complemented the flower petal phenotype whereas the short isoform complemented the root growth (Zhang and Mount, 2009).

The lack of an *in vitro* splicing system derived from plant cells has hampered progress in understanding regulated splicing in plants as compared to animal systems. Since the pre-mRNAs of plant SRs show extensive AS, they can serve as excellent candidates to elucidate the mechanisms(s) involved in regulated AS in plants. The pre-mRNAs of an Arabidopsis *SR* gene, *SCL33,* undergo AS and produce at least nine splice variants, eight of them are generated due to AS of the third intron (Palusa and Reddy, 2010). To determine if the sequence elements within the intron are sufficient for regulation of AS we developed an *in vivo* splicing assay using a splicing reporter in protoplasts. To identify which SR proteins are involved in regulated splicing of this intron, we analyzed AS of the splicing reporter in protoplasts from three single and three double mutants of Arabidopsis *SR* genes. The results presented here show that all the signals necessary for AS of the third intron of *SCL33* are present within the intron. Our work also shows that two related SR proteins (SCL33 and SCL30a) are functionally redundant in producing specific isoforms. Furthermore, using RNA binding studies we identified a 92 nt region with multiple GAAG repeats in the *SCL33* intron that binds to SCL33 protein, suggesting autoregulation of *SCL33* AS. Mutational analysis of the GAAG repeats confirmed the importance of these elements in AS.

RESULTS

Signals for alternative splicing of the *SCL33* intron reside within the intron

We have previously shown that pre-mRNA from *SCL33* undergoes AS in different tissues, including leaves, and all the splice variants except one are generated by intron retention, alternative 3' splice site selection or by using both alternate 3' and 5' splice sites (Palusa et al., 2007a). The *SCL33* third intron was chosen as a model to study AS regulation and identify splicing factors that regulate AS events using an *in vivo* protoplast system. We first tested if the AS pattern of *SCL33* in leaves is identical in mesophyll protoplasts by RT-PCR using RNA from leaves and protoplasts. As shown in Figure 3.1, the splicing pattern of *SCL33* in protoplasts is similar to leaves, indicating that the protoplast system was suitable for studying the regulation of AS of this gene. To identify the splice sites used in producing these splice variants, cDNA prepared from protoplasts was used to amplify all splice variants from the SCL33 3rd intron, which were then sequenced.

To determine if the signals necessary for AS of the third intron reside within the intron, we developed a splicing reporter construct that can be used to monitor AS of this intron *in vivo* using a protoplast system. The splicing reporter construct was made by cloning the entire third intron of *SCL33* (765bp) into the coding region of GFP to generate *GFP-SCL33 intron-GFP*

(*GFP-INT-GFP*) (Figure 3.2). Splicing of the intron would result in GFP fluorescence. Furthermore, the production of splice variants can be monitored by RT-PCR using GFP-specific primers. We transfected protoplasts with this construct or the uninterrupted *GFP* (Figure 3.2) as a control. Expression of the splicing reporter in wild-type protoplasts showed GFP fluorescence (Figure 3.2) although less fluorescent than the uninterrupted GFP (Figure 3.2), suggesting that the intron is excised properly from the splicing reporter.

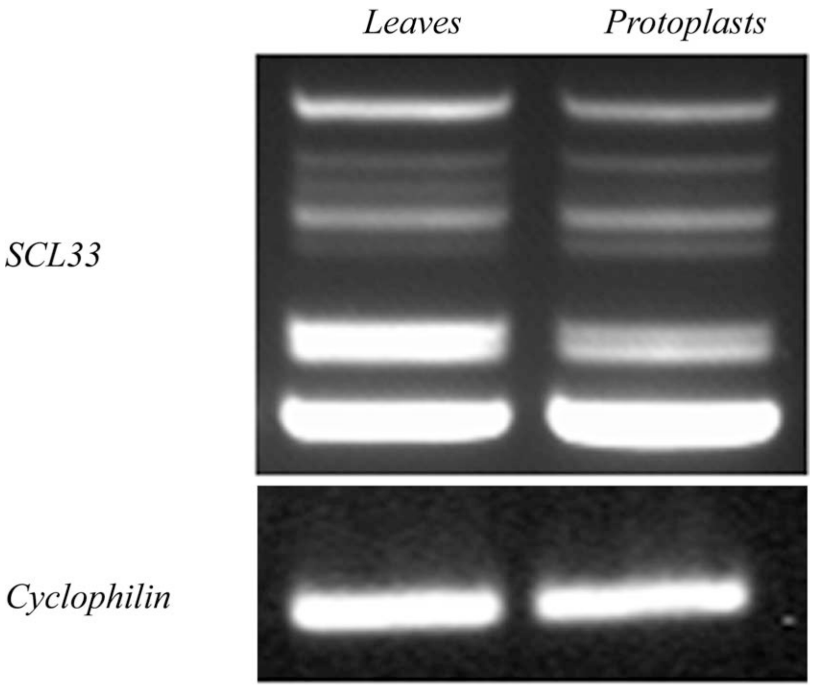


Figure 3.1. Alternative splicing of *SCL33* pre-mRNA in three week-old leaves and protoplasts isolated from the same age leaves.

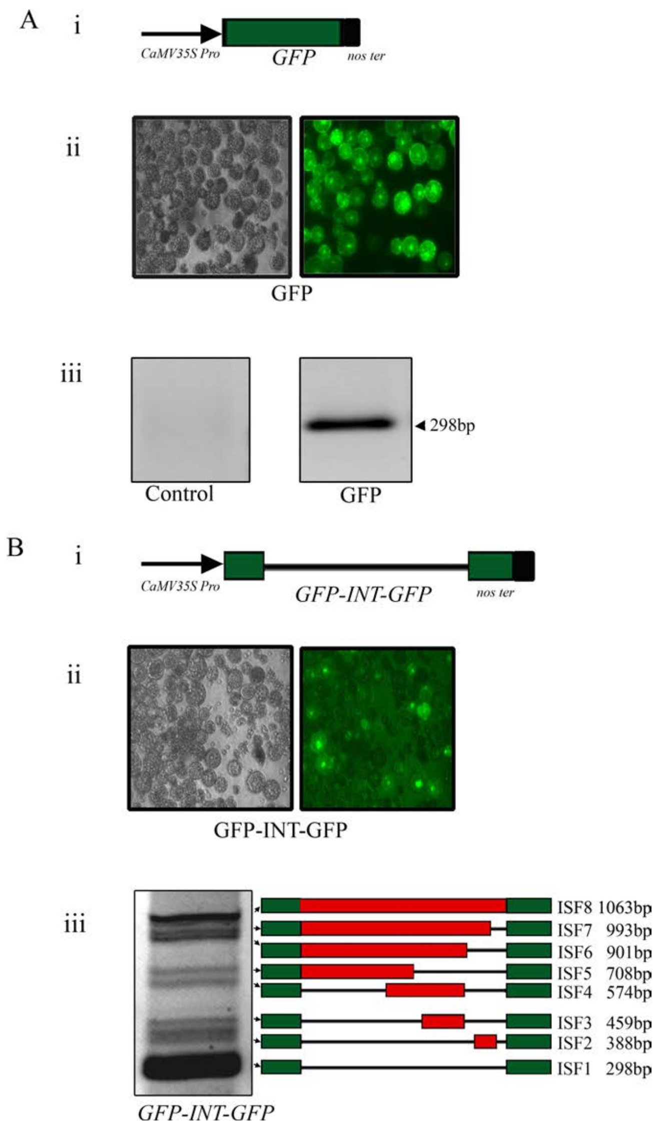


Figure 3.2: Analysis of AS of *SCL33* intron 3 in a reporter gene using Arabidopsis protoplasts. (A) Expression of *CaMV35S-GFP* construct in protoplasts i) Schematic diagram of *CaMV35S* promoter-GFP construct; *nos ter*, *nos* terminator ii) Light microscope image (left) and fluorescence image (right) of protoplasts transformed with *CaMV35S-GFP* construct; iii) RT-PCR with RNA from untransformed protoplasts (control) showed no GFP transcript (left) whereas RNA from protoplasts transformed with *CaMV35S-GFP* construct showed a 298 bp fragment (right). **B)** AS of *SCL33*-intron in Arabidopsis protoplasts; i) Diagram of *CaMV35S-GFP-SCL33 intron-GFP* (*GFP-INT-GFP*) construct; ii) Light microscope image (left) and fluorescence image (right) of protoplasts transformed with *GFP-INT-GFP* construct; iii) Detection of splice variants generated from *GFP-INT-GFP* by RT-PCR using *GFP*-specific primers. Schematic diagram of each splice isoform (ISF) is shown alongside of each band. The numbers indicate the length of each ISF. Green,exon; red, included intron; black, excluded intron.

We then asked if the intron is alternatively spliced as in the endogenous gene and if so are the AS events and the sites used to generate the splice variants identical to the native gene. To address these questions, we isolated RNA from the protoplasts expressing the *GFP* and *GFP-INT-GFP* construct and performed RT-PCR using *GFP* specific forward and reverse primers. As expected, a single product was seen in GFP transfected protoplasts (Figure 3.2A iii). In the *GFP-INT-GFP* all expected eight isoforms of the native *SCL33* gene were obtained (Figure 3.2B iii). To determine if the splice variants are identical to those that are produced from the native gene, we cloned and sequenced all splice variants from the reporter gene. Alignment of isoforms generated from intron 3 in the native *SCL33* gene with those generated from *GFP-INT-GFP* has revealed that, remarkably, all eight isoforms are identical to the splice forms from the endogenous gene (Figure 3.2Biii, schematic diagram). The same isoforms were obtained from the leaves of the stable transgenic lines expressing the *GFP-INT-GFP*, validating the use of the transient expression system for AS. Isoform 8, the largest one, is produced by intron retention, and isoform 1 is just the GFP product produced by complete removal of the intron. Three isoforms (5 to 7) are generated by an alternative 5’ SS selection whereas isoforms 2, 3 and 4 are produced by using both alternative 5’ and 3’ SS. Among the eight forms, three isoforms (3, 4, 6) share the same 5’ splice site for the splicing of the second part of the intron, but have different 3’ splice sites for splicing of the first part of the intron (Figure 3.2Biii). These results demonstrate experimentally that the third intron of *SCL33* has all the necessary signals to faithfully undergo AS and suggest that the sequences in other parts (exons or other introns) of *SCL33* are not required for AS of this intron.

SCL33 protein binds to a 92-nucleotide segment of the *SCL33* intronic RNA

Some SR proteins in animals and plants are known to autoregulate pre-mRNA splicing (Long and Caceres, 2009; Reddy and Ali, 2011). However, in plants the binding regions in native pre-mRNAs have not been identified for any of the SR proteins. To investigate whether the SCL33 protein interacts directly with intron 3 of *SCL33*, we performed electrophoretic mobility shift assays (EMSA) using different regions of labelled *SCL33* intron (Figure 3.3A) and purified SCL33 protein. As shown in Figure 3.3B, the 5' side of the intron (P1) did not bind to the SCL33 protein whereas P2 RNA shifted when recombinant SCL33 was added. The formation of RNA-protein complex increased with increasing concentrations of purified protein. This result suggests at least one binding region for SCL33 in the 3' segment (P2) of the intron. To map the binding region in P2, labelled RNA from two shorter fragments (P3 and P4) were used in EMSA. Both RNAs bound to SCL33 and the extent of binding also increased with an increase in protein concentration. Since P4 is the smallest fragment, with 92 nt, we conclude that it has a binding site for SCL33. To determine the specificity of the P4 binding to SCL33 we performed a competition assay with increasing amount of cold P4 RNA. As shown in Figure 3.3C, protein-RNA complex formation was observed between the SCL33 protein and P4 RNA (Lane 2), whereas addition of increasing concentrations of cold competitor RNA reduced binding. At a concentration of 50X (lane 7), the competitor RNA completely abolished P4 binding indicating that the interaction between SCL33 and this intron segment is specific. To further demonstrate the specificity of the SCL33 interaction with P4, we performed EMSA with P4 RNA with purified SR45, an SR-like protein (Golovkin and Reddy, 1999). The P4 RNA showed no binding to SR45 (Figure 3.4).

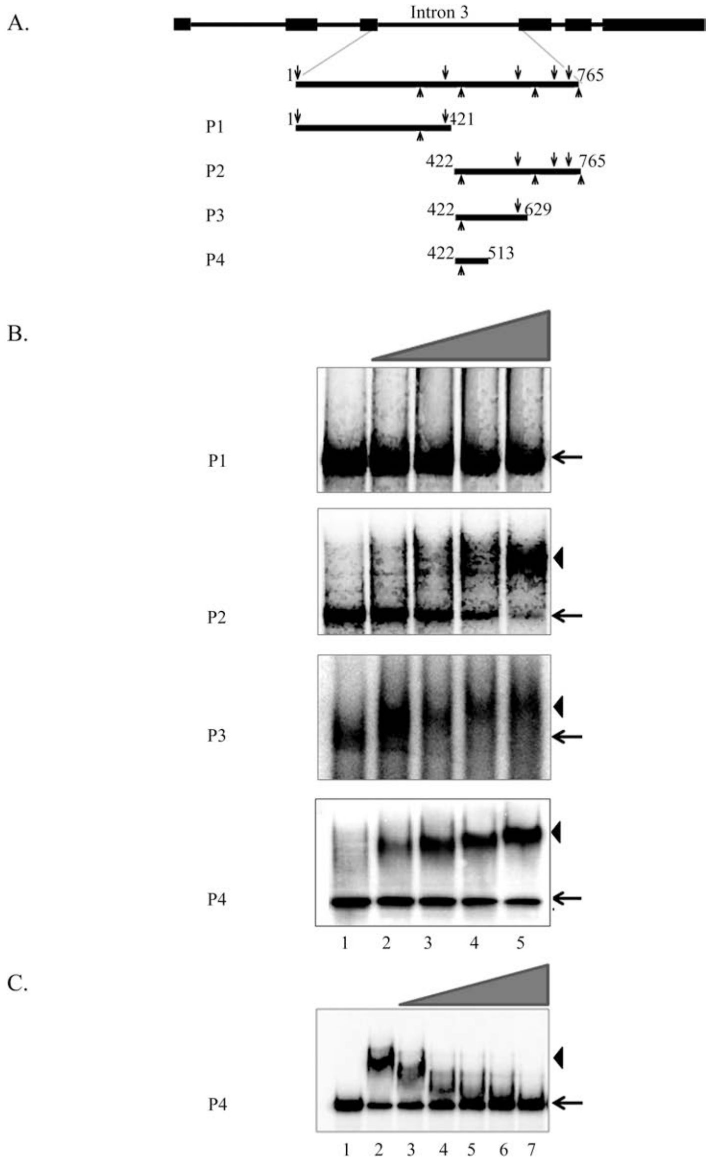


Figure 3.3: SCL33 protein binds to a specific region of SCL33 intron. A) Schematic diagram of different regions of the third intron of *SCL33* that were used to generate RNA probes for Electrophoretic Mobility Shift Assay (EMSA). P1 to P4 represent different parts of intron 3 as illustrated. The number indicates the start and end nt position of each probe relative to the intron 5'SS. Arrows and arrowheads show the 5' and 3' splice sites of different isoforms generated by AS, respectively. **B)** EMSA with P1, P2, P3 and P4 RNA probes using purified SCL33 protein. Lane1, free probe, lanes 2-5 have increasing concentration of SCL33 (60, 120, 180, 300ng). An arrow indicates free probe and the RNA-protein complexes are indicated by an arrowhead. **C)** The binding of SCL33 to P4 is competed by cold P4. Lane 1, free probe, Lane 2, Probe + SCL33 protein (300 ng). Lanes 3-7, same as lane2 with increasing concentration of cold P4 (10X, 20X, 30X, 40X and 50X).

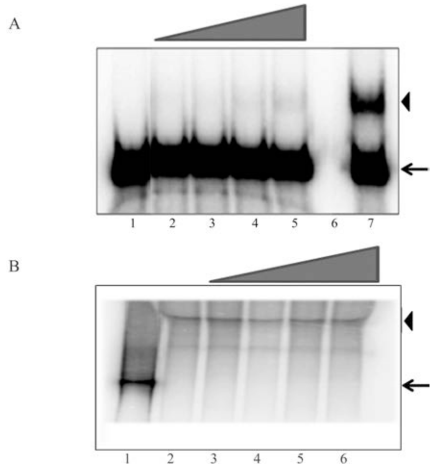


Figure 3.4. A) SR45 does not bind P4 RNA B) Excess amount of cold P1 RNA does not abolish binding of P4 RNA to SCL33.

To further confirm the binding specificity between P4 and SCL33, we performed a competition assay where increasing concentrations of cold P1 RNA were added to the SCL33-P4 labeled complex. Addition of excess P1 (50X) did not reduce the amount of SCL33-P4 complex (Figure 3.4). Together, these results establish that SCL33 binds specifically to a 92 nt region in the *SCL33* third intron and suggests SCL33 may auto-regulate AS by binding its own intronic RNA.

SCL30a is the closest paralog of SCL33 (Richardson et al., 2011) and the third intron of SCL30a also undergoes AS, and the location of AS sites in some isoforms is similar to SCL33 (Palusa et al., 2007a). To see if there is any sequence conservation between the third intron of *SCL33* and *SCL30a*, we aligned the nucleotide sequences of these introns using TCOFFEE (http://tcoffee.vital-it.ch/). Interestingly, the 92 nt region that bound to SCL33 is highly conserved (90% identity) between *SCL33* and *SCL30a* and all the alternative splice sites in both genes are near or within the conserved 92 nt region (see Figure 3.5). In addition, we found that four closely spaced purine-rich GAAG repeats, which are known exonic splicing regulatory elements in plants and animals, (Chasin, 2007; Pertea et al., 2007) are found in the third intron of both genes, suggesting that they may be important for SCL33 interaction and AS.

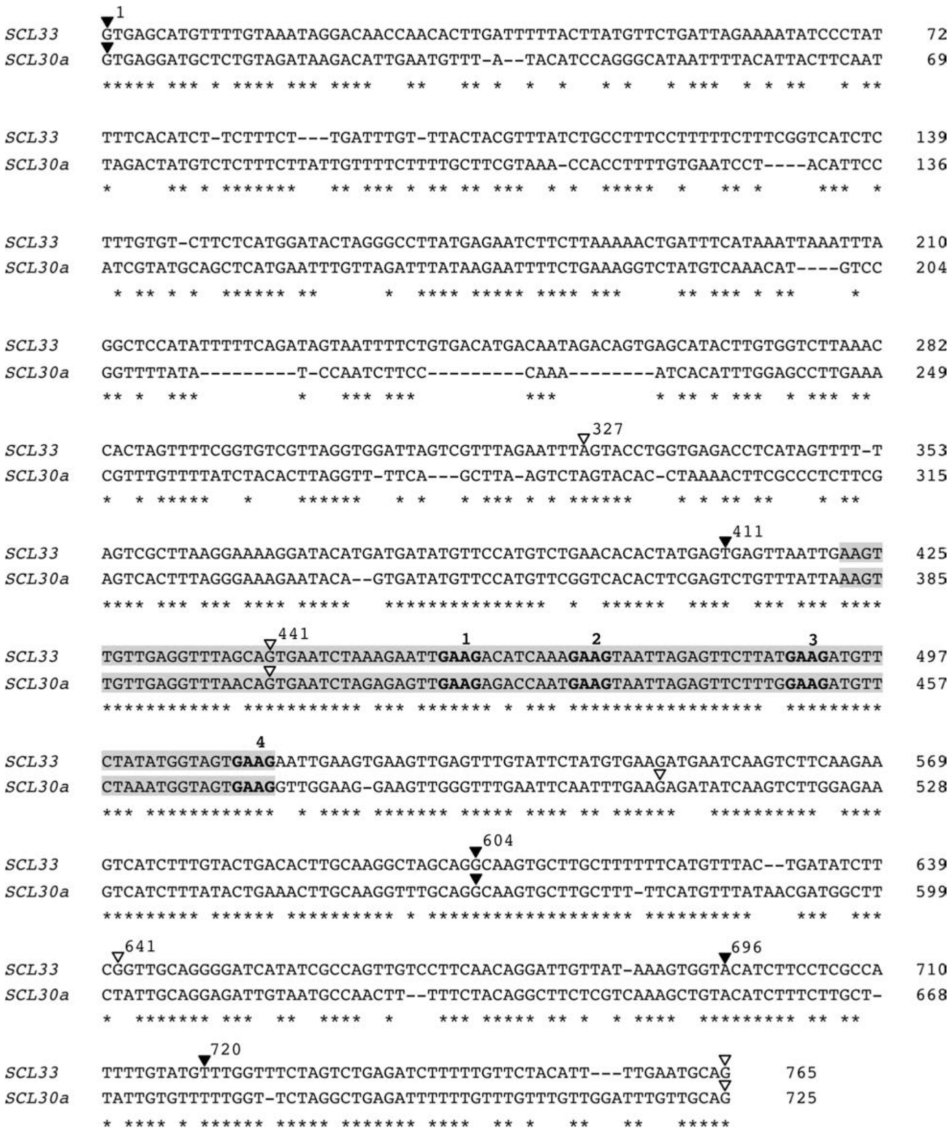

1
SCL33 GTGAGCATGTTTTGTAAATAGGACAACCAACACTTGATTTTTACTTATGTTCTGATTAGAAAATATCCCTAT 72
SCL30a GTGAGGATGCTCTGTAGATAAGACATTGAATGTTT-A--TACATCCAGGGCATAATTTTACATTACTTCAAT 69
SCL33 TTTCACATCT-TCTTTCT---TGATTTGT-TTACTACGTTTATCTGCCTTTCCTTTTTCTTTCGGTCATCTC 139
SCL30a TAGACTATGTCTCTTTCTTATTGTTTTCTTTTGCTTCGTAAA-CCACCTTTTGTGAATCCT----ACATTCC 136
SCL33 TTTGTGT-CTTCTCATGGATACTAGGGCCTTATGAGAATCTTCTTAAAAACTGATTTCATAAATTAAATTTA 210
SCL30a ATCGTATGCAGCTCATGAATTTGTTAGATTTATAAGAATTTTCTGAAAGGTCTATGTCAAACAT----GTCC 204
SCL33 GGCTCCATATTTTTCAGATAGTAATTTTCTGTGACATGACAATAGACAGTGAGCATACTTGTGGTCTTAAAC 282
SCL30a GGTTTTATA---------T-CCAATCTTCC---------CAAA--------ATCACATTTGGAGCCTTGAAA 249
327
SCL33 CACTAGTTTTCGGTGTCGTTAGGTGGATTAGTCGTTTAGAATTTAGTACCTGGTGAGACCTCATAGTTTT-T 353
SCL30a CGTTTGTTTTATCTACACTTAGGTT-TTCA---GCTTA-AGTCTAGTACAC-CTAAAACTTCGCCCTCTTCG 315
411
SCL33 AGTCGCTTAAGGAAAAGGATACATGATGATATGTTCCATGTCTGAACACACTATGAGTGAGTTAATTGAAGT 425
SCL30a AGTCACTTTAGGGAAAGAATACA--GTGATATGTTCCATGTTCGGTCACACTTCGAGTCTGTTTATTAAAGT 385
441
1
2
3
SCL33 TGTTGAGGTTTAGCAGTGAATCTAAAGAATTGAAGACATCAAAGAAGTAATTAGAGTTCTTATGAAGATGTT 497
SCL30a TGTTGAGGTTTAACAGTGAATCTAGAGAGTTGAAGAGACCAATGAAGTAATTAGAGTTCTTTGGAAGATGTT 457
4
SCL33 CTATATGGTAGTGAAGAATTGAAGTGAAGTTGAGTTTGTATTCTATGTGAAGATGAATCAAGTCTTCAAGAA 569
SCL30a CTAAATGGTAGTGAAGGTTGGAAG-GAAGTTGGGTTTGAATTCAATTTGAAGAGATATCAAGTCTTGGAGAA 528
604
SCL33 GTCATCTTTGTACTGACACTTGCAAGGCTAGCAGGCAAGTGCTTGCTTTTTTCATGTTTAC--TGATATCTT 639
SCL30a GTCATCTTTATACTGAAACTTGCAAGGTTTGCAGGCAAGTGCTTGCTTT-TTCATGTTTATAACGATGGCTT 599
641
696
SCL33 CGGTTGCAGGGGATCATATCGCCAGTTGTCCTTCAACAGGATTGTTAT-AAAGTGGTACATCTTCCTCGCCA 710
SCL30a CTATTGCAGGAGATTGTAATGCCAACTT--TTTCTACAGGCTTCTCGTCAAAGCTGTACATCTTTCTTGCT- 668
720
SCL33 TTTTGTATGTTTGGTTTCTAGTCTGAGATCTTTTTGTTCTACATT---TTGAATGCAG 765
SCL30a TATTGTGTTTTTGGT-TCTAGGCTGAGATTTTTTGTTTGTTTGTTGGATTTGTTGCAG 725

Figure 3.5: Alignment of nucleotide sequence of the third intron of the *SCL33* and *SCL30a* genes showing sequence identity and experimentally determined alternative splice sites. Asterisks indicate the same nucleotide in both introns. The conserved 92 nt sequence highlighted in grey is sufficient for binding to SCL33 protein as shown in Fig. 3.3 and contains four GAAG motifs named 1 to 4. These motifs were mutated as described in Figure 3.8. Filled arrowheads above the *SCL33* sequence indicate the 5' splice sites and hollow arrowheads depict the 3' splice site of various splice variants of *SCL33*. The nucleotide positions in *SCL33* and *SCL30a* are indicated at the right side of the sequence. The position of 5' and 3' splice sites of *SCL30a* splice variants are also marked above the *SCL30a* sequence with arrowheads.

SCL33 and SCL30a are functionally redundant in regulating AS of *SCL33* intron 3

SR proteins in animals are known to regulate both CS and AS by binding to SREs and recruiting spliceosomal components (Allo et al., 2009). Although many pre-mRNAs of plant SR genes are alternatively spliced (Palusa et al., 2007a; Reddy and Ali, 2011), little is know about the *trans*-acting splicing factors that regulate AS in these genes. To identify the SR proteins that may regulate AS of *SCL33*, we used a genetic approach to address the role of three of the eighteen Arabidopsis SR genes (Barta et al., 2010) in AS of *SCL33*. We identified loss-of-function T-DNA insertion mutants of *scl33*, *sc35*, and *scl30a*, with insertions in the exon of each gene (Figure 3.6A) and generated three double mutants (*scl33 scl30a*, *scl33 sc35*, *scl33 scl30a*) since SR proteins, especially the closely related ones have redundant functions. We confirmed homozygosity of the mutants by genomic PCR (Figure 3.6B) using primers from the genes and T-DNA insert. The homozygous lines showed no expression of transcripts corresponding to the mutated gene(s) in single and double mutants (Figure 3.6C), suggesting that these mutants are complete knockouts. Since the T-DNA insertion in the *scl33* mutant is in the last exon, we performed RT-PCR with primers corresponding to exons flanking intron 3, which lie upstream of the T-DNA insertion site, to see if any splice variants are produced in the mutant. No splice variants were detected with these primers, confirming that *scl33* is a complete loss-of-function mutant (Figure 3.7). We then used the protoplasts from all the mutant lines to monitor AS of the splicing reporter to determine the functions of these SRs in AS.

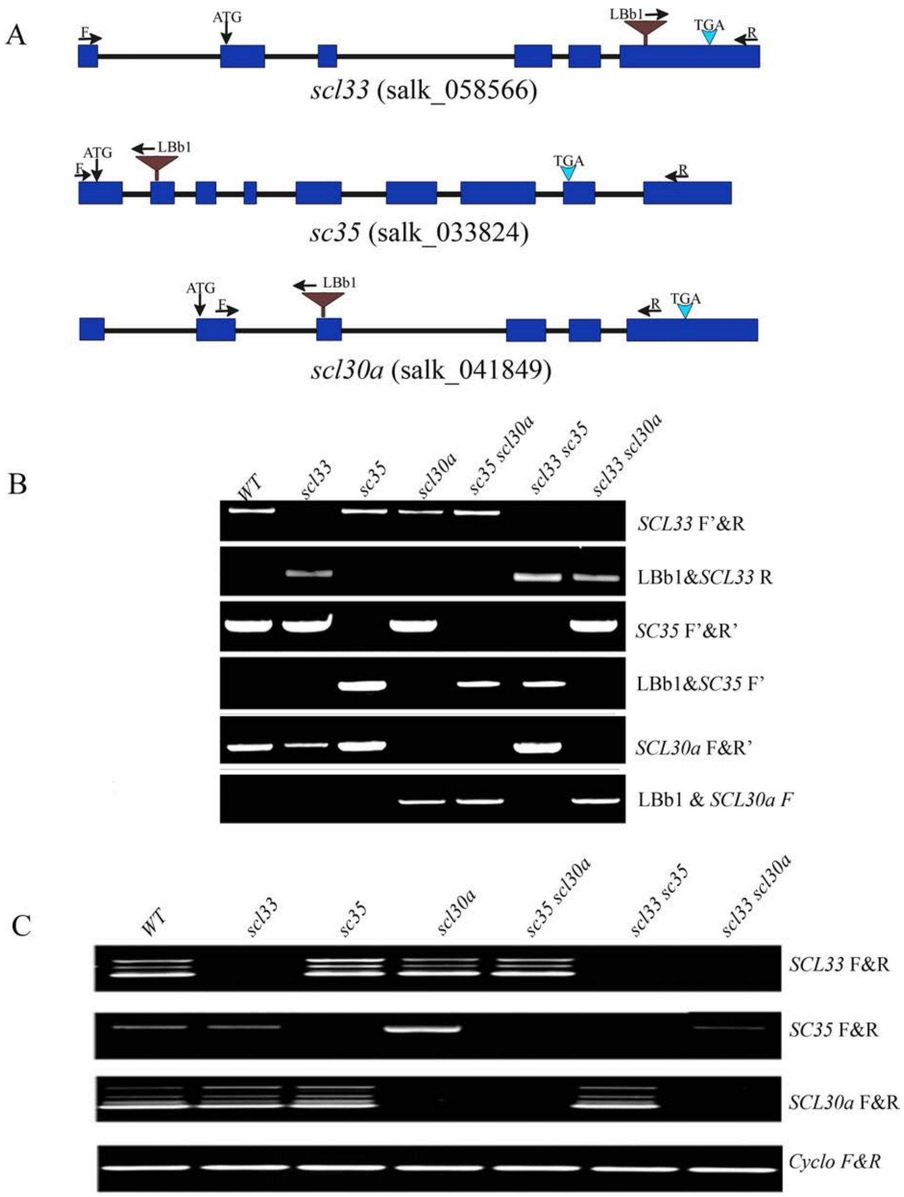


Figure 3.6: **Genotypic characterization of single (*scl33, sc35, scl30a)* and double *(sc35 scl30a, scl33 sc35, scl33 scl30a*) mutants using genomic PCR and RT-PCR. (A)** Schematic diagram showing T-DNA insertion in *SCL33, SC35, SCL30a*. Blue boxes represent exons and lines between exons indicate introns. The triangle represents the T-DNA insertion site. *SCL33, SC35,* and *SCL30a* genes have insertions in the last exon, second and third exon, respectively. LBb1 primer is in the T-DNA insert. FP and RP are the gene-specific primers. **(B)** Verification of T-DNA insertion in each of these genes by genomic PCR using LBb1 and gene-specific primers. **(C)** RT-PCR analysis of expression of all three genes in wild type (WT), single mutants (*scl33, sc35, scl30a)* and double mutants *(sc35 scl30a*, *scl33 sc35*, *scl33 scl30a*) using gene-specific primers.

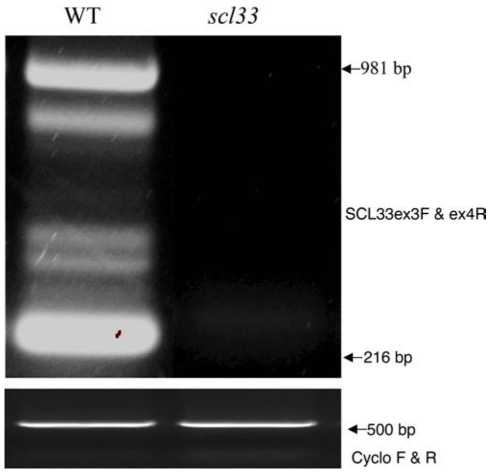


Figure 3.7. Transcripts corresponding to the region prior to the T-DNA insertion site are not present in the *scl33* mutant.

Protoplasts from wild type and three single and double mutants were transfected with the *GFP-INT-GFP* construct and splicing was analyzed by RT-PCR. Splicing of this reporter in five of the six mutants is similar to that of wild type. Only the *scl33 scl30a* double mutant showed altered splicing where two isoforms (isoform 3 and 6) were missing (Figure 3.8A). These results indicate that the *SCL33* and *SCL30a* genes are functionally redundant, and generation of all splice variants from the *SCL33* intron can occur in the presence of either *SCL33* or *SCL30a* but not in the absence of both. Since these two SRs are the closet paralogs and share 66% identity and 74% similarity in amino acid sequence it is not surprising that the lack of one SR is compensated for by the other SR.

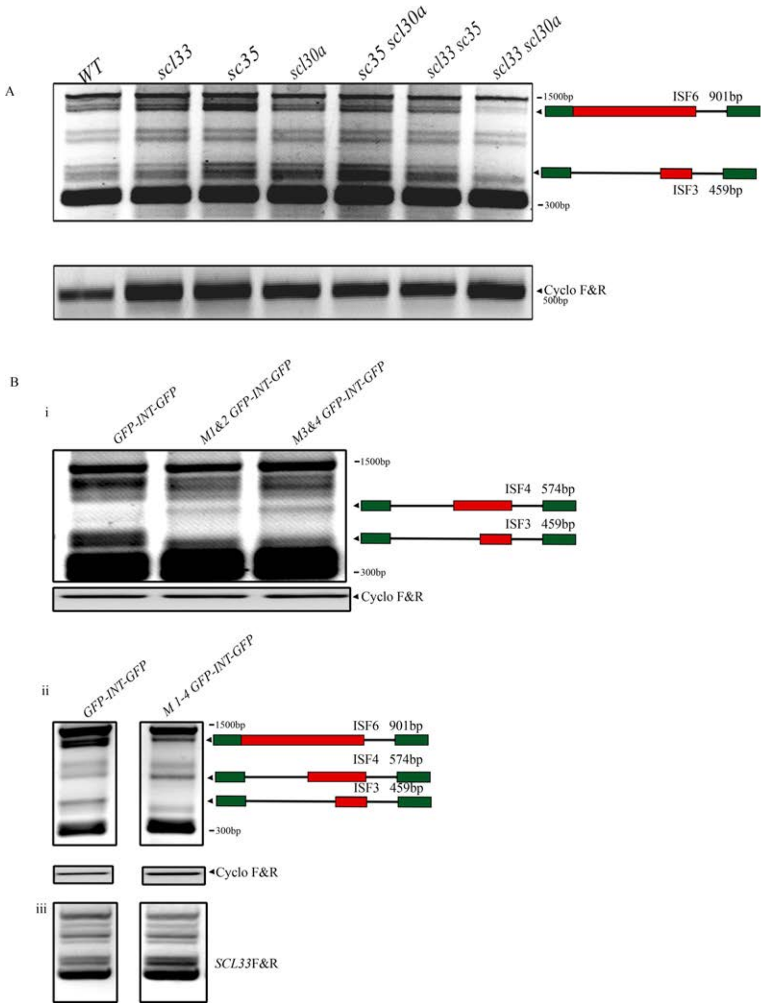


Figure 3.8: Analysis of AS of *SCL33* intron in wild type (WT) and six mutants of the Arabidopsis *SR* genes. (A) Protoplasts from WT, three single *(scl33, sc35, scl30a*) and three double (*sc35 scl30a, scl33 sc35, scl33 scl30a)* mutants were transformed with *GFP-INT-GFP* construct and splice variants were analyzed by RT-PCR using GFP primers. The gel shows RT-PCR products with changes in the splicing pattern in double *scl33 scl30* mutant as compared to WT, and other single and double mutants. The missing isoforms with sizes are shown along the side. **(B)** Wild-type protoplasts were transformed with either *GFP-INT-GFP* or mutant *GFP-INT-GFP*s forms M1&2, M3&4 and M 1-4, in which 4 GAAG elements 1 and 2, 3 and 4, or all 4, respectively (See Figure 3.5) were changed to CTTC. (i) AS of the mutated *GFP-INT-GFP* (M1&2 and M3&4) and *GFP-INT-GFP* in WT protoplasts. Schematic diagram of altered isoforms are shown next to the gel; ii) AS of mutated *GFP-INT-GFP* (M 1 to 4) and *GFP-INT-GFP* in WT protoplasts. Schematic diagram of altered isoforms are shown alongside the gel ; iii) Splicing of the endogenous *SCL33* is not changed in protoplasts transformed with either *GFP-INT-GFP* or *GFP-INT-GFP* M1-4 mutant . The cyclophilin control for each experiment is shown below, and for all gels the isoform structure is represented alongside. Numbers next to arrowheads indicate the size of amplified products.

The conserved GAAG repeats are required for producing specific isoforms

As described above the alignment of the third intron of *SCL33* and *SCL30a*, which undergoes AS, revealed considerable sequence conservation especially at the 3' end where AS takes place, and contains four closely spaced GAAG sequence elements in the region that was shown to bind SCL33 (Figures 3.3, 3.4). To test if these GAAG elements are important for generation of one or more splice variants, we generated three mutants (M1&2, M3&4 and M1-4) where the GAAG sequence is changed to CTTC (Figure 3.5). In M1&2, the first two GAAG elements are mutated, in M3&4 the last two elements are changed whereas in M1-4 all four elements were mutated. The wild type *SCL33* intron in *GFP-INT-GFP* splicing reporter was replaced with the three mutated forms to monitor their splicing in protoplasts from wild type. In experiments comparing wild-type and mutated introns splicing, the M1&2 and M3&4 mutants with the first or last two GAAG elements mutated, respectively, showed almost complete loss of isoform 3 and increase in isoform 4 (Figure 3.8B i) whereas M1-4 intron with mutations in all four GAAG elements resulted in the loss of isoforms 3 and 6 and an increase in isoform 4 (Figure 3.8B ii). Analysis of AS of endogenous *SCL33* splicing by RT-PCR using the SCL33 specific primer showed no change in its splicing pattern (Figure 3.8B iii), confirming that the splicing pattern is changed only in the mutated *GFP-INT-GFP* constructs. Interestingly, all three isoforms that are affected in M1-4 transfected cells have common 5' and 3' splice sites at the 3' end of the intron (Figure 3.8Bii, right). These results suggest that generation of these three isoforms requires a transacting splicing factor that recognizes the GAAG elements for accurate splicing of the 3' region of the intron. From this data we can also conclude that the number of GAAG elements plays a role in the reduction or increase of splice variants, as we observed that when two GAAG elements are mutated at a time, only two isoforms are affected, but mutating all four GAAG elements affected three isoforms. Interestingly, the double mutant (*scl33 scl30a*) displays a similar splicing pattern to M1-4-*GFP-INT-GFP*, suggesting that the SCL33 or SCL30a proteins bind to these *cis* elements and are necessary for accurate splicing of the *SCL33* intron. The binding of the SCL33 protein to the 92 nt fragment (Figure 3.3) supports this hypothesis.

DISCUSSION

Alternative splicing (AS) is highly prevalent in plants and is thought to play an important role in increasing proteome diversity as well as in regulating gene expression at the post-transcriptional level (Reddy et al., 2012b). Some of the regulators of AS are RNA-binding proteins such as the members of the SR family that bind to specific RNA sequences in pre-mRNA and aid in spliceosome assembly at the weak splice sites and contribute to regulated AS. The pre-mRNAs from plant SR genes themselves undergo extensive AS but the mechanisms that regulate AS are poorly understood. Splicing of pre-mRNA substrates in a cell-extract, which has been used extensively in animals (Chasin, 2007; Long and Caceres, 2009), is a powerful method to study SREs and splicing factors. However, the lack of such a system derived from plant cells hampered progress in this area in plants. The use of protoplasts to transiently express splicing reporters offers an alternate and powerful approach to identify *trans*-acting factors that regulate AS. Here we have developed an *in vivo* splicing reporter assay to study regulation of AS of the *SCL33* intron. Use of this splicing reporter in protoplasts from mutant plants lacking one or more SRs or other putative splicing regulators offers a novel and tractable way to study the function of a given SR in AS.

Using *in vivo* splicing assays with a splicing reporter containing the *SCL33* intron we show that all splice variants from this intron are accurately generated, suggesting that all signals required for AS reside within the intron of the *SCL33* gene. This supports the intron definition model for AS regulation of this intron. Using this reporter system, we performed further experimental characterization of the sequence elements located within this intron and identified SR proteins that are involved in regulating AS. RNA-protein interaction studies and AS analysis in mutant protoplasts presented here indicate that SCL33 autoregulates its AS. EMSA analysis revealed that the purified SCL33 protein binds to a 92 nt region in the *SCL33* third intron. Four lines of evidence indicate that the observed interaction between the 92 nt fragment (P4) and SCL33 is specific. First, the 5' end of the intron (part 1) did not bind SCL33. Second, the binding of SCL33 to P4 can be eliminated with excess cold RNA. Third, SR45, an SR-like protein, does not bind to this fragment (Figure 3.4A). Finally, the binding of SCL33 to P4 was not affected by adding excess amount of cold P1 RNA (Figure 3.4B). Although, the AS of the *SCL33* intron is not altered in the *scl33* mutant, in the double mutant (*scl33 scl30a*) where a closely-related paralog is also lost, specific isoforms are missing, suggesting that SCL33 and SCL30a have a redundant function. Mutations in the GAAG elements of the 92 nt segment in the *SCL33* intron, which binds to SCL33 protein, resulted in an altered splicing pattern similar to that observed with the wild-type intron in the *scl33 scl30a* double mutant. This suggests that the GAAG-element containing intron sequence is critical for SCL33 binding and regulation of normal AS. Furthermore, the affected 5' splice site at nucleotide position 604 is adjacent to the SCL33 binding region. A model illustrating the mechanism(s) by which SCL33 autoregulates AS is presented in Figure 3.9. Since the affected isoforms have the same 5' splice site, it is possible that SCL33 binds to the 92 nt region in the middle of the intron and recruits U1 snRNP to the 5' splice site either by directly interacting with one of the U1 snRNP proteins or other SRs that interact with U1 snRNP. There is prior evidence for such interactions with SCL33. We have previously shown that SCL33 can directly interact with U1-70K, one of the U1 snRNP proteins (Golovkin and Reddy, 1999) and also an SR-like protein (SR45), which is known to interact with U170K (Golovkin and Reddy, 1999; Reddy, 2007). Our results also suggest that SC35 alone or in combination with either SCL33 or SCL30a does not regulate AS of the *SCL33* intron, as the pattern of AS in single (*sc35)*, and double (*sc35 scl33* or *sc35 scl30a)* mutants is not altered. The observation that only certain isoforms are affected in the *scl33 scl30a* double mutant indicates that other SRs may also be involved in regulating the splicing of this intron.

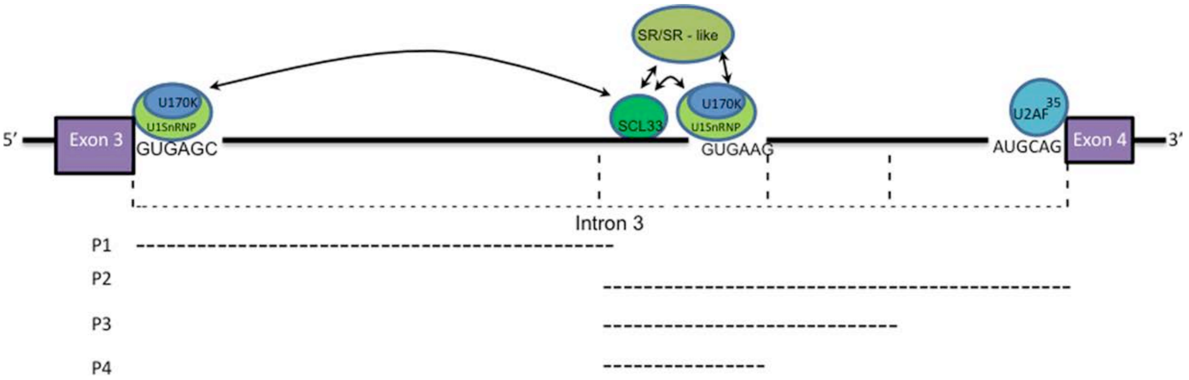


Figure 3.9. Model illustrating the role of SCL33 in regulation of its own pre-mRNA splicing. It is based on the data in this article and published reports (Golovkin and Reddy, 1996; Reddy, 2007). Boxes indicate exons 3 and 4 and line indicates intron 3. See discussion for details.

Many of the splice variants from SR genes have a premature termination codon and are targets for nonsense-mediated decay (NMD) (Lareau et al., 2007b; Palusa et al., 2007a; Palusa and Reddy, 2010). In fact, all splice variants that contain all or any part of the third intron of *SCL33* are potential targets of NMD (Palusa et al., 2007a), and some of them were experimentally shown to be degraded by NMD (Palusa and Reddy, 2010). The generation of splice variants with NMD in SRs and several other RNA-binding proteins is autoregulated so that high levels of protein result in generation of premature termination codon (PTC)-containing transcripts to tightly regulate the levels of splicing factors (Jumaa et al., 1997; Lopato et al., 1999a; Sureau et al., 2001; Lareau et al., 2007b; Schoning et al., 2008). It is likely that autoregulation of *SCL33* AS that generates PTC-containing isoforms plays a role in controlling the levels of the functional SCL33 transcript and protein. The fine balance of the PTC-isoforms with respect to the functional transcript and protein is an important feature in gene regulation (Mitrovich and Anderson, 2000; Sureau et al., 2001; Lareau et al., 2007b; Schoning et al., 2008).

Mutational studies with GAAG elements in the SCL33 binding region indicate that the number and sequence of these repeats is important for some isoforms. In humans, studies have shown that the GAAGAA hexamer, the highest scoring ESE motif, functions as an exonic splicing enhancer (Fairbrother et al., 2002) and such purine-rich elements are reported to function as ESEs in other vertebrates (Tacke and Manley, 1995; Chasin, 2007). A computational analysis of the Arabidopsis exons for candidate ESE identified GAAGAAGAA as one the ESEs (Pertea et al., 2007) and our results show that GAAG repeats can also function as intronic splicing regulators. In addition to the GAAG elements in the 92 nt fragment, there are a few other GAAG repeats on the 3' side of the Arabidopsis SCL33 intron. To see if the third intron of SCL33 from other species also contains the GAAG repeats, we aligned the nucleotide sequences of the third intron from *Arabidopsis, Brassica, Capsella and Populus*. Remarkably the 3' end of the intron where all of the AS events take place in Arabidopsis is conserved. Furthermore, most of the GAAG repeats are highly conserved across different dicots (Figure 3.10). The 3' region of the third intron of *Brachypodium*, a monocot, also contains multiple GAAG elements (Figure 3.11), suggesting conservation of this element in angiosperms. These, together with our experimental results, indicate that GAAG repeats function in regulating alternative splicing.

```
At  GTGAGCA--TGTTTT----GT------AAA-------------------T
Cr  -TGAGCA--TGTTTT----GT------AAA-------------------T
Br  GTATATACATACTTTATTTCTCTCTCTACAACTCCCTTCTTCTTCTGCTT
Pt  GTGAGCAATTACTTTAGCTGTTCATGTATGAAG-----------------
     *    *  *  ***     *      *

At  AGGAC-----AAC---------------C----AA-----CACTTGATTT
Cr  AGAAC-----ATT---------------T----AA-----CACTTGATTT
Br  AGGACTGTATAATTGTTTGGATACCTCTTAAGCAAATCTAGTCATGATGT
Pt  ---TT-----ATT---------TGC---T----GA-----TATTTCATTT
              *                       *         * ** *

At  TTAC--TTATGTTCTG--A------TTAGAAAAT----------------
Cr  TTAC--TTATGTTCTG--A------TTAGAGAAT----------------
Br  TCTCTTTTATATACTGTGAAGACGTTAAGAAAGTTTATGTATGTCATCAA
Pt  TTTC--TTACTATTTC--G------TTAAAAAAA----------------
    *  *  ***     *          * * * *

At  ----ATCC------CT--ATTTTC------------------ACAT---C
Cr  ----ATCT------CT--ATTTTT------------------TCAT---C
Br  AGCTATCTTGATAACT--ATGTTTGGTGATAGAATTCTGGCAACATGATA
Pt  ----TTCCA-----TTGGGGTTTT------------------GCTT---T
         **        *     **                    * *

At  TTCT-------------TTCTTGA------------T---------TTG-
Cr  TTCT-------------TTCTTGAA--------TC-T---------TTG-
Br  CTCTTACGTTTTTGCGGTTCATGAAAACCATACTTCTCTG-GTG--TAG-
Pt  TTCT----A--------ATCATGGT--------GTCTTTATGTTCCTTGG
    ***              ** **              *         * *

At  --TTTACTACG---------TTTATCT-------GCCTTTCCTTTTTCTT
Cr  --TTTACTACG---------TTTATCC-------CCCTTTCCTTTTTCTT
Br  --TTGACTATGACTCCAATCTTTATTCATTTTTTGCCTTTTTTT-TACTT
Pt   AGTTAATGGAG---------TCAAT-AGTCTT----GTTTGGTTTCTTTT
       ** *    *         *  **            ***  **    **

At  TCG------GTCAT---CTCTTTGTGTC----------TT------CTCA
Cr  TCT------TTCAT---CTCTGTGTGTC----------TC------CTCA
Br  TCTCTCTAATTTATCTTCCCTTTCTTTCAGGCAAGAAGAT------CTCA
Pt  TCAA--A-ATTAAT--TCCCCTTGTGAAAC--------AGATGGGACAAA
    **        * **   * *  *  *                     *  *

At  TGG----ATAC-------TAGGGCCTTATGAGA--ATCT----TCTTAAA
Cr  TG-----ATAC-------TAGGGCCTTATTAGA--ATCT----TCTTAAA
Br  GGAGGTCATTCGAGCAGTTCGGGCCTCTTAAGGACATCTATATGCCTAGG
Pt  TA-----ATGT-------TAAGTTCTTACCACA--GCTT----TCTGTTT
           **         *  *  **    *       *     *

At  A-------A----CTG---------ATTTCATAA--ATTAA---------
Cr  T-------G--CGCTG---------GTTTCATAA--ATCAA---------
Br  GATTACT-A--TACTGGGTGAGCATGTTTGATAGTTATCTATTGTTTTAT
Pt  T---TCTTTCTTCCTT---TTTCCTCTCCCTTAG--AAGTA---------
                 **           *    **   *   *

At  ATTTAGG----CT-CCATATTTTTCAGATAG--TA---------------
Cr  ATTTAGG----CT-CCATATTTTTTTCA-----TA---------------
Br  TTTTATTTTACTT-GTGTACTGATTAGATAA--TATCTCTTTTTCATCCT
Pt  ACTCAGC----TTGTCCTATTTTTTGTATTGTCTG---------------
      * *       *    ** *  *   *     *
```

```
At     ----------------------------------------------ATTT
Cr     ----------------------------------------------ATTT
Br     TGAATCTTCTAATATGCGCTTATTCCATAAACCAAAATTAGCCTCCATTT
Pt     ----------------------------------------------ATTT
                                                     ****

At     TCTG-TGACAT-GACAATAGACAG---TGAGCAT--ACTTG---TGGTCT
Cr     TCTG-TGATAT-G-AACCAT-------------C--ACTTG---TGGTCT
Br     TCTT-TGAGAT-GAGATCTT-------------C--ATTTG---AAGCCT
Pt     CCTTCCTGCACTG-CATCAGAAACTTATTAGCTTGAACTATCAATTTTTT
        **      *  *  *                    * *          *

At     TAAACC---ACTA---GTTTTCGGT--------GTCGTT----------A
Cr     TAAACC---ACTA---GTTTTTAGT--------GTCTTTCAAAGAAAGAA
Br     TAAACCCCCCCTA---GTTTTCAGTTATCCAGTGTCGAT----AAAAGTG
Pt     GAAAC---TTTTTGATGTTATCATT--------GTCATA----A--TTTT
        ****      *    *** *   *        ***

At     G--GTGGATT--AGT---CGTTTAGAA-T----TTAGTA-CC--TGGTGA
Cr     A--GTGGATT--AGT---CGTTTTGAA-T----TTAGTA-CC--TGTTGA
Br     TGTATGGATT--AGG---TTTTTTGAA-T----TTAGCA-CC--TGTTCA
Pt     TG--TGAAATTCACACAATGTTCAGCAATCTGTTTGGTTTTCTCCAAGCA
           ** * *  *       **  * * *    ** *    *       *

At     GACCTCATAGT----------TTTTAG-TCGCTTAAGGAAAAGGATACAT
Cr     GACCTCATACA----------TTTTAG-TCCCTTAAGGAGATACA---AT
Br     GAC-TTGTACT----------TTTTAG-TCGCTTAAGGAGAAGGATTC--
Pt     GGCCCTATAACCCCCCCTTCCTCCAAATTTCTTTTAGGTCATCTC---TT
       * *    **            *   *  *   ** ***  *

At     GATGATATGTTCCATGTCTGAACACACTATGAGTGAGTTAATTGAAGTTG
Cr     GATGATATGTTCCATGTCTGAACACACTATGAGTGAGTTAATTGAAGTTG
Br     -AT----TGTTGCTTGAT---ATGTTTCATGTGTGTGTTAATTGAAGTTG
Pt     AATGATATACTCC---TGTAATCATAAGCTGGGTTCCTTGATTGAAGTTC
        **    *  * *                ** **   ** *********

At     TTGAGGTTTAGCAGTGAATCTAAAGAATTGAAGACATCAAAGAAGTAATT
Cr     TTGAGGTTTAGCAGTGAATCTAAAGAATTGAAGACTTCAAA----TAATT
Br     TTGAGGTTTAGCAGTGAATCTAAAGAATTGAAGACATCA--------ATT
Pt     TTGAGATTTAACAGTGAAATTAAAGTTTTGAAGACATCAATCAAGTAATT
       ***** **** *******  *****  ******** ***        ***

At     AGAGTTCTTATGAAGATGTTCTATATGGTAGTGA-AGAATTGAAGTGAAG
Cr     AGAGATCTTATGAAGATGTTGTATATGGTAGTGA-AGAATTGAAGTGAAG
Br     AGAGTTTTTATGAAGATGTTGTGTATCGTAGTGAAAGAGTTGAAGTGAAG
Pt     AGAGTTCTTTTGAAGATGATGTTAATGATAGTGA-AGATTTGAAG-GAAG
       **** * ** ******** * *  **  ****** *** ****** ****

At     TTGAGTTTGTATTCTATGTGAAGATGAATCAA-GTCTTCAAGA------A
Cr     TTGAGTTTGTATTCAATTTGAAGATGAATCAA-GTCTTCAAGA------A
Br     TTGAGTTTTTCTTCGATTTGAAGATGAATCAAGGTCTTCAATA-------
Pt     TTGGGTTTGAATACATTGTGAAGAGATTCCAA-GTCTTGAAGAGTTTCAA
       *** ****   * *  * ******     *** ***** ** *

At     GTCATCTTTGTACTGACACTTGCAAGGCTAGCAGGCAAGTGCTTGCTTTT
Cr     GTCATCTTTATACTGAAACTTGCAAGGCTAGCAGGCAAGTGCTTGCTTTT
Br     -----------------------------AGCAGGCAAGTGCTTGCTTAT
Pt     GTCATCTTTAAACTGACACTTGCAAGGTCAGCAG---TGTGCTTGCTTT-
                                    *****    **********
```

```
At    TTCATGTTTAC----TGATATCTTCGGTTGCAGG-GGATCATATCGCCAG
Cr    TTCATGTTTAC----TGATGTCTTCTGTTGCAGG-GGATCATATAGCCAG
Br    TTCATGTTTATCCTGTGATATTGTCTGTTGCAGGGGGATCATATTGCCAA
Pt    TTCATATC------ATGA--CCTTCATATAATGG-GGATTATGTTGGTAC
      ***** *        ***     **   *   ** **** ** * *  *

At    TTGTCCTT--CAACAGGATT--GTTATAAAGTGGTACATCTTC-------
Cr    TTGTCCTT--CAACAGGATT--GTTGTAAAGTGGTACATCTTC-------
Br    CTGTCCCT--TAACAGGCTTGTGTTGTAAAGTGGTGCATCTTTTTTTTTT
Pt    ATTTCATTGGCAAAAGGCTT--ATGGTTCCCGTGCTCAGC--C-------
       * **  *   ** *** **   *  *      *  ** *

At    ------------------------------CTCGCCATTTTGTAT-GTTT
Cr    ------------------------------CTCGTCATTTTGTAT-GTTT
Br    TTTTTTTGTCAAACAAGTGGTGCATCTATGTCTGTCATTTTGTAT-GTTT
Pt    --------------------------------CCTGTTTTTGTAGCAGAT
                                           *******     *

At    GGT--TTCTAGTCTGAGAT-CTTTTTG-TTCTACATT--TTGAATGCAG
Cr    GGT--TTCTAGTCTGAGAC-CTGTTTTGTTCTACATT--TTAAATGCAG
Br    GGT--TCTTAGTCTGAGATCCTGTTTT-TTTTTCA------ATATGCAG
Pt    GGTGACACTCATCTTCAAC-TTGTTTC-TTTATCATCTTCTTTTGACAG
      ***     *  ***   *   * ***  **   **           ***
```

Figure 3.10. Alignment of the nucleotide sequence of intron 3 of SCL33 homologs from *Arabidopsis thaliana* (*At*), *Capsella rubella* (*Cr*), *Brassica rapa* (*Br*) and *Populus trichocarpa* (*Pt*).

```
GTAAGATTTTCTAAAAGCTTTTTCAGAATGTAAACGCAGCTGTGACATTA
TGTACCCATTCCATTTTGTGGCAAGGGTGTGTTCATCTTGAAGTAAGCTA 100
GTTTTGTCTGTATTTACATGTGGTTTCGAGTATCACTTGGTTTTGATCCC
ATCTCTCTCACACACTCTTCTCCCAATCGTTCCTCTATTTCTTTGGTTGC 200
GACTTTTTTGTCTGCATATAATGGGAATATCTTCTTCCCTCTTGGCCCTT
TTTAATTCATCCTCCATCATGAATGCAGGCCAGAAGACCTTCGTCGGCCA 300
TTTGGACAATTTGGCCGCCTTAAAGATGTATATATTCCAAGGGATTACTA
TACTCGGTAAGACTATCTATCTGGACTATTGTTAGTTGCCTCTACATCTG 400
GAGCTACAAGCATTTTATTAATCTAGTGGCAGTTTTATAGTTGTTATTTG
AGCTGGGTAGGATAGCATAAATTGCTTTAGCCCTATGATAATGTTACATA 500
CAGAACCACTCATTGTTATATAATGCACTTTGTTTCTCAAGTACAAAATA
GTAAAGTGCACTTTGGAGGGTAGTTTTTCTCTGTAAGGTAGTTGTTTGCG 600
CTCTAGTTTTATTGGCAGAATTAGATACTTCTGTCAATCTGTTGAAATTA
ACAGTGGAAATAATTTTTTGAAGACAAGTCTTAAGAGTACCCTTGAAGAG 700
AAGGGAAGCATAGAGAAGTTGCAAGTTAAGTAATGTGCTGTTAAAGATTA
TTGTGGAAGAGAAAGGAAGAAATAGATTTGCAAGTTAAGGGATGTGCAAG 800
TAAGGTTACTGTTGAAGAGAACAGAAGATATAGTTACAGATTTAAGAGTT
AAGAAGTGATGAAGAGTTTGAAGTCCGTATTAGAACTGTGGACAGTGCAA 900
GGTCAGCAGCCGATTGACTACATTGTTAATTTGTCTGTGTTTGCAGTCTT
TGAGAATTTGACTCTGCTTTGGTTGCTCTATATTCGGAAGAGCTTCTGTT 1000
ATATTGCCATGGACCAGAGGCTTTATCCTTTTTGAAGGAGAGATGATTTA
TTTTCTTATTTACAG 1065
```

Figure 3.11. Nucleotide sequence of SCL33 intron 3 of *Brachypodium distachyon*. GAAG repeats in the 3' region of the intron are highlighted in blue.

Since intron retention is common in plants, we analyzed for the presence of two or more GAAG repeats (with spacing of 3 to 15 nucleotides) in retained and constitutively spliced introns in the Arabidopsis genome. Among a total of 2780 introns that are retained (TAIR 10 annotations) we found that 59 retained introns have 2-12 GAAG repeats. A list of these genes along with the number of GAAG occurrences is presented in Table 3.1. This fraction (0.021) is statistically significantly higher than in constitutively spliced introns ($p < 2 .10^{-5}$ in a binomial exact test), where 0.011 of introns have such a repeat. Since retained introns tend to be shorter than constitutive introns, the *p*-value is conservative. This suggests that GAAG repeats could be one of the signals that contribute to intron retention.

Table 3.1: Introns with GAAG repeats in the Arabidopsis genome. Loci are with respect to TAIR 10 annotations.

Gene locus	Donor position	Acceptor position	Number of occurrences of GAAG
AT5G14320	4618232	4618006	2
AT2G33847	14320174	14319299	4
AT3G26180	9579656	9579255	4
AT5G16880	5550909	5550985	2
AT2G17670	7675464	7675569	2
AT2G20010	8641010	8640910	3
AT4G20260	10942948	10943124	7
AT1G48450	17909058	17908701	5
AT3G45638	16756143	16755973	2
AT4G01070	462281	461881	11
AT5G12840	4053411	4053302	2
AT3G13080	4196707	4196633	3
AT4G28260	14004914	14005002	2
AT5G51630	20970810	20970953	2
AT2G07708	3387728	3387996	4
AT4G10550	6520213	6520138	2
AT1G77180	29000274	28999969	5
AT3G46600	17158209	17158300	4
AT5G01810	311308	311442	3
AT3G10915	3417471	3417036	2
AT1G21750	7648411	7648633	7
AT3G54890	20340068	20339967	6
AT1G20920	7285435	7285559	7
AT3G13810	4544501	4544952	4
AT5G42020	16808172	16808008	7
AT4G25390	12978883	12979344	2
AT5G26760	9405963	9405754	2
AT1G14380	4919222	4919037	5
AT3G04930	1364187	1364618	3
AT5G65210	26058659	26059018	2
AT1G09840	3200181	3200099	2
AT5G55670	22545581	22545207	2
AT4G36980	17434318	17434142	3
AT1G12080	4084567	4084668	5
AT1G27370	9508145	9507023	4
AT5G28500	10478595	10478920	9
AT4G26110	13235027	13235429	3
AT1G08570	2713142	2713234	3
AT1G79000	29716936	29716775	6
AT5G03190	758660	758841	3
AT3G15095	5082226	5082608	8
AT1G79940	30072859	30073137	4
AT5G58320	23579429	23579522	6
AT3G11773	3723397	3723330	2
AT1G14170	4845224	4845126	2
AT3G29575	11383922	11383661	3
AT5G35603	13785818	13785380	3
AT5G22640	7532574	7532930	12
AT2G41430	17269402	17269718	2
AT5G07530	2382906	2382765	5
AT5G02020	387039	386721	3
AT5G21160	7199199	7198847	7
AT2G42280	17611907	17611795	2
AT1G07660	2369256	2369306	2
AT5G54600	22184027	22184083	2
AT2G45380	18701535	18701145	4
AT1G72510	27303520	27303737	2
AT1G17780	6124157	6124252	2
AT4G12850	7537203	7537281	2

The protoplast system with splicing reporters together with putative splicing factor mutants, as described here, can be employed to identify splicing regulators involved in AS. This, combined with *in vitro* RNA binding studies can provide further insights into direct or indirect regulation of AS of a given pre-mRNA. Although transgenic lines overexpressing a splicing factor have been used to study the role of splicing regulators (Wang et al., 1996; Lopato et al., 1999a; Kalyna et al., 2003), results from such experiments may not provide accurate insights since the effects of many splicing regulators are dosage dependent. The use of protoplasts from knockout mutants offers an alternative approach and has an advantage in that the cells lack one or more splicing regulators. As several SRs have paralogs, the functional redundancy can also be addressed by using protoplasts from double or triple mutants as has been demonstrated here.

MATERIAL AND METHODS

Construction of a splicing reporter and generation of a transgenic line expressing the splicing reporter

Genomic DNA was isolated from *Arabidopsis thaliana* Columbia (Col-0) using a Plant DNAeasy kit (Qiagen, USA) and used as a template in PCR. The third intron of the *SCL33* gene was amplified with Hot Start Pfu polymerase using intron specific primers (SR33-IN-FP and SR33-IN-RP, see Table 3.2) and cloned into pGFP(GA5)II at the *MscI* site within the coding region of GFP to generate the *GFP-SCL33 INTRON-GFP* (*GFP-INT-GFP*) construct driven by the *CaMV35S* promoter. Correct orientation of the intron was verified by sequencing. The *GFP-INT-GFP* and the control *GFP* plasmids were used to transform leaf protoplasts. To generate *GFP-INT-GFP* stable lines, the *GFP-INT-GFP* region was isolated from the transient expression vector and cloned into *Sac*I-*Xho*I sites in a binary vector (pBA002) and used to transform

Arabidopsis. Transgenic lines were selected on BASTA (5 µg/ml) and the F2 plants expressing *GFP* or *GFP-INT-GFP* were used to monitor splicing.

Table 3.2. Primers used for genotyping SR mutants, for amplifying splice variants from the splicing reporter and for generating *SR33* intron RNA probes. Restriction enzyme sequences are underlined.

Name	**Sequence**	**Restriction Enzyme**
Genotyping		
SCL33F'	5'-GGTAGATCTCGGTCACGGAG-3'	
SCL33R	5'-GTTCCCCACATGTTCC-3'	
SC35F'	5'-ATCGCTGCTGAACCGATACGAAC-3'	
SC35R’	5'-CTCCTACGAGGACTGCGGCTTC-3'	
SCL30aF	5'-CATGATTGCAGGCAAGAAGA-3'	
SCL30aR'	5'-CCAGTAGTAATCCCTAGGA-3'	
LBb1	5'-GCGTGGACCGCTTGCTGCAACT-3'	
RT-PCR		
SCL33F	5'-CTCCGTCGTTCCTCACCACCG-3'	
SCL33R	5'-GTTCCCCACATGTTCC-3'	
SCL30aF	5'-CATGATTGCAGGCAAGAAGA-3'	
SCL30aR	5'-CTTTGGCTCCTTGCTTGTTC-3'	
Clones for RNA probes		
SR33Intron_P1F	5’-CGCGGATCCGTGAGCATGTTTTGTAAATA-3’	*Bam*HI
SR33Intron_P1R	5’-CCCAAGCTTCAATTAACTCACTCATAGTG-3’	*Hind*III
SR33Intron_P2F	5’-CGCGGATCCAAGTTGTTGAGGTTTAGCAG-3’	*Bam*HI
SR33Intron_P2R	5’-CCCAAGCTTCTGCATTCAAAATGTAGAAC-3’	*Hind*III
SR33Intron_ P3F	5’-CGCGGATCCAAGTTGTTGAGGTTTAGCAG-3’	*Bam*HI
SR33Intron_ P3R	5’-CCCAAGCTTTAAACATGAAAAAAGCAAGC-3’	*Hind*III
SR33Intron_ P4F	5’-CGCGGATCCAAGTTGTTGAGGTTTAGCAG -3’	*Bam*HI
SR33Intron_ P4R	5’-CCCAAGCTTCTTCACTACCATATAGAACA-3’	*Hind*III
For analysis of splice variants from the splicing reporter and from the native gene		
pGFP-F	5'-AATTCTTGTTGAATTAGATGGTGATG-3'	
pGFP-R	5'-GACTTCAGCACGTGTCTTGTAG-3'	
SCL33-ex3F	5’-CGTTTGAGCAGTTTGGTCCT-3’	
SCL33-ex4R	5’-GCCTCCCTTGCTCTCATTTCA-3’	

Generation of single and double mutants of Arabidopsis

The Arabidopsis T-DNA insertion lines for the genes *SCL33* (Salk_058566), *SC35* (Salk_033824), *SCl30a* (Salk_041849) in Columbia background were obtained from the Arabidopsis Biological Resource Center. The T-DNA insertion in each gene was verified by genomic PCR using a gene-specific primer and T-DNA specific primer (LBb1). Expression of the *SR* genes in the mutants was analyzed by RT-PCR using gene specific primers (Table 3.2). The following PCR conditions were used for genotyping: The initial denaturation at 94^0C for 2 min, followed by 29 cycles at 94^0C for 30 sec, 56^0C for 30 sec and 72^0C for 1 min. The final extension at 72^0C for 10 min. DNase-treated RNA from two-week-old seedlings from wild type and mutant lines was used for RT-PCR analysis as described earlier (Palusa et al., 2007a). Wild type and all mutant lines were grown under long-day conditions (16 hrs light and 8hrs dark; 100 μmol/m^2/s^2 light intensity, 22^0C). Three double mutants (*sc35 scl30a*, *scl33 sc35*, and *scl33 scl30a*) were generated by crossing the single mutants. All double mutants were genotyped for T-DNA insertion and homozygosity using genomic PCR and RT-PCR as described above.

Analysis of *GFP-INT-GFP* splicing in Arabidopsis mesophyll protoplasts

Splicing of *GFP-INT-GF*P pre-mRNA was analyzed in mesophyll protoplasts obtained from the leaves of Arabidopsis wild type, single (*scl33*, *sc35* and *scl30a*), and double (*sc35 scl30a*, *scl33 sc35* and *scl33 scl30a*) mutants. Protoplasts from rosette leaves of 3 to 4-week-old plants grown in a greenhouse at 22°C under (16 hrs light and 8 hrs dark) were prepared and transfected as described earlier (Yoo et al., 2007). Equal amounts (20 µg/ml) of *GFP* or *GFP-INT-GFP* plasmid were used to transfect 2 mL ($2x10^6$ protoplasts) of protoplasts from wild type, three single, and three double mutants. Transfected protoplasts were incubated in Petri dishes in the growth chamber in dark at 22^0C for 15 to 16 hrs. The protoplasts were then visualized under the fluorescent microscope for GFP expression and RT-PCR analysis.

RNA isolation and RT-PCR analysis

RNA from the transfected protoplasts, wild type and mutant plants was isolated using RNAeasy plant mini kit (Qiagen, USA). On-column DNAase (Qiagen, USA) digestion was performed to remove any genomic or plasmid DNA contamination before cDNA synthesis. DNAse treated RNA (200ng) was used to synthesize first-strand cDNA using Superscript II reverse transcriptase (Invitrogen, USA) and 2 μl of the first-strand cDNA was used for PCR in a reaction volume of 20 μl. The *GFP* specific forward and reverse primers (Table 3.2) were used for amplification. To monitor the levels of *SCL33* splice variants generated from the endogenous gene, *SCL33* F & R primers were used (Table 3.2). All splice variants generated from the *SCL33* 3rd intron of the endogenous gene in protoplasts were amplified using primers corresponding to exon 3 and 4 (Table 3.2). The PCR products were gel purified, cloned into PCR2.1 TOPO vector and sequenced.

Preparation of ^{32}P-labeled RNA probes and cold competitor RNAs

Intron 3 of *SCL33* was divided into four parts (P1, P2, P3 and P4) and cloned into pGEM4 vector (Promega, USA) using the *Bam*H1 and *Hin*dIII restriction sites. P1 consists of the first 421 nucleotides (nt), P2 the remaining 344 nt fragment of the third intron, P3 the first 208 nt of P2 and P4 the first 92 nt of P2. All fragments were generated by PCR amplification using the primers listed in Table 3.2, and the clones were verified by sequencing. Each of these constructs were linearized by digesting with *Hin*dIII and used as a template to prepare labelled P1, P2, P3 and P4 RNA probes as follows. Capped RNAs were transcribed *in vitro* and labeled with 45 μCi of [α-^{32}P] UTP (800 Ci/mmol, Perkin-Elmer, USA) using SP6 RNA polymerase (Fermentas, USA) in the presence of 500 μM ATP, 500 μM CTP, 50 μM GTP, 50 μM UTP and 7mGpppG from linearized pGEM4 plasmid DNA templates and gel purified as previously described (Wilusz and Shenk, 1988). Unlabeled competitor RNAs were generated in the same manner, but without 7mGpppG, radiolabeled nucleotide, and the concentrations of UTP and GTP were increased to 500 μM.

Expression and purification of recombinant SCL33 and SR45 proteins

The *SCL33* clone in pET32 expression vector was used to prepare purified SCL33 protein (Golovkin and Reddy, 1999) with minor modifications. The bacteria were grown at 37^0C until OD_{600} 0.6 after which 0.5 mM isopropyl-B-D- thiogalactopyranoside (IPTG) was added and the culture was incubated for 4 hrs at 30^0C to induce protein expression. Subsequently, the bacteria were centrifuged, the pellet was resuspended in 1/10 of the culture volume of binding buffer [50 mM Tris-HCL (pH 8.0), 2 mM EDTA, 100 μg/ml lysozyme and 0.1% of Triton X-100] containing protease inhibitors and incubated at 4^0C for 15 min. The sample was then sonicated, centrifuged, and the supernatant was collected. S-protein agarose beads were added to the supernatant and the mixture was incubated for 1 hr at 4^0C. After washing the beads with binding buffer, the bound protein was eluted with 0.2 M citrate buffer (pH 2) and neutralized by adding 1/20th volume of 2 M Trisbase (pH 10.4). The eluted proteins were dialyzed using the phosphate buffer (10 mM Na_2HP0_4, 2 mM KH_2PO_4, 2.7 mM KCl, 137mM NaCl pH 7.4). SR45 was purified as described earlier (Golovkin and Reddy, 1999).

Electrophoretic mobility shift assays

Four to twenty fmols of indicated internally radiolabeled RNAs (P1, P2, P3 and P4) were incubated with increasing amounts of purified recombinant SCL33 in the presence of 20 units of RNAse inhibitor (Invitrogen, USA) 0.15 mM spermidine and gel shift buffer [15 mM HEPES (pH 7.9), 8% glycerol, 100 mM KCl and 2 mM $MgCl_2$] for 5 minutes at 30^0 C in a 14 µl reaction. Following incubation, 4 µg/µl of heparin sulfate (Sigma, USA) was added to each reaction, the samples were then chilled on ice for 5 minutes and 1.5 µl of 10X loading dye (30% glycerol, 0.5% bromophenol blue, 0.5% xylene cyanol) was added. RNA-protein complexes were run on a 5% native polyacrylamide gel at room temperature in 1X TBE buffer (200 Volts for 2-6 hrs). Gels were then dried, exposed to a phosphor screen, and visualized by Phosphor Imaging using Storm 840 (Molecular Dynamics, USA).

Generation of mutations in the *SCL33* Intron

The shortest fragment (92 nt) of the *SCL33* intron that bound to SCL33 protein contains four conserved GAAG elements. To test if these sequences are important for AS we mutated them, two at a time or all four, to CTTC using QuickChange Lightning site-directed mutagenesis kit (Stratagene, USA) using the *GFP-INT-GFP* construct in pGFP(GA5)II vector as a template. The mutants were confirmed by sequencing. The first two mutated GAAG elements are represented as M1&2, the last two as M3&4 and the mutation in all four elements is designated as M1-4. Protoplasts from wild type and *SR* mutants were transfected with the mutated introns and RNA was extracted for splicing analysis using GFP-specific primers.

Chapter 4

GLOBAL ANALYSIS OF GENE EXPRESSION AND ALTERNATIVE SPLICING IN A TRIPLE MUTANT OF ARABIDOPSIS THAT LACKS THREE SERINE/ARGININE (SR)-RICH SPLICEOSOMAL PROTEINS USING RNA-SEQ

SUMMARY

In animals, the serine/arginine-rich (SR) family of proteins is implicated in regulating gene expression at multiple levels. Auto- and cross-regulation of splicing of several pre-mRNAs was shown by overexpression of some plant SR proteins. However, studies to investigate global changes in gene expression and splicing in any loss-of-function mutants of plant SR proteins have not been carried out. Here we performed a transcriptome analysis in a triple mutant lacking three SR genes (*SC35*, *SCL30a* and *SCL33*) using next generation sequencing to monitor transcriptome-wide changes. About 80 million reads (40 M from WT and 40 M from mutant) from two-week old seedlings of wild type and the triple mutant were obtained and analyzed for changes in gene expression and pre-mRNA splicing. Analyses of this RNA-Seq data show that loss of SC35, SCL30a and SCL33 results in significant changes in expression levels of protein-coding and non-protein coding genes (miRNAs and other non-coding RNAs) and splicing patterns of many genes. Expression of 737 genes including 5 miRNAs was altered in the mutant, of which 351 are up-regulated whereas 386 are down-regulated. Our analysis also identified 13 novel transcriptional units. In addition, splicing patterns of several genes are altered in the mutant. There is also considerable overlap between differentially expressed and differentially spliced genes in the mutant. Among differentially spliced genes both qualitative and quantitative

changes in splicing were observed in the mutant. We validated over 30 genes that are either differentially expressed or spliced using RT-PCR. Grouping of all differentially expressed genes using gene ontology (GO) terms has revealed that genes involved in plant immunity and in iron and phosphorous homeostasis and stress responses especially in are overrepresented. The set of differentially expressed/spliced genes that we identified here represents direct and indirect targets of these SR proteins. Identification of direct targets of each of the SRs using methods such as PAR-CLIP will help us not only to identify which of these are indirect targets but also pave the way to use computational tools to identify potential splicing regulatory elements in direct targets.

INTRODUCTION

Alternative splicing (AS) of precursor-mRNAs (pre-mRNAs) is an important process in generating transcriptome and proteome diversity and it adds a new level of regulation of gene expression. Alternative splicing can generate multiple transcripts from a single gene that could encode functionally distinct proteins or regulate the levels of functional transcripts by modulating RNA stability (Reddy, 2007; Kalyna et al., 2012; Syed et al., 2012) and hence can have subtle or opposing functional consequences in regulating biological processes. With the advent of high throughput next generations sequencing technologies the frequency of AS documented in various organisms has increased tremendously, with about 95% of pre-mRNAs from multiexon genes in humans undergoing AS whereas in Drosophila 60.7% of the 12,295 expressed multi-exon genes also display AS (Pan et al., 2008; Graveley et al., 2011). The frequency of AS in Arabidopsis has risen from 1.2% in 2003 based on EST/cDNAs to 61% with RNA-Seq data using the next generation sequencing (NGS) methods (Marquez et al., 2012; Syed et al., 2012). In rice, RNA-Seq studies have shown that about 50% of intron-containing genes are alternatively spliced (Lu et al., 2010). Alternative splicing is a complex process that is coupled to transcription and epigenetic modifications (Das et al., 2007; Lyko et al., 2010; Schor et al., 2010; Reddy et al., 2012b).

AS can alter the intracellular localization of proteins by altering localization signals, change sequences for post-translational modification or interaction sites with other proteins, and enzymatic activity (Rao et al., 2005; Wang et al., 2005). Given the extensive functions of AS, it is not surprising that aberrant regulation of AS leads to many human diseases (Kelemen et al., 2012). AS contributes to species-specific differences in gene expression and function, with also disparities among individuals of the same species (Hull et al., 2007; Mola et al., 2007). In plants, alternative splicing of pre-mRNAs is implicated in a range of plant functions, including growth and development, disease resistance, signal transduction, biotic and abiotic stress responses, flowering time and the circadian clock (Reddy, 2007; Ali and Reddy, 2008b; Syed et al., 2012). The circadian clock regulates alternative splicing in such a manner that circadian transcripts are synchronized with certain tissue types and environmental cues (Staiger and Green, 2011; Filichkin and Mockler, 2012; James et al., 2012; McGlincy et al., 2012).

The process of constitutive splicing (CS) and AS requires assembly of spliceosomes on the pre-mRNA and this involves numerous RNA-RNA, RNA-protein and protein-protein interactions. The short splicing signals in pre-mRNAs at 5'(GU) and 3'(AG) splice sites are quite conserved between plants and animals and are necessary for splice site selection, but two other signals, the branch point and the polypyrimidine tract, are not as well-defined in plants as in vertebrates (Chasin, 2007; Schuler, 2008; Reddy et al., 2012b)}. In animals, in addition to these four core signals, numerous loosely conserved splicing regulatory elements (SREs) in the intronic and exonic regions have been identified, and these in coordination with trans-acting proteins influence splice site selection for both CS and AS (Chasin, 2007; Barash et al., 2010; Reddy et al., 2012b). Very few SREs have been identified experimentally in plants (Day et al., 2012; Reddy et al., 2012b; Thomas et al., 2012). The differences, in intron length and composition and in some core splice signals between plants and animals suggest that some regulatory mechanisms involved in pre-mRNA splicing may be unique to plants (Reddy, 2004; Reddy, 2007). The size of introns seems to be a determining factor in the prevalence of certain types of AS events. It appears that smaller sizes of introns may lead to intron retention, which is more prevalent in plants than animals (Reddy, 2007; Graveley et al., 2011; Marquez et al., 2012).

Excision of introns and joining of exons in pre-mRNAs takes place in the spliceosome, a huge RNP complex consisting of a family of five small nuclear RNAs and about 180 proteins (snRNP and non-snRNP) (Wahl et al., 2009). The major spliceosome (U2-type) removes introns that have canonical GT-AG splice sites (Wang and Brendel, 2004; Reddy, 2007; Ru et al., 2008; Reddy et al., 2012a). Another class of “non-canonical” introns with AT-AC boundaries is excised by the minor spliceosome (U12-type) (Shukla and Padgett, 1999). The non-snRNP proteins, mainly serine-arginine (SR)-rich proteins, SR protein kinases and hnRNP proteins modulate snRNPs interaction with pre-mRNAs and hence regulate splicing in a concentration dependent manner (Long and Caceres, 2009; Reddy and Ali, 2011). These proteins have a modular structure with one or more RNA-binding domains and protein-protein interaction domains. In animals SRs have been shown to recruit U1 snRNP to the 5’splice site, U2AF to the 3 ‘splice site, or U2snRNP to the branch point, and also bind splicing regulatory elements (SREs) for both CS and AS (Reddy, 2007; Long and Caceres, 2009: Lam, 2002 #13574). In plants also, several SR proteins interact with components of snRNPs (Golovkin and Reddy, 1996, 1998, 1999; Lopato et al., 2002; Lopato et al., 2006; Reddy, 2007; Day et al., 2012). Extensive studies with animals have shown that SR proteins in combination with other splicing factors play a major role not only in CS and AS but numerous other processes like mRNA export, RNA stability, nonsense mediated decay (NMD), mRNA surveillance, and also as carbohydrate binding proteins on the cell surface (Bourgeois et al., 2004; Hatakeyama et al., 2009; Long and Caceres, 2009; Twyffels et al., 2011). The analysis of splicing regulation in plants has been thwarted due to lack of a plant in vitro splicing assay and also because plant introns are not processed accurately in the widely used mammalian *in vitro* splicing system (McCullough et al., 1991; Reddy, 2001b; Schuler, 2008).

Pre-mRNAs of plant SR genes display high levels of AS. It was experimentally shown in Arabidopsis that there is a six-fold increase in the SR gene transcriptome (14 SR genes generate 93 distinct AS isoforms) and that the splicing patterns of these are significantly altered during development and in response to various stresses and hormones (Palusa et al., 2007a). The prevalence of AS of SR pre-mRNAs in plants is much higher than what was observed with mammalian SR genes (Lareau et al., 2007b). Some splice variants of *SR45a* pre-mRNA, which encodes an SR-like protein, are increased in abundance with heat and drought stress (Gulledge et al., 2012). Global studies on AS of plant pre-mRNAs have also indicated that stresses or changes in environmental conditions are major factors in altering splicing patterns of many other genes. GO (Gene Ontology) analysis of AS genes revealed that the majority of genes with AS are associated with biotic or abiotic stresses (Iida et al., 2004; Iida and Go, 2006; Wang and Brendel, 2006a; Filichkin et al., 2010; Gan et al., 2011).

Interestingly many of the SR splice variants (about 53 representing over half of SR splice variants) in Arabidopsis contain a premature termination codon (PTC) due to intron retention.

By analyzing the levels of SR splice variants in a mutant (*upf3*) in which nonsense-mediated decay (NMD) is impaired, it was shown that many of the PTC-containing transcripts are the candidates for NMD (Palusa and Reddy, 2010). Since SRs undergo AS producing multiple transcripts with potentially different functional roles, they can cause changes in the amount of AS and expression of other multi-exon pre-mRNAs from SRs and other genes. Overexpression of Arabidopsis SR30 resulted in morphological and developmental phenotypes including late flowering and changed the splicing of other SR genes (Lopato et al., 1999b). An increase in levels of atRSZ33 protein levels caused severe pleiotropic changes in plant development resulting from increased cell expansion and changed polarization of cell elongation (Kalyna et al., 2003). Recently, single, double and triple mutants of the *SCL33, SCL30a* and *SC35* showed significant flowering time phenotypes (Thomas and Reddy, 2012). Overexpression of a rice SR (RSZ36) gene caused changes in splicing of its own pre-mRNA (Isshiki et al., 2006). Autoregulation of AS by SRs is quite well established both in plants and animals (Jumaa and Nielsen, 1997; Lareau et al., 2007b; Dreumont et al., 2010; Thomas et al., 2012). In Arabidopsis three SR proteins (SR30, RSZ33 and SCL33) have been shown to regulate splicing of their own pre-mRNAs (Lopato et al., 1999b; Kalyna et al., 2003; Thomas et al., 2012).

The roles of SRs in animals have been studied quite extensively primarily using *in vitro* splicing assays, but similar studies have not been performed in plants because of the lack of a plant-derived in vitro splicing system. Hence, efforts are focused on using alternate *in vivo* approaches to address the roles of plant SRs. Arabidopsis serves as an excellent model for in-vivo studies to address the role of plant SRs because of the availability of mutant resources, and well-studied biological pathways. The well-annotated Arabidopsis gene models help in the analysis of AS using bioinformatics tools, so that changes in AS due to loss of one or more SR proteins can be precisely determined using a candidate gene approach as well as more global analysis using the high throughput next generation sequencing technologies. Although humans are considered to be the most complex among multicellular organisms and have a very high percentage of AS (found in 95% of intron-containing genes), they contain only 10 SR proteins. However, plants possess the largest number of SR proteins among all multicellular organisms, with 18 SRs in Arabidopsis, 22 in rice and 25 in soybean (Richardson et al., 2011). In Arabidopsis the SR proteins are divided into six major sub-families with several plant-specific ones (Barta et al., 2010). All subfamilies except one have two or more paralogs and this raises the question of whether these have different functions or are redundant in their functions. The SCL subfamily in Arabidopsis has 4 genes (*SCL28, SCL30, SCL30a, SCL33*) and this family is similiar to SC35 subfamily, which has one SR protein in Arabidopsis. Recent studies show that plant SRs perform both unique and redundant roles depending on the process. For instance, analyses of mutants of *SCL33* and *SCL30a* showed functional overlap of these two protein in regulating splicing of *SCL33* pre-mRNA (Thomas et al., 2012). However, phenotypic analysis of *SCL33* and *SCL30a* mutants revealed unique and opposing roles in regulating flowering time in Arabidopsis (Thomas and Reddy, 2012).

Global transcriptome (RNA-Seq) studies using next generation sequencing (NGS) methods have been extremely useful in uncovering changes in gene expression and AS, and also in identifying novel transcripts and splice variants and non-coding RNAs in animals and plants (Pan et al., 2008; Wang et al., 2008; Fox et al., 2009; Filichkin et al., 2010; Trapnell et al., 2010; Zhang et al., 2010; Ghazalpour et al., 2011; Graveley et al., 2011; Marquez et al., 2012; Reddy et al., 2012b). RNA-Seq is based on the idea that read counts for each transcript from the NGS reflect relative transcript abundance, and this relative quantification is reproducible and highly accurate (Martin and Wang, 2011; Reddy et al., 2012b; Trapnell et al., 2012). In recent years one of the NGS platforms (Illumina) has been extensively used in sequencing of short fragments of cDNA in large numbers, in tens of millions. These reads are then aligned to a reference genome to quantify transcript isoform frequency using a variety of newly developed bioinformatics tools (Trapnell et al., 2010; Wang et al., 2010; Rogers et al., 2012; Trapnell et al., 2012).   In an early study of the human transcriptome (Wang et al., 2008) analysis of 15 tissue types using Illumina reads revealed 92-94% AS. In another study with the Drosophila transcriptome, RNA-Seq, tiling microarrays, and cDNA sequencing was used to analyze the transcriptome of 30 developmental stages (Graveley et al., 2011). The aligned reads were quantified by Cufflinks (Trapnell et al., 2010) and AS products were identified by JuncBASE (Brooks et al., 2011). In addition to discovery of 1938 novel transcribed regions, small non-coding RNAs, and primary miRNAs, about 61% of the 12,295 multiple exon genes showed AS. Among plants, Arabidopsis was first systematically analyzed for AS by NGS using multiple tissues sampled over different time points and stresses. This study reported an AS in 42% of intron-containing genes (Filichkin et al., 2010). The Arabidopsis work was soon followed by RNA-Seq analysis in rice showing 48% of genes with AS (Lu et al., 2010; Zhang et al., 2010). In perennial orange, comparison of an early flowering mutant to wild-type revealed AS in flowering time genes including three AS products for *FY,* a gene involved in promoting flowering, in the mutant (Ai et al., 2012). While most of the bioinformatics tools for RNA-Seq data analysis focus on addressing quantitative gene expression (Trapnell et al., 2010), newer tools such as Cuffdiff and SpliceGrapher (Rogers et al., 2012; Trapnell et al., 2012) have been developed and are being refined for identification and quantification of AS events.

Some of the important unanswered questions about plant SRs are to what extent each SR protein controls AS globally, what AS events are unique to a given SR and how much functional overlap there is between different SRs, especially among the members of the same subfamily, in their pre-mRNA targets. Using the available loss-of-function mutants that lack one or more SRs that are generated in this study for global transcriptome studies one can begin to address some of these questions. Global transcriptome studies (RNA-Seq) coupled with global identification of RNA targets for each SR protein using powerful approaches such as HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) and PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced CrossLinking and ImmunoPrecipitation) (Licatalosi et al., 2008; Darnell, 2010; Hafner et al., 2010a; Xiao et al., 2012), it is possible to identify direct and indirect pre-mRNA targets of each SR protein and identify potential *cis*-elements in RNA that bind SR proteins. Integration of RNA-Seq and PAR-CLIP results will not only provide new insights into the roles of SR proteins but also aid in understanding regulatory networks of AS and eventually contribute to deciphering the splicing code in plants (Barash et al., 2010; Reddy et al., 2012b). Towards this goal we have initiated global RNA-Seq studies with all single, double and triple *SR* mutants that we have generated. RNA-Seq analysis with double and triple *SR* mutants will permit us to analyze combinatorial effects of the lack of multiple SRs in different combinations in regulating gene expression and AS.

The objective of the work presented in this chapter is to find the global alteration in gene expression and large-scale changes in AS, both qualitative and quantitative variation in isoforms, of various genes in a triple mutant (*sc35, scl33* and s*cl30a*) of SR proteins. We selected *SC35*, a unique member of the SC35 subfamily and an ortholog of mammalian SC35 the lack of which results in embryo lethality, and two plant-specific *SR*s (SCL33 and SCL30a) that are related to

SC35 because they have almost similar RRM domains. Since SR genes regulate their own splicing and subsequently affect AS of other genes, a triple mutant seedlings is expected to reveal significant global changes in gene expression and AS. RNA-Seq data from wild type and triple mutant were analyzed using a suite of bioinformatics tools (Rogers et al., 2012; Trapnell et al., 2012) to uncover changes in gene expression and AS in the mutant. Results of this analysis are presented below.

RESULTS AND DISCUSSION

Analysis of the transcriptome, which allows detection and quantification of all transcripts in a given cell, tissue or organism, is critical to understand the relationship between genotype and phenotype as well as in quantifying changes in gene expression during development and in response to changes in environmental conditions. With recent high-throughput NGS technologies it is possible to quantify transcriptomes including rare transcripts rapidly and at a reasonable cost. More recently, RNA-Seq technology is increasingly used to identify alternative splicing, novel transcripts, rare non-coding transcripts and trans-splicing events. Since SRs are key regulators of splicing, loss of an SR protein is expected to result in changes in expression of many genes. In order to understand splicing regulatory networks, it is necessary to investigate changes in gene expression and splicing in mutants that lack one or more SR. By performing such studies with all SRs one can identify unique and redundant roles of SR. However, in the case of plant SR proteins no RNA-Seq studies have been performed so far. To address this, we performed RNA-Seq analysis with SR mutant of Arabidopsis.

Quality of RNA-Seq reads from WT and TM

A triple mutant (TM) of Arabidopsis *SR* genes (*sc35 scl33 scl30a*) was generated by making crosses as described earlier in Chapter 2. RNA from two week-old seedlings of WT and TM grown under long day conditions were used for RNA-Seq analysis. For each genotype two biological replicates were performed. Genotypes of seedlings that were used for RNA-Seq were confirmed by genomic PCR and RT-PCR (Figure 4.1).

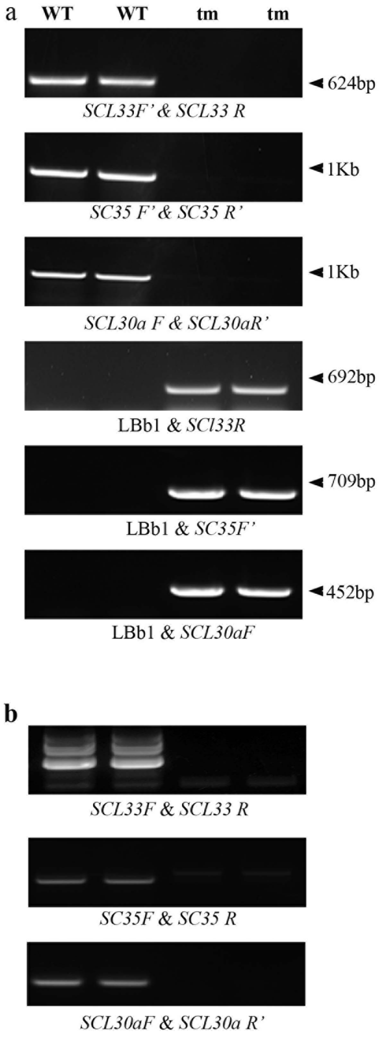
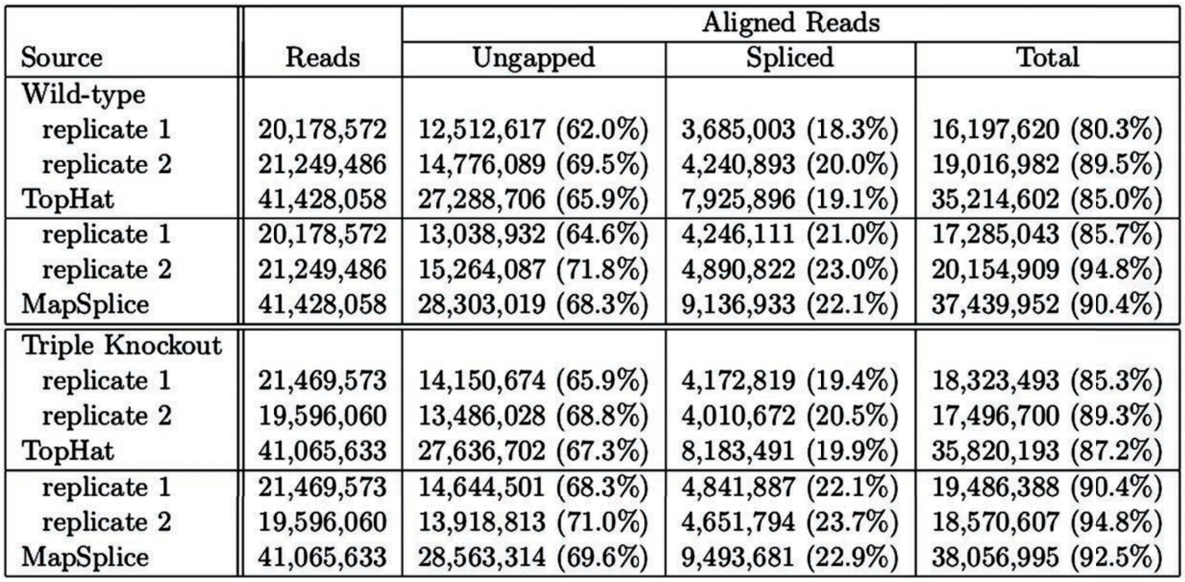


Figure 4.1: Genotypic characterization of triple knockout mutant (tm) (*scl33 sc35 scl30a)* using genomic PCR and RT-PCR. (a) Wild-type (wt) and tm plant DNA samples screened by PCR using gene primer combinations (F and R refer to forward and reverse primers) as shown below in the first three panels. Bottom three panels show PCR using the T-DNA left border LBb1 primer and gene specific primers. **(b)** RT-PCR analysis of expression of all three genes in wt and tm lines using the gene primers as shown below.

The quality of RNA, as judged by the RNA integrity (RIN) value, was found to be high (8 on scale of 10). Poly (A) RNA was isolated from total RNA, randomly sheared to about 200 nt and cDNA was synthesized using random primers. Unlike synthesizing cDNA first with oligo d(T) primer and shearing cDNA, the method we used in our study eliminates bias in reads toward the 3’ end of the transcript. After adding adapters to cDNA, single end reads of 75 nt were generated on the Illumina GAIIx platform. An equal number of reads (about 20 million) for each replicate of WT and TM were obtained (Table 1). To assess the quality of sequence reads, a FastQC was performed. The quality scores of reads from all samples were high and ranged between 34-40 (40 being the highest score) (Figure 4.2). The distribution of the read depth for the genes in WT and TM is very similar (Figure 4.3), suggestive that there is no sequencing bias towards specific groups of genes in either TM and WT.

Table 4.1: Read alignment statistics for the two WT- type replicates (top) and the tm replicates (bottom). Reads were aligned to the TAIR 10 version of the Arabidopsis genome using TopHat and MapSplice. False positive spliced alignments were removed using SpliceGrapher.



		Aligned Reads		
Source	Reads	Ungapped	Spliced	Total
Wild-type				
replicate 1	20,178,572	12,512,617 (62.0%)	3,685,003 (18.3%)	16,197,620 (80.3%)
replicate 2	21,249,486	14,776,089 (69.5%)	4,240,893 (20.0%)	19,016,982 (89.5%)
TopHat	41,428,058	27,288,706 (65.9%)	7,925,896 (19.1%)	35,214,602 (85.0%)
replicate 1	20,178,572	13,038,932 (64.6%)	4,246,111 (21.0%)	17,285,043 (85.7%)
replicate 2	21,249,486	15,264,087 (71.8%)	4,890,822 (23.0%)	20,154,909 (94.8%)
MapSplice	41,428,058	28,303,019 (68.3%)	9,136,933 (22.1%)	37,439,952 (90.4%)
Triple Knockout				
replicate 1	21,469,573	14,150,674 (65.9%)	4,172,819 (19.4%)	18,323,493 (85.3%)
replicate 2	19,596,060	13,486,028 (68.8%)	4,010,672 (20.5%)	17,496,700 (89.3%)
TopHat	41,065,633	27,636,702 (67.3%)	8,183,491 (19.9%)	35,820,193 (87.2%)
replicate 1	21,469,573	14,644,501 (68.3%)	4,841,887 (22.1%)	19,486,388 (90.4%)
replicate 2	19,596,060	13,918,813 (71.0%)	4,651,794 (23.7%)	18,570,607 (94.8%)
MapSplice	41,065,633	28,563,314 (69.6%)	9,493,681 (22.9%)	38,056,995 (92.5%)

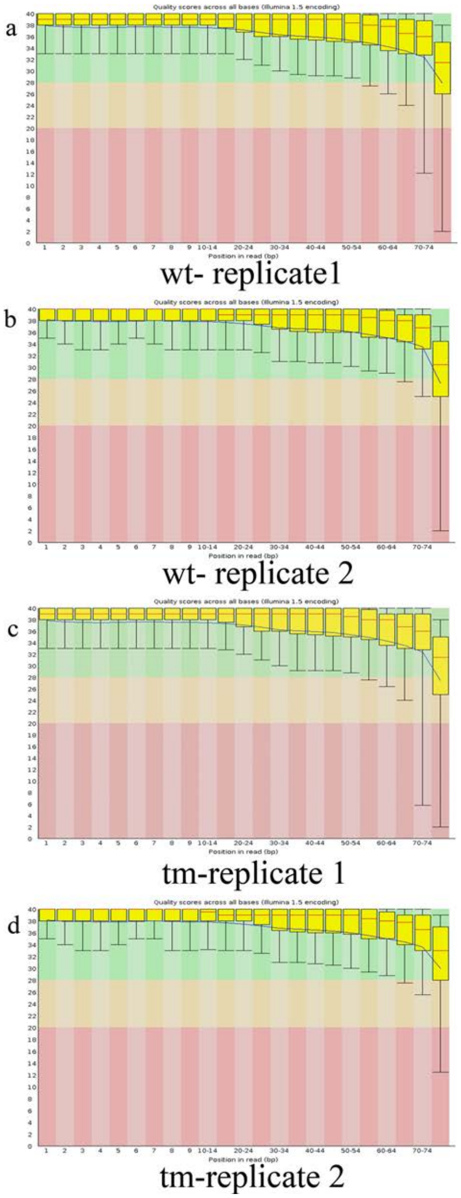


Figure 4.2: Per base quality assessment using FastQC of RNA-Seq data of WT and tm. The 4 graphs for the samples are illustrated by Box-Whisker plots. The elements of the plot are: central red line is the median value, yellow box represents the inter-quartile range (25-75%), upper and lower whiskers represent the 10% and 90% points. The y-axis on the graph shows the quality scores, with higher scores having better base call. The background of the graph divides the y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). In Figure 4.2 a-d most of the reads are distributed in the green area.

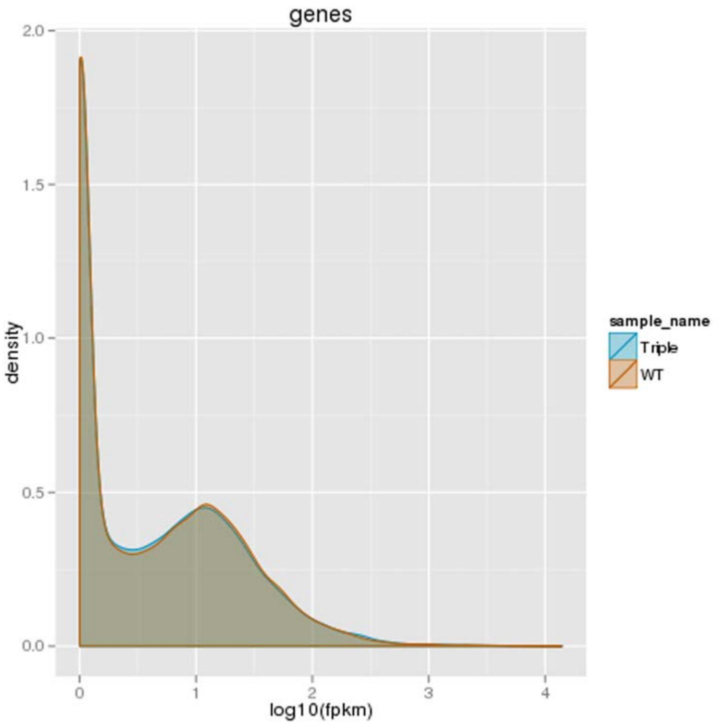


Figure 4.3: Density distribution of the genes analyzed by RNA-Seq.
Genes are shown with variation in expression levels, measured by log10 (fpkm) for tm and WT. For a majority of the genes, their expression levels (log10(fpkm)) are from 0.5 to 2 (FPKM around 3 to 100) centered around 1 (or 10 as FPKM) (the 2nd broad peak). The genes, which were either not expressed or at a very low expression level, form the very first thin peak. The genes from tm and WT showed a very similar distribution, suggesting no sequencing bias for the two samples.

To assess the uniformity of read coverage within each data set, we combined the data for the replicates in each genotype and measured the combined median read coverage within 1000-nt windows across each chromosome. We compared the read coverage in the wild-type (in blue) and the mutant (red) data sets and identified windows where read coverage was present in one data set but not the other (Figure 4.4). The reads were distributed throughout the length of five chromosomes except for the centromeric regions, which are known to be not very active transcriptionally (Figure 4.4).

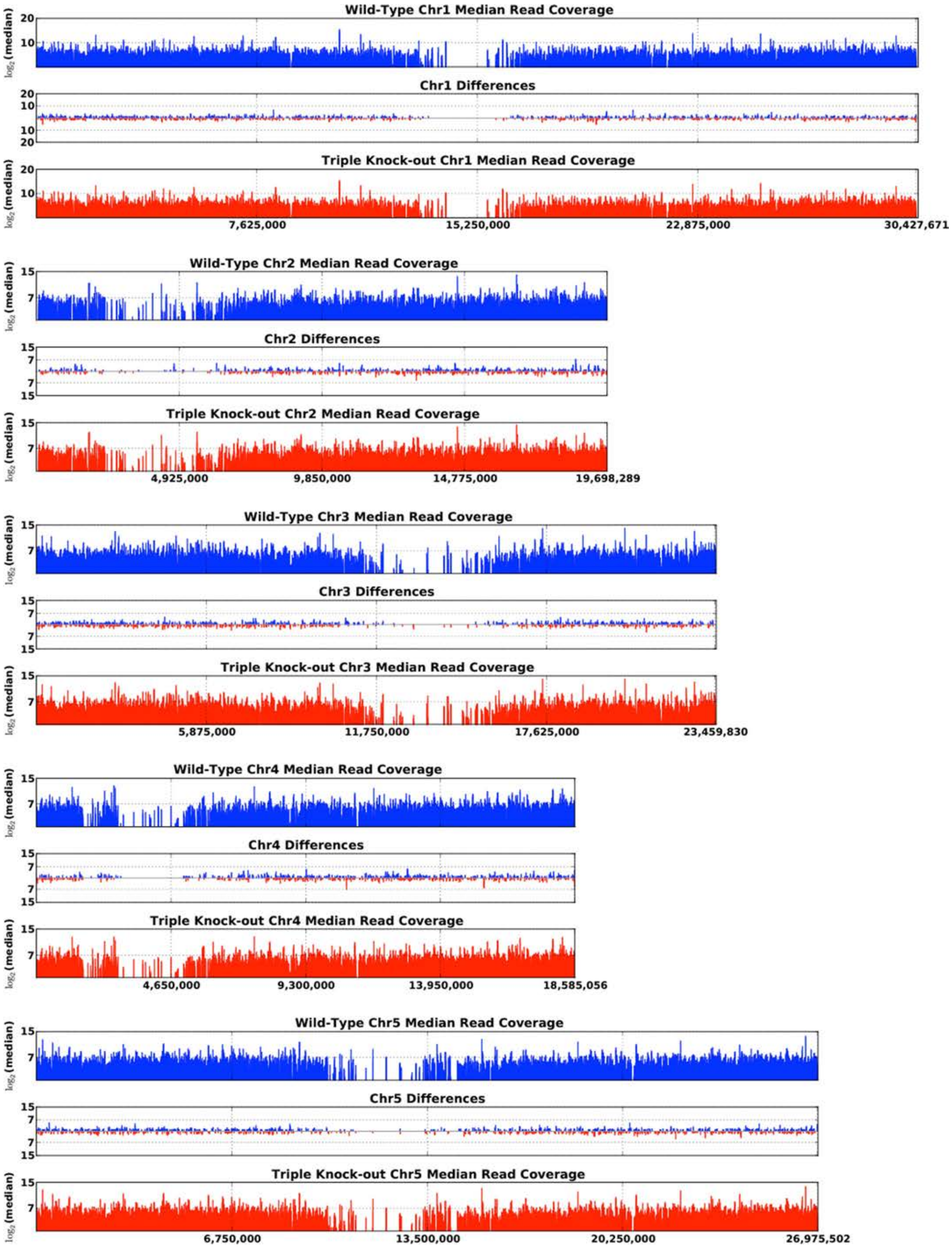


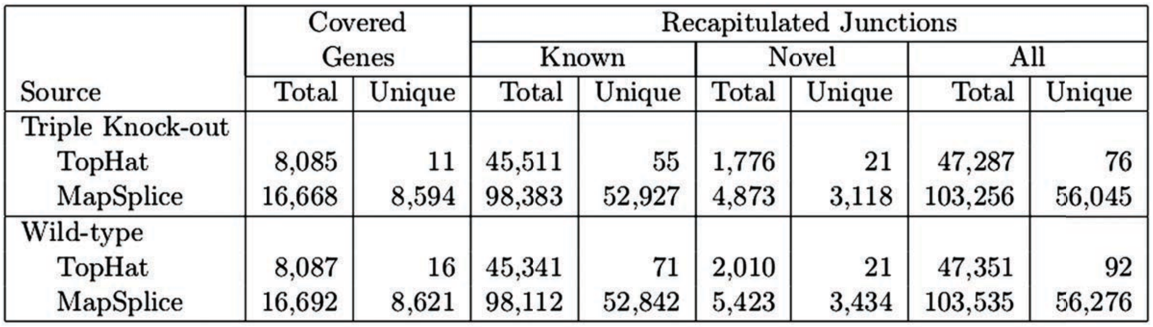
Figure 4.4: Uniformity of read coverage across the genome in WT and tm.
Reads were mapped on the chromosome labeled 1-5. The median number of reads were mapped in 1000 nucleotide windows across the chromosomes, the bar heights indicate read counts. For each chromosome the read counts (combined between replications) are given for the wild-type (in blue) and the triple mutant (red), with the differences in reads present in one genotype but not the other shown in between.

Mapping of RNA-Seq reads to the Arabidopsis genome

We aligned the RNA-Seq reads from each of our data sets to the TAIR10 version of the Arabidopsis genome with TopHat (Trapnell et al., 2012), an extensively used program to align spliced reads, and MapSplice (Wang et al., 2010), which is reported to perform better than TopHat in terms of specificity and sensitivity in detecting splice junctions. We found that MapSplice provided better read coverage than Tophat (Trapnell et al., 2012), and greater sensitivity for spliced reads. On average, about 85-87% of the reads were mapped with TopHat whereas around 90-92% reads were mapped with MapSplice. About 3 to 4 million mapped reads out of about 20 million in each sample showed gapped alignment i.e, mapped to splice site junctions and the rest (about 13 to 14 million) showed ungapped alignment (Table 4.1). Both ungapped and spliced aligned reads for WT and TM were better with MapSplice than TopHat (Table 4.2). MapSplice identified more than twice as many splice junctions in the RNA-Seq data, including several thousands of novel junctions (Table 4.2). In addition, we found that SpliceGrapher's classifiers (Rogers et al., 2012) identified 33% of novel TopHat junctions as false-positives whereas for MapSplice the false-positive rate was 30%, demonstrating that the increased splice-junction sensitivity with MapSplice does not appear to increase the proportion of false-positive junctions.

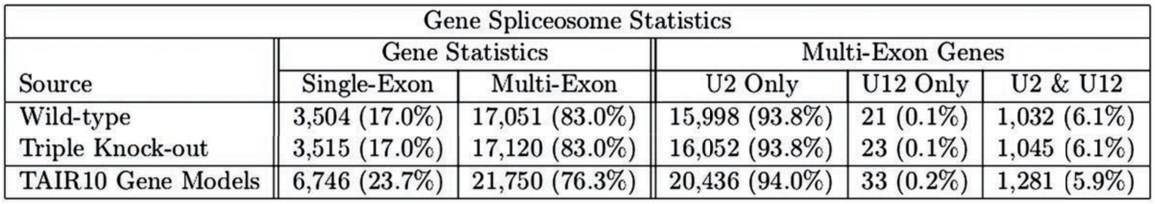
The TAIR10 release of Arabidopsis has about 28496 genes of which 6,746 are intronless genes and the remaining 21,750 genes have one or more introns. RNA-Seq reads from each genotype mapped to about half of single exon genes (~3500) and about 78% of multi-exon genes, suggesting that about 70% of all genes are expressed in two week-old seedlings (Table 4.3). Most introns in Arabidopsis (over 94%) and other plants have canonical GT/AG splice sites and are spliced by the major U2 spliceosome whereas a small fraction have non-canonical splice sites such as AT/AC donor/acceptor sites are spliced by a minor U12 spliceosome. To assess splicing activity associated with the U2 and U12 spliceosomes, we counted the number of splice junctions recapitulated in the RNA-Seq data and distinguished between canonical GT donor sites and semi-canonical (GC) or non-canonical splice sites (Table 4.3). We considered a splice junction to be recapitulated if there were at least two reads across the junction. The analysis of the splice junctions for U2 and U12 type splicing shows that transcripts processed by both types of spliceosomes are represented in both wild type and mutant RNA-Seq data (Table 4.3). To assess the abundance of highly expressed transcripts, we analyzed the read coverage of the top 9 most abundantly expressed genes in our RNA-Seq data and found that they accounted for 7.29 to 7.81% of all reads in wild type and TM, respectively (see Table 4.4). This level of coverage is consistent with non-normalized data in other RNA-Seq experiments (Filichkin et al., 2010; Marquez et al., 2012). All nine highly expressed genes are involved in photosynthetic processes (Table 4.4). Prior to analyzing the gene expression, we verified expression of the three mutated SR genes. As expected, *SC35* and *SCL30a* are not expressed in the mutant whereas some reads of *SCL33* are found as the T-DNA insertion is in the last exon. In many knockouts, if the T-DNA insertion is in the later part of the gene, transcripts 5’ to the insertion site are observed (Ali et al., 2007; Du et al., 2009).

Table 4.2: Comparison of spliced alignment statistics for TopHat and MapSplice alignments. Shown are the numbers of genes, the number of annotated splice junctions and the number of novel splice junctions identified by each method. Unique columns show the genes and splice junctions unique to each method. MapSplice was able to align twice as many spliced reads as TopHat, including 99.8% of the junctions identified by TopHat.



	Covered Genes		Recapitulated Junctions					
			Known		Novel		All	
Source	Total	Unique	Total	Unique	Total	Unique	Total	Unique
Triple Knock-out								
TopHat	8,085	11	45,511	55	1,776	21	47,287	76
MapSplice	16,668	8,594	98,383	52,927	4,873	3,118	103,256	56,045
Wild-type								
TopHat	8,087	16	45,341	71	2,010	21	47,351	92
MapSplice	16,692	8,621	98,112	52,842	5,423	3,434	103,535	56,276

Table 4.3: Summary of spliceosome activity found in RNA-Seq data for WT and tm data sets. Shown are the number of single-exon and multi-exon genes with read coverage in the two data sets. Multi-exon genes are further divided into those with only U2 and U12 splicing activity in the data, and those with both kinds. For comparison, we show for each category the counts from the TAIR 10 gene models.



Gene Spliceosome Statistics					
	Gene Statistics		Multi-Exon Genes		
Source	Single-Exon	Multi-Exon	U2 Only	U12 Only	U2 & U12
Wild-type	3,504 (17.0%)	17,051 (83.0%)	15,998 (93.8%)	21 (0.1%)	1,032 (6.1%)
Triple Knock-out	3,515 (17.0%)	17,120 (83.0%)	16,052 (93.8%)	23 (0.1%)	1,045 (6.1%)
TAIR10 Gene Models	6,746 (23.7%)	21,750 (76.3%)	20,436 (94.0%)	33 (0.2%)	1,281 (5.9%)

Table 4.4: The top ten genes with the highest expression values in the wild-type and triple knock-out data. A total of 23,088 genes had some read coverage in the wild-type data and 23,033 genes had coverage in the triple knockout. In each case, the top ten genes accounted for no more than 7.8% of all reads.

GENE	WT		TM		Annotation
	Reads	% Reads	Reads	% Reads	
AT1G29930	523,989	1.13%	482,003	1.02%	CAB1, CHLOROPHYLL A/B BINDING PROTEIN 1
AT2G39730	453,042	0.98%	550,186	1.16%	RCA, RUBISCO ACTIVASE
AT3G47470	414,566	0.90%	409,027	0.86%	LHCA4, LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT A4
AT1G67090	385,941	0.83%	544,583	1.15%	RBCS1A, RIBULOSE BISPHOSPHATE CARBOXYLASE SMALL CHAIN 1A
AT4G10340	327,210	0.71%	350,831	0.74%	LHCB5, LIGHT HARVESTING COMPLEX OF PHOTOSYSTEM II 5
AT5G54270	304,221	0.66%	303,213	0.64%	LHCB3, LIGHT-HARVESTING CHLOROPHYLL B-BINDING PROTEIN 3
AT2G34420	276,378	0.60%	287,896	0.61%	LHCB1.5, PHOTOSYSTEM II LIGHT HARVESTING COMPLEX GENE 1.5,
AT1G29920	249,814	0.54%	290,345	0.61%	CAB2, CHLOROPHYLL A/B-BINDING PROTEIN 2
AT3G54890	227,574	0.49%	243,981	0.51%	LHCA1, PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 1
AT1G61520	209,653	0.45%	0	0.00%	LHCA3, PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 3
AT1G29910	0	0	239,856	0.51%	CAB3, CHLOROPHYLL A/B BINDING PROTEIN 3

Quantitative difference in gene expression in TM

The RNA-Seq data was analyzed through a workflow of bioinformatics tools to assemble transcripts, identify differentially expressed transcripts and characterize isoforms. The overall workflow is shown in Figure 4.5, including the pipelines for MapSplice and Tophat/Cufflinks. MapSplice followed by SpliceGrapher specifically provided a qualitative analysis of alternative splicing and isoforms. Cufflinks, a suit of programs, is widely used to analyze RNA-Seq data to identify differentially expressed genes (Trapnell et al., 2012). The Cufflinks pipeline was used to identify genes that are differentially expressed in the tm as outlined in Figure 4.6. The alignment files from TopHat were used to assemble the transcript units using Cufflinks and the merged files were analyzed using Cuffdiff to identify differentially expressed genes. The Integrated Genome Browser (IGB) was then used to visualize the results. We have used EdgeR together with Mapslice alignment to verify the results from Cufflinks (see Figure 4.5).

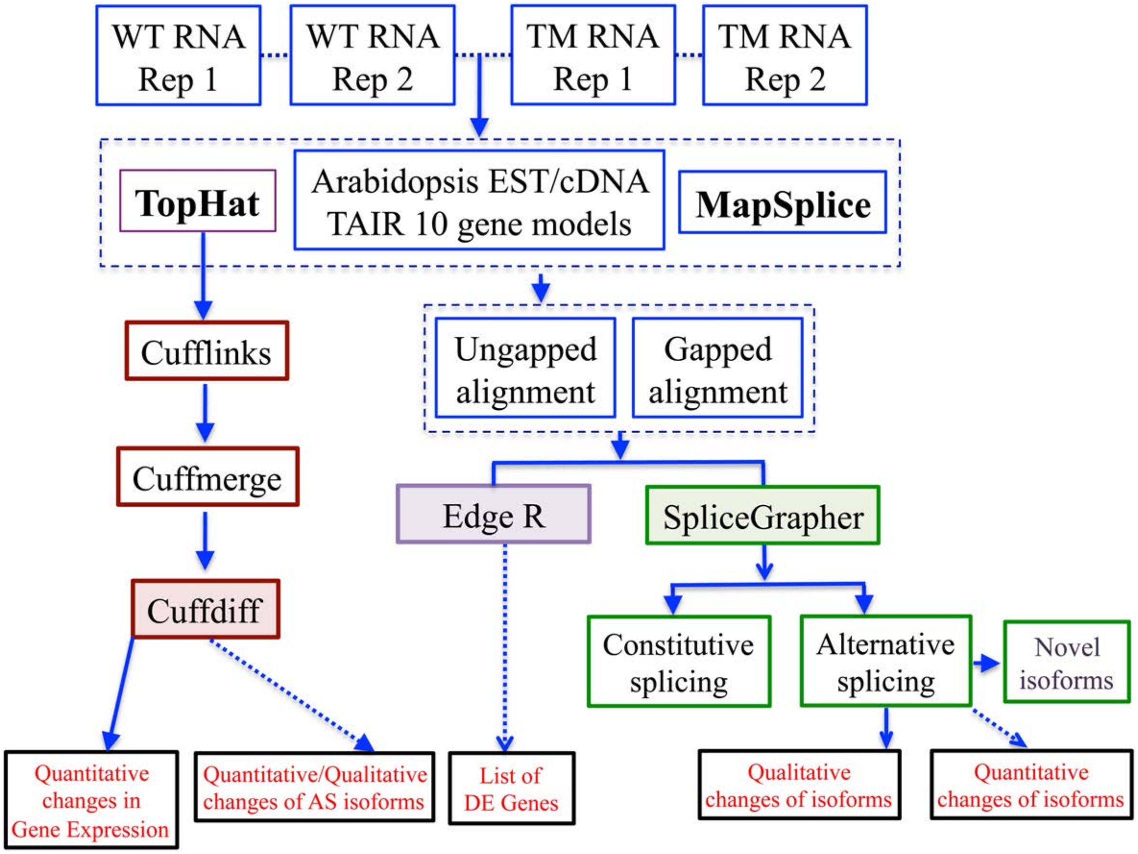


Figure 4.5: Pipeline for the analysis of gene expression using EdgeR and alternative splicing using SpliceGrapher.

The RNA sample from two replicates for each wild type and mutant line was used for RNA-Seq analysis. The data processing is outlined in boxes and arrows from the read counts generated from Illumina GAIIx platform. The reads are aligned to Arabidopsis Tair10 and EST/cDNA datasets using MapSplice and Tophat. The data from MapSplice is further analyzed using EdgeR for differential expression of genes or used by SpliceGrapher to draw out as a splice graph to depict the qualitative and quantitative differences in isoforms shown as differential AS events in WT versus tm mutant.

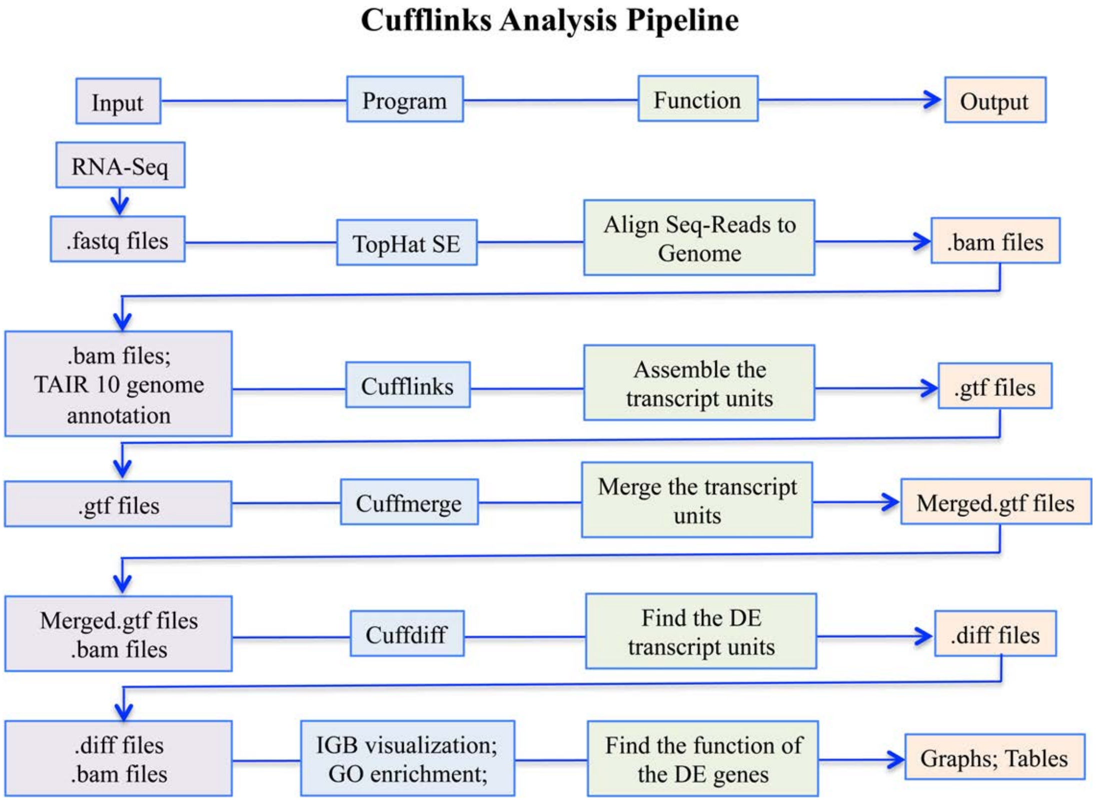

Cufflinks Analysis Pipeline
Input
Program
Function
Output
RNA-Seq
.fastq files
TopHat SE
Align Seq-Reads to Genome
.bam files
.bam files; TAIR 10 genome annotation
Cufflinks
Assemble the transcript units
.gtf files
.gtf files
Cuffmerge
Merge the transcript units
Merged.gtf files
Merged.gtf files .bam files
Cuffdiff
Find the DE transcript units
.diff files
.diff files .bam files
IGB visualization; GO enrichment;
Find the function of the DE genes
Graphs; Tables

Figure 4.6: Cufflinks pipeline used for analysis of RNA-Seq reads.
The Cufflinks tool (Trapnell 2010) was used for analysis of RNA-seq data of the 4 samples (2 replications of WT and tm genotypes). The pipeline shows the steps from input to output of each of the tools used within Cufflinks: TopHat for alignment, Cufflinks to assemble transcription units, Cuffmerge merges data for analysis, and Cuffdiff finds the differential expressed genes and individual isoforms.

Figure 4.7 shows a scattered plot of differentially expressed genes between wild type and TM and the volcano plot (Figure 4.8) shows differentially expressed genes with their *p* values. In the volcano plot the red dots represent genes, and their expression are similar in both genotypes. The blue dots indicate differentially expressed genes with reads towards the left higher in TM, and the reads towards right higher in WT.

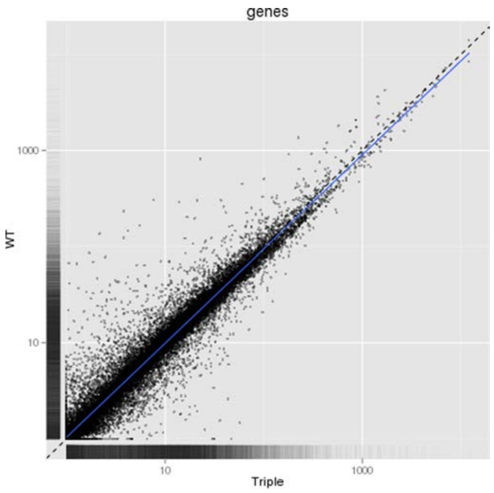


Figure 4.7: Scatter plot of reads from RNA-Seq.

The expression scatter plot is shown of all the genes in tm (X axis) and WT (Y axis). (*Each spot corresponds to a single gene and defined by its expression level in mutant (its x value) and wild type (its y value)*). Most of the spots are in the center around the diagonal line, suggesting that the expression of those genes is the same in both conditions (wt and tm). However, a fraction of the spots was clearly deviated from the line, suggesting differential expression of these genes between WT and tm.

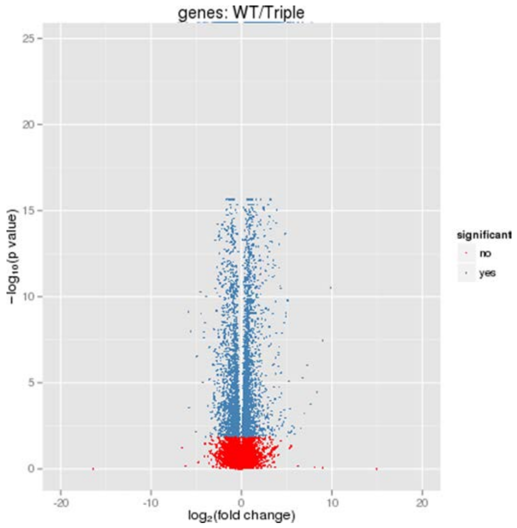


Figure 4.8: Volcano plot of the differentially expressed genes.

The ratios $\log_2$(fold change) of WT/TM of each gene were plotted against their statistical significance in $\log_{10}$(p-value). The red dots correspond to the genes with same expression between WT and tm, and the blue dots to the genes whose expressions are different with the p-value < 0.05

A total of 737 genes were identified that are differentially expressed with a cutoff of 0.05 *p* values and three fold change. This includes 351 up-regulated and 386 down-regulated genes. Five microRNAs and thirteen novel transcriptional units are also differentially regulated. The one hundred up-regulated and down-regulated genes are presented in Table 4.5A and 4.5B. The list of affected miRNAs and coordinates of novel transcripts are presented in Table 4.6A and 4.6B. A complete list of all differentially expressed genes with fold change and significance values will be posted on the TM-RNA-Seq web page.

The read coverage of some down and up-regulated protein coding genes in the wild type and mutant as visualized using the Integrated Genome Browser (IGB) is shown in Figure 4.9. A similar graphic view of expression of miRNAs and novel transcripts is presented in Figures 4.10 and 11, respectively. To further evaluate differential gene expression results obtained with Cufflinks, we are using another pipeline that uses MapSplice and EdgeR (Robinson et al., 2010) to quantify the reads between WT and TM. This analysis is in progress. However, preliminary results indicate that there is considerable overlap between the differentially expressed gene lists generated from Cufflinks and EdgeR.

Table 4.5a: List of 100 Up-regulated genes in the triple mutant

Genes	Fold change	Genes	Fold change
AT4G23690	1.60048	AT1G01580	4.02169
AT1G12040	1.60567	AT4G14690	4.07147
AT3G05890	1.6092	AT5G56080	4.7262
AT4G08410	1.63297	AT2G30766	4.74976
AT1G51860	1.6385	AT1G13609	5.99101
AT3G13610	1.66746	AT4G31940	6.66405
AT5G64100	1.72595	AT4G19690	6.90151
AT3G24460	1.73391	AT4G30450	1.58604
AT4G15160	1.73881	AT3G25930	1.58771
AT2G46750	1.75945	AT4G08400	1.80788
AT5G17820	1.77256	AT1G30730	2.13779
AT5G35190	1.77834	AT5G14650	1.78456
AT1G08430	1.86132	AT2G41480	2.40152
AT4G05200	1.86189	AT3G11550	2.38905
AT3G11340	1.86724	AT1G74500	2.44603
AT2G15020	1.93785	AT5G45070	2.18142
AT1G51420	2.00822	AT5G47980	2.36744
AT1G51830	2.06607	AT2G27370	2.83383
AT2G32300	2.08136	AT2G35380	1.60985
AT2G28670	2.08907	AT4G12545	2.37828
AT1G15380	2.10262	AT1G72230	1.93103
AT3G53480	2.11372	AT5G60660	1.78584
AT5G42590	2.15438	AT3G50300	1.96702
AT2G21045	2.16525	AT2G41660	1.84436
AT1G80240	2.2211	AT3G02885	1.74972
AT2G18980	2.22956	AT3G16440	2.34248
AT5G05730	2.28394	AT3G32030	2.22035
AT1G73600	2.30197	AT1G29025	2.4875
AT5G43580	2.35414	AT2G27000	1.67807
AT3G57010	2.37199	AT3G61270	1.80178
AT4G13580	2.41048	AT5G01040	1.75773
AT5G35935	2.41671	AT3G54770	2.46062
AT5G05250	2.44994	AT2G24610	2.29862
AT1G23720	2.47255	AT4G31870	1.70282
AT3G23470	2.52928	AT5G13580	1.89463
AT5G66390	2.53219	AT5G56320	2.08722
AT2G39430	2.60049	AT1G78090	2.79476
AT2G36100	2.62121	AT3G19430	4.43968
AT1G51470	2.6234	AT5G58310	2.33117
AT1G12740	2.65222	AT2G18800	3.84977
AT1G47600	2.69237	AT1G02360	1.81935
AT1G53830	2.7076	AT5G06839	1.62641
AT3G01190	2.73882	AT1G44970	1.74922
AT1G24580	2.81232	AT1G02810	1.89587
AT1G74770	2.89243	AT3G50400	1.93255
AT5G59090	2.93288	AT2G40113	2.03773
AT5G04150	2.99889	AT2G46740	2.68682
AT5G42180	3.08645	AT2G47540	1.91633
AT1G78340	3.23166	AT2G39110	1.58545
AT5G60530	3.9235	AT1G47395	4.4279

Table 4.5b: List of 100 Down-regulated genes in the triple mutant

Genes	Fold change	Genes	Fold change
AT2G18660	-7.84138	AT1G67600	-2.55639
AT2G34210	-5.80543	AT5G54610	-2.53555
AT2G46880	-5.76682	AT3G61410	-2.47768
AT1G23110	-5.42229	AT2G43535	-2.4765
AT5G02200	-5.28409	AT2G15042	-2.44313
AT2G36724	-5.13561	AT5G01740	-2.44228
AT2G26560	-4.67893	AT4G32280	-2.42942
AT3G22235	-4.4706	AT3G61280	-2.40192
AT1G21520	-4.42233	AT4G28790	-2.38069
AT2G40750	-4.35032	AT2G46430	-2.35514
AT5G64000	-4.34637	AT2G25510	-2.33938
AT5G10760	-4.25226	AT4G01380	-2.31795
AT3G13570	-4.12275	AT5G01600	-2.31703
AT2G32680	-4.08081	AT2G46440	-2.29668
AT2G04040	-4.03523	AT2G47015	-2.27428
AT3G05630	-3.92205	AT1G22220	-2.27037
AT1G35710	-3.84798	AT3G05660	-2.23259
AT5G25250	-3.50849	AT1G17420	-2.2017
AT2G36970	-3.48485	AT4G23000	-2.1931
AT5G46330	-3.47707	AT1G15040	-2.1757
AT4G11890	-3.41767	AT4G27300	-2.16704
AT5G50335	-3.38513	AT3G47420	-2.15491
AT5G20410	-3.36341	AT2G23790	-2.15485
AT2G24600	-3.3415	AT5G25440	-2.14077
AT1G21250	-3.14219	AT5G17860	-2.09746
AT3G44510	-3.11258	AT3G25760	-2.08124
AT1G23730	-3.06362	AT2G17040	-2.07506
AT4G14400	-3.06359	AT1G24575	-2.05452
AT5G01220	-3.05821	AT4G37370	-2.05079
AT5G52750	-3.01252	AT2G30250	-2.04075
AT3G49570	-2.96748	AT4G36648	-2.01027
AT2G30540	-2.95389	AT2G31865	-2.00855
AT3G49160	-2.93675	AT4G33770	-2.00687
AT1G09350	-2.92004	AT1G66920	-1.9997
AT4G30270	-2.89788	AT5G47220	-1.98422
AT5G39610	-2.89716	AT2G40300	-1.9818
AT1G23140	-2.7853	AT2G28400	-1.97464
AT4G12290	-2.76811	AT4G38550	-1.97397
AT4G23220	-2.76366	AT1G56600	-1.97277
AT4G16780	-2.76269	AT3G28540	-1.96651
AT1G52890	-2.75788	AT4G03960	-1.96277
AT2G41090	-2.74267	AT1G02340	-1.93779
AT2G29460	-2.73442	AT1G06080	-1.92554
AT4G16540	-2.70426	AT3G22060	-1.89346
AT1G07620	-2.64617	AT5G57240	-1.8886
AT1G13750	-2.63061	AT3G45730	-1.87609
AT1G11210	-2.6213	AT4G33030	-1.87223
AT5G42530	-2.59313	AT5G07100	-1.85978
AT3G26830	-2.57892	AT3G51430	-1.83234
AT3G26840	-2.57277	AT5G24660	-1.8259

Table 4.6a: List of differentially expressed miRNAs with fold change and significance

Gene	Gene ID	Log_2(FC)	q value
MIR408	AT2G47015	-2.27428	0
MIR399D	AT2G34202	-9.95462	6.58E-10
MIR167A	AT3G22886	-1.60847	0.000137755
MIR399C	AT5G62162	-7.71234	0.00123805
MIR171C	AT1G62035	-1.90022	0.0311283

Table 4.6b: List of differentially expressed novel transcripts with fold change and significance

Gene Symbol	Locus	Log_2(FC)	q-value
Novel_12	1:11453622-11455260	-3.88881	0
Novel_1	1:28515202-28515898	1.79348	8.10E-09
Novel_10	2:7632584-7633782	-3.44666	0.00441307
Novel_9	3:10601415-10601846	-2.87174	0.00471068
Novel_13	3:4372540-4373208	-4.26318	0.00709564
Novel_5	3:15611314-15612223	-2.086	0.0162963
Novel_7	1:1480148-1480881	-2.31242	0.0287319
Novel_6	1:29328033-29329121	-2.15688	0.0410125
Novel_2	5:1041891-1042827	1.68184	0.0412271
Novel_8	5:25980258-25981099	-2.67668	0.0412271
Novel_4	1:29740235-29741294	-2.05621	0.0424771
Novel_11	3:22556308-22557024	-3.86086	0.0442921
Novel_3	3:15613454-15614807	-1.92067	0.0443835

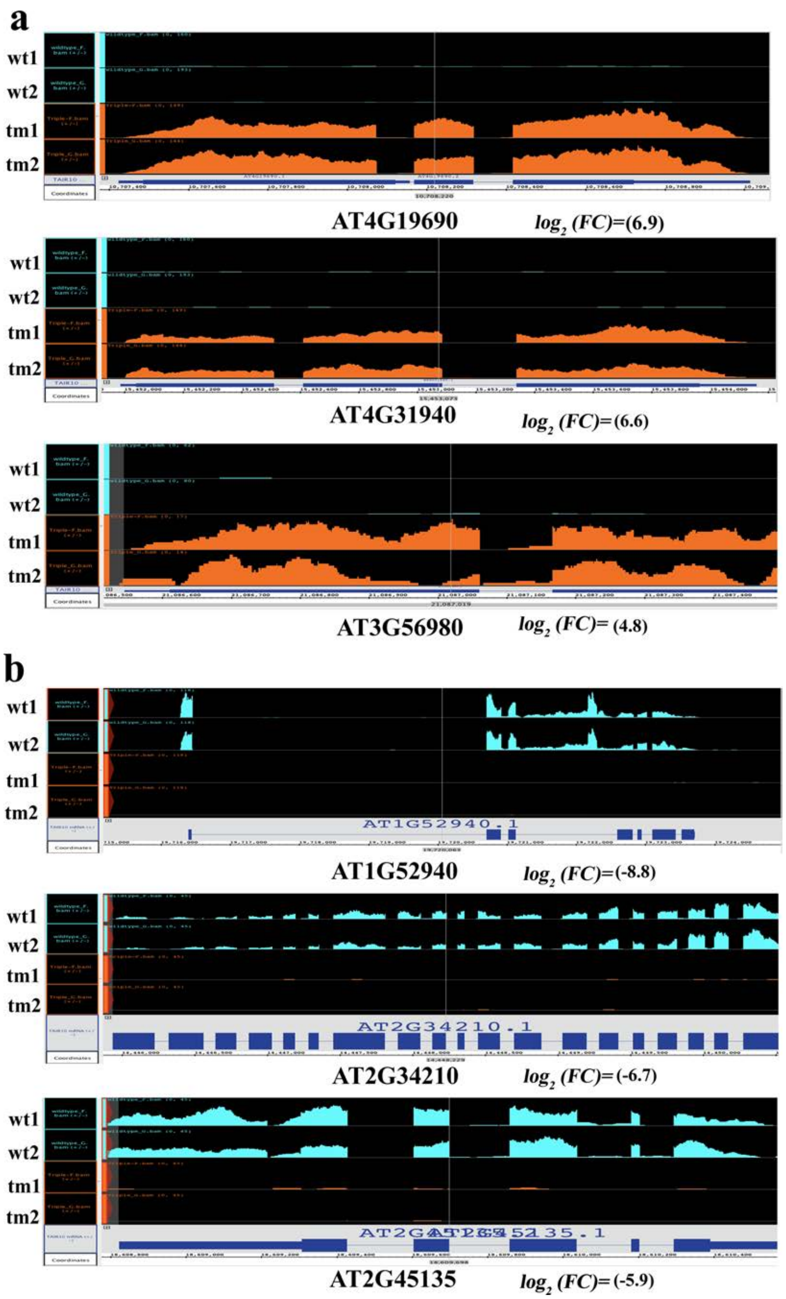


Figure 4.9: Visualization of differential expression of three up-regulated and three down-regulated protein-coding genes in the mutants using IGB (integrated genome browser) a) The three up-regulated genes show higher read counts in tm (orange) compared to wt (blue) and a $\log_2$ fold change of gene expression for each gene AT4g19690, At4g31940 and At3g56980 is shown. b) The lower three genes show very few read counts in tm (orange) compared to wt (blue) and the level of expression measured as a $\log_2$ fold change for each gene At1g52940, At2g34210 and At2g45135 in tm versus wt is also shown. The uniformity within each replicates of mutant line (tm1; tm2) and wild line (wt1; wt2) is shown for each gene.

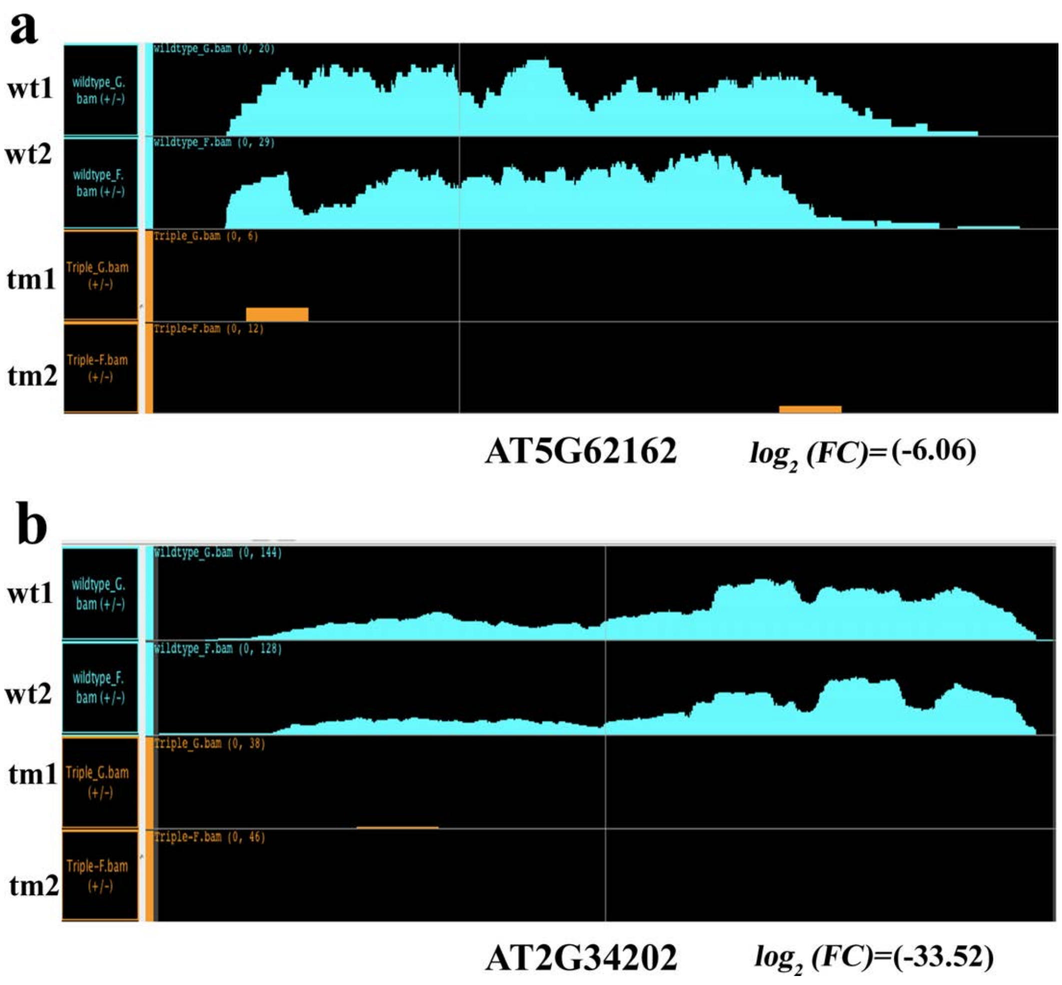


Figure 4.10: Reduced expression of two micro-RNA (miRNA) genes in triple mutant from Integrated Genome viewer (IGB)

The up-regulation of miRNA transcripts shown in wt compared to tm for two (a, b) miRNA genes. The uniformity of reads between the replicates and the log2 fold change in gene expression for AT5g62162 and At2G34202 genes in wt and tm lines is shown.

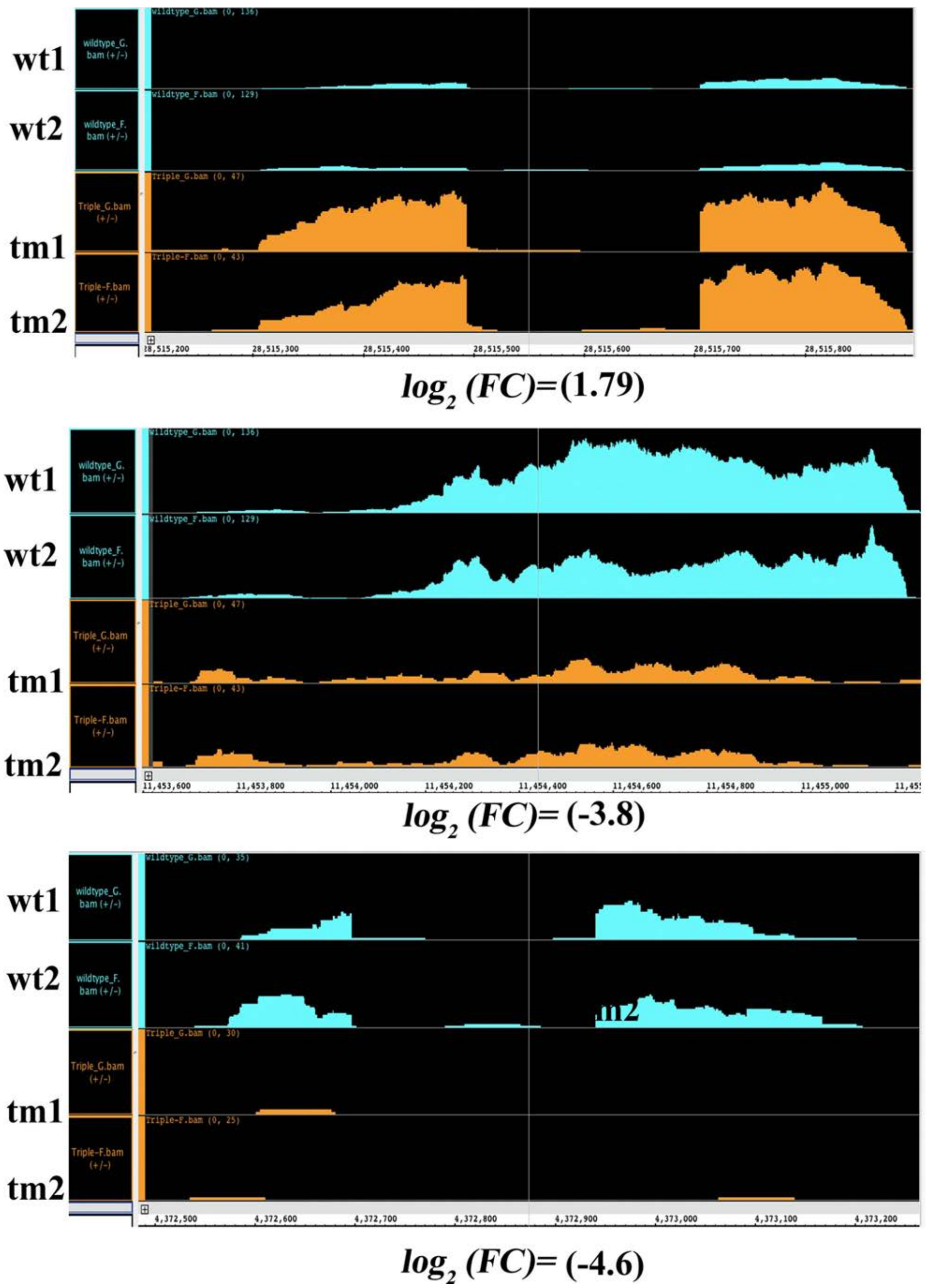


Figure 4.11: Differential expression of transcripts from thee novel genes (unannotated in TAIR10 version) in wild type compared to the mutant line IGB

The first panel shows more read counts in tm replicates compared to wt and lower two panels show reduced expression in tm as compared wt**.**

Validation of RNA-Seq results using RT-PCR

We selected 16 upregulated and 20 down-regulated genes in the mutant including several protein coding genes, some miRNAs and a few novel transcripts for validation using RT-PCR. Validation of these results by RT-PCR is presented in Figure 4.12. The expression pattern of 35 out of 36 genes is confirmed by RT-PCR. Interestingly, several of the selected genes are alternatively spliced and splicing patterns of these pre-mRNAs is altered in the mutant. Of the 36, 18 genes showed changes in alternative splicing that represented both qualitative and quantitative changes (discussed below). Among the induced genes, isoform 1 of *IRT1* (At4g19690) is present only in the mutant where isoform 2 is present in wild type and mutant with much higher levels in the mutant (Figure 4.12a). One of two transcripts variants of a novel gene (At2g35637) is significantly higher in the wild type and the other isoform level is similar in both wild type and the mutant (Figure 4.12b). In the case of Phy A, one of the three splice variants is present only in the wild type (Figure 4.12b).

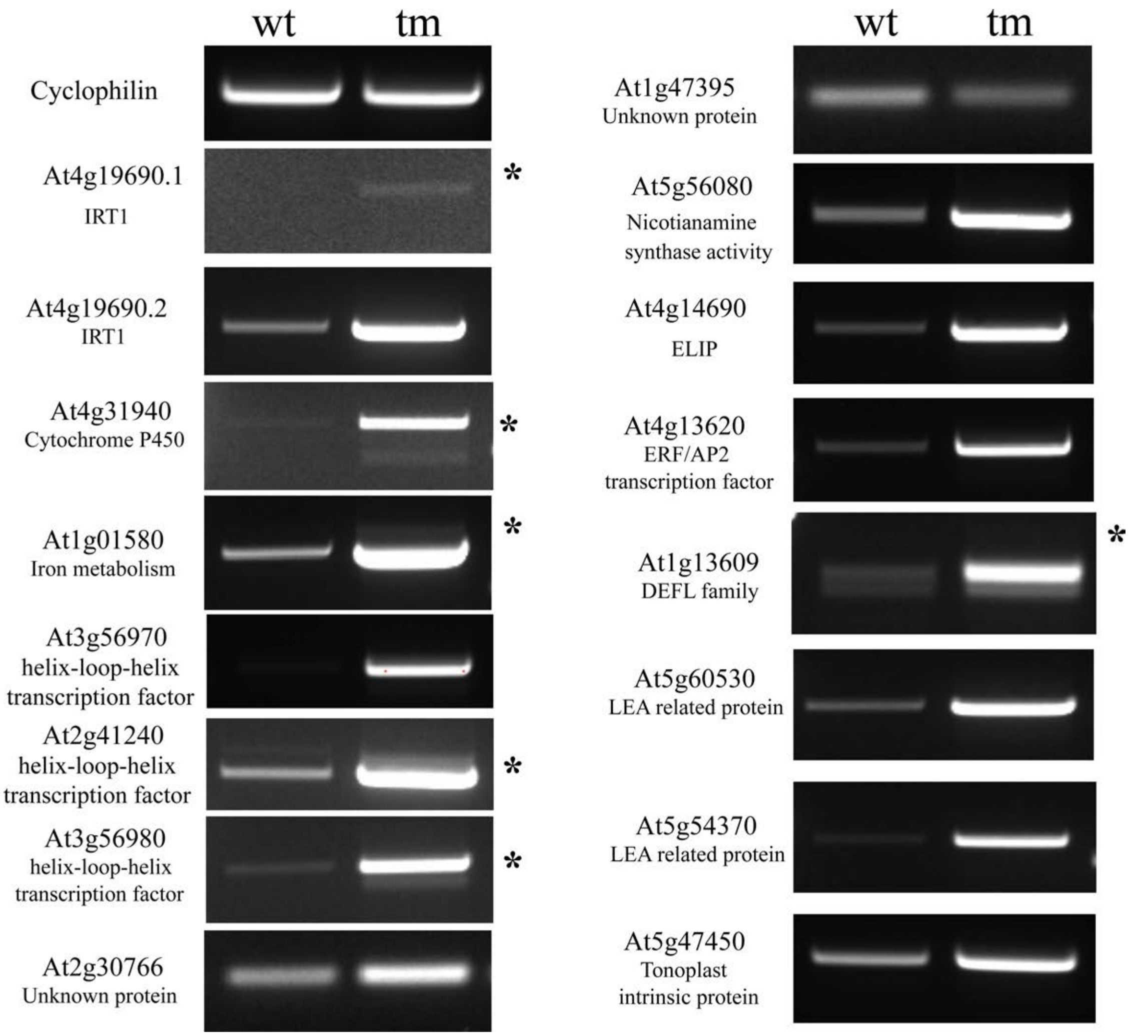


Figure 4.12a: Validation of 16 up-regulated genes by RT-PCR.

Expression of 16 up-regulated genes in the mutant are verified by RT-PCR. Cyclophilin expression (top panel) shows equal amount of cDNA in both samples. Asterisks indicate gene that show splice variants.

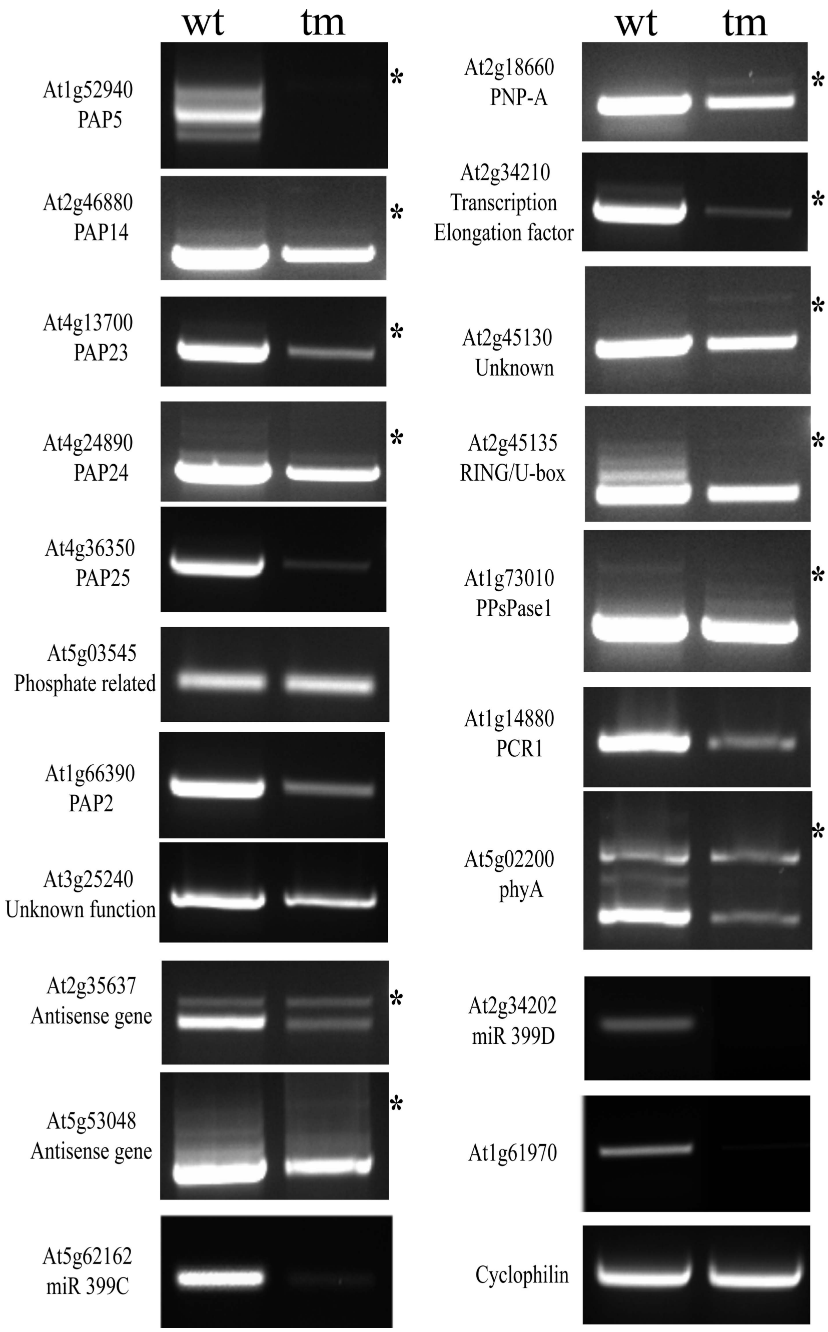


Figure 4.12b: Validation of 20 genes that are down regulated in the triple mutant. The antisense gene 9 (At2g35637) not only shows an over-all change in expression but also a decrease in the lower isoform in tm compared to wt, although the level of upper isoform is equal in both lines. The cyclophilin gene is used as a control to verify equal amounts of RNA in both samples. The asterisks indicate the genes undergoing alternative splicing. The fold change and description of these genes are shown in Table 4.5 and is highlighted in yellow.

Analysis of differential splicing

Two types of changes in AS of pre-mRNA are expected when comparing AS in wild type and mutants of splicing regulators. These include i) qualitative changes in AS where a particular isoform is absent in either wild type or in the mutant, and /or ii) quantitative changes in AS where the levels of one or more splice variants is altered either in the wild type or mutant. Analysis of qualitative and quantitative changes in alternative splicing is not trivial and computations tools for this analysis are still being developed and refined. Some of the tools that are in use for mammalian systems are not readily applicable to plants as there are major differences in the prevalence of AS types between plants and animals. Our collaborators in Computer Sciences at CSU has recently developed a SpliceGrapher tool (Rogers et al., 2010) that used RNA-Seq data from plants to predict splicegraphs. This tools is being refined to determine qualitative changes in gene expression in our TM mutant. Preliminary results from this analysis are presented below.

Qualitative changes in AS in TM

To identify qualitative changes in AS between wild type and TM, we used the SpliceGrapher tool (Rogers et al., 2010). To account for variation within each data set, for every gene we compared the splice forms between replicates. Genes whose splice forms differed between replicates were discarded from our analysis. This yielded two sets of genes that showed consistent splicing patterns across replicates: one set for wild-type (WT) and one set for the triple knockout mutant (TM). For those genes common to both sets, we compared splice forms between the WT and TM to identify genes that showed qualitative changes in AS between WT and mutant. This comparison yielded about 30 genes that exhibited differences in annotated splice forms between the wild type and TM. In addition, several non-annotated (i.e, novel) splice forms also changed and this list is being compiled. A web site is being developed to post these results.

All types of AS events (intron retention, exon skipping, alternative 5’, alternative 3’, both alternative 5’ and 3’) were represented in the differentially splice variants. Three qualitative changes in alternative splicing examples that involve intron retention, use of alternative 5’ splice or alternative 3’ splice site are presented here. Figure 4.13a shows that pre-mRNA from At1G61970 produces two splice variants with or without retaining the 2nd intron in the wild type. Splice junction reads corresponding to this splicing event show that the intron excised variant is produced only in the wild type but not in the mutant. Seven splice junction reads are found in wild type but none in the mutant (Figure 4.13a). Using a reverse primer corresponding to the splice junction of the intron excised variant together with a forward primer corresponding to the upstream region we verified that this isoform is produced only in the wild type (Figure 4.12b), suggesting the excision of this intron requires one or more of the SRs that are absent in the triple mutant. In the case of At2G23930, which encodes a small RNP SNRNP-G, two splice variants are produced by using alternative 5’ splice sites in exon 1. Both isoforms are present in the triple mutant whereas one isoform is present in the wild type (Figure 4.13b.). The splice junction reads for isoform 1 between exon 1 and 2 for WT are 198, and for TM are 223. For isoform 2 of At2G23930 the splice junction reads (9) between the alternative 5’ for exon 1 and 2 is only present in TM, while the splice junction reads for the other isoform are more than 200. The use of the alternative 3’ splice site in At5G40550, which encodes a protein with a SGF29 tudor-like domain with biological functions in histone-3 and -4 acetylation and response to salt (Kaldis et al., 2011), generates two isoforms (Figure 4.13c). Splice junction reads provide evidence for the presence of both forms in wild type whereas there is evidence for only one isoform in the mutant.

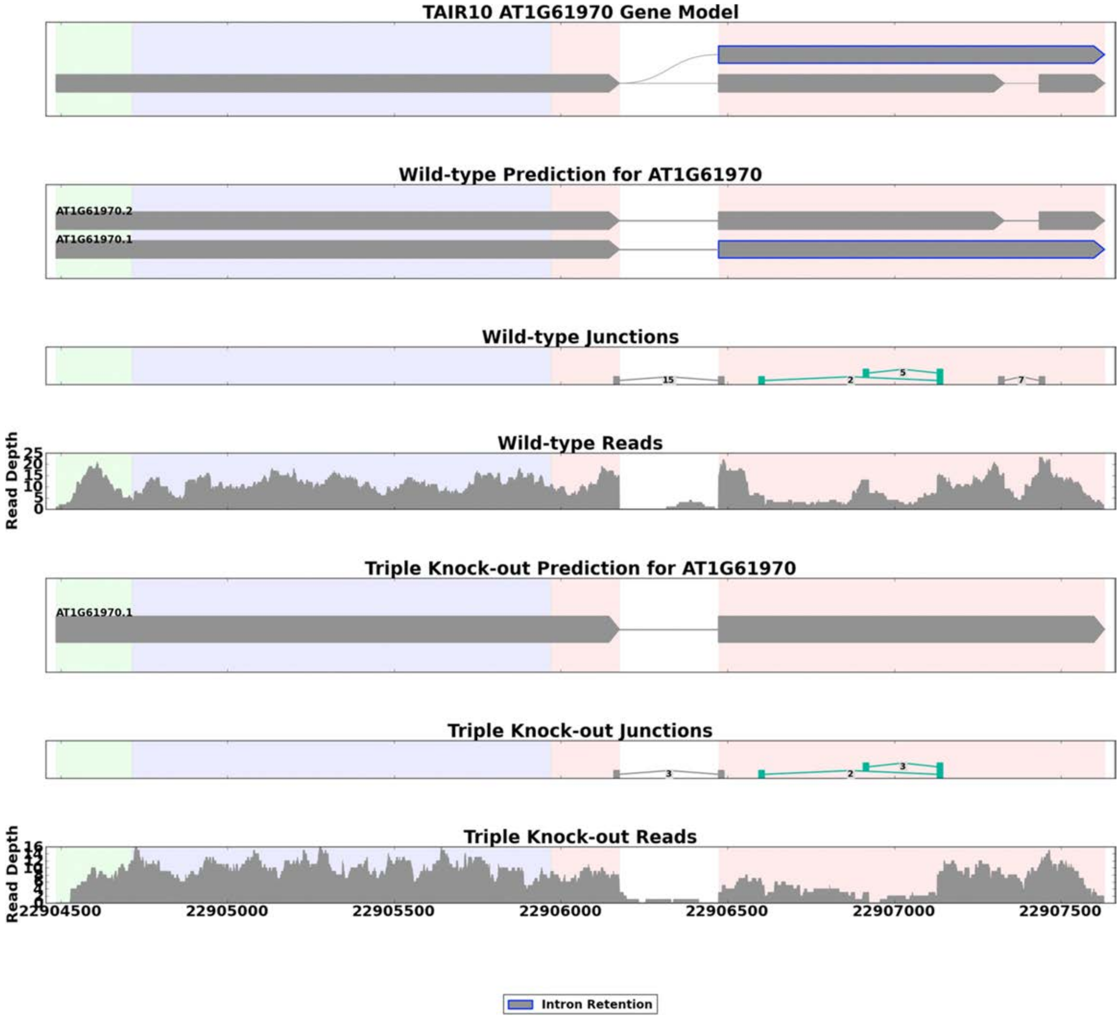


Figure 4.13: Qualitative changes in splice variants between wt type and mutant. The SpliceGraph shows alignment of transcript reads of the genes to the TAIR 10 model. Exons are boxed and the lines joining the exons are introns. Top splicegraph is based on TAIR 10 annotation and the splicegraphs of WT and tm are based on RNA-seq reads.

Figure 4.13a) Gene AT1G61970 shows an intron excision event of 2nd intron in wt with 7 splice junction reads and no splice junction reads at that location in tm line, suggesting that the intron excised form is absent in the mutant.

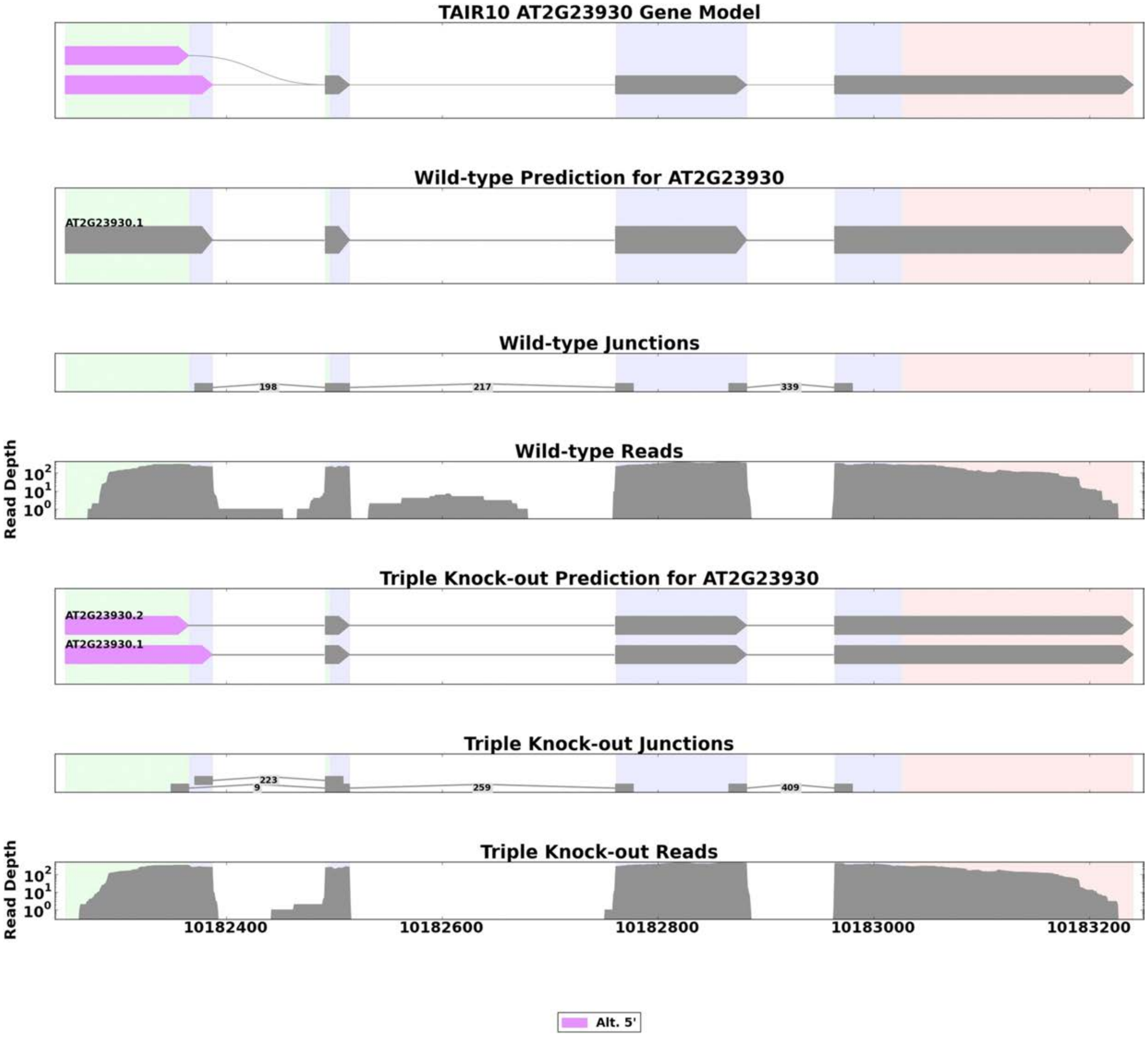


Figure 4.13b: AT2G23930 shows an example of qualitative AS with an alternate 5’ splice site in intron 1 that generates two isoforms in tm and only one isoform in wt. The two exons in bright purple show an alternate 5’ splice change. The number of reads for isoform 2 splice junction is 9 in tm line as compared to no read counts for that junction in wild type.

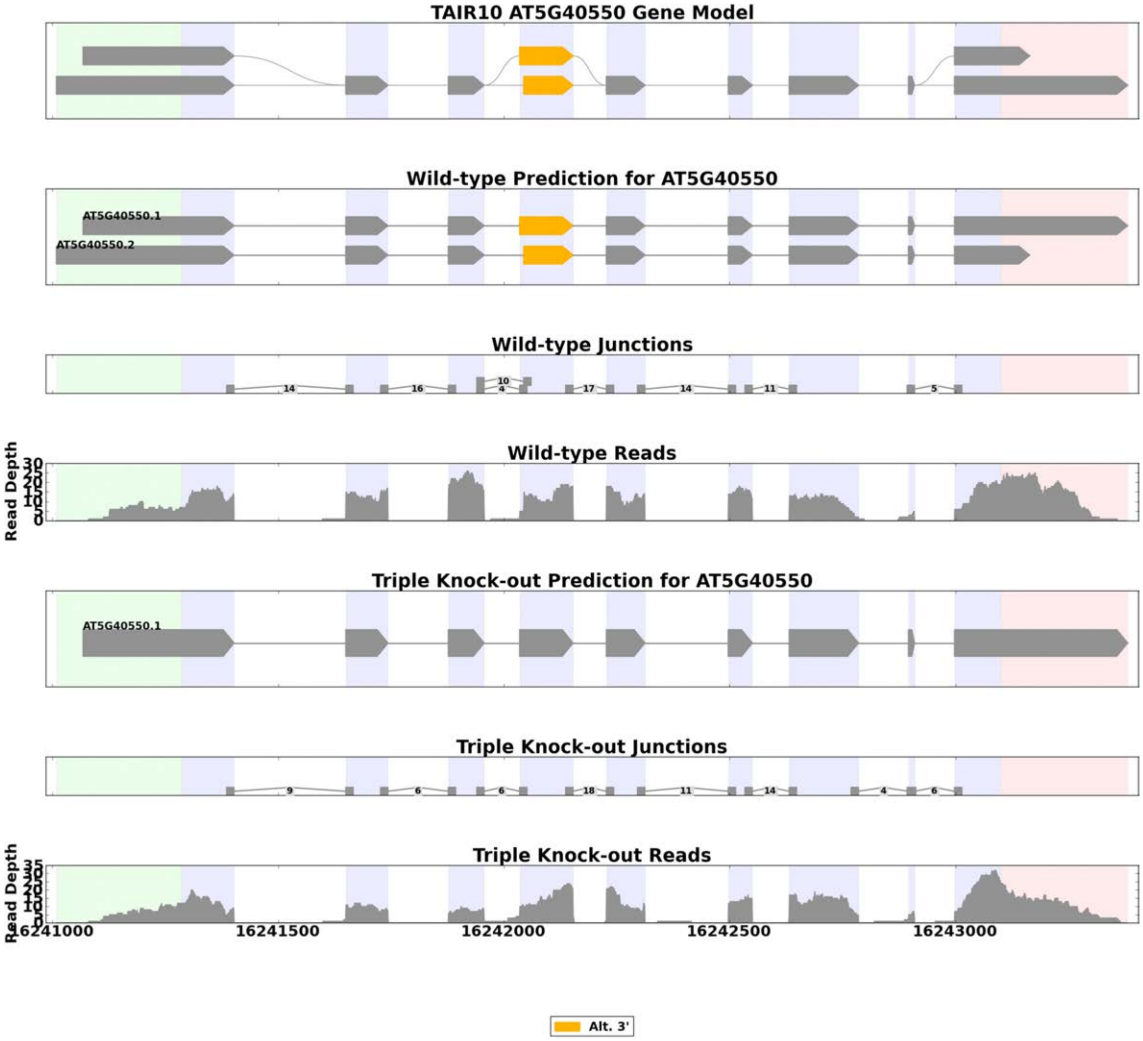


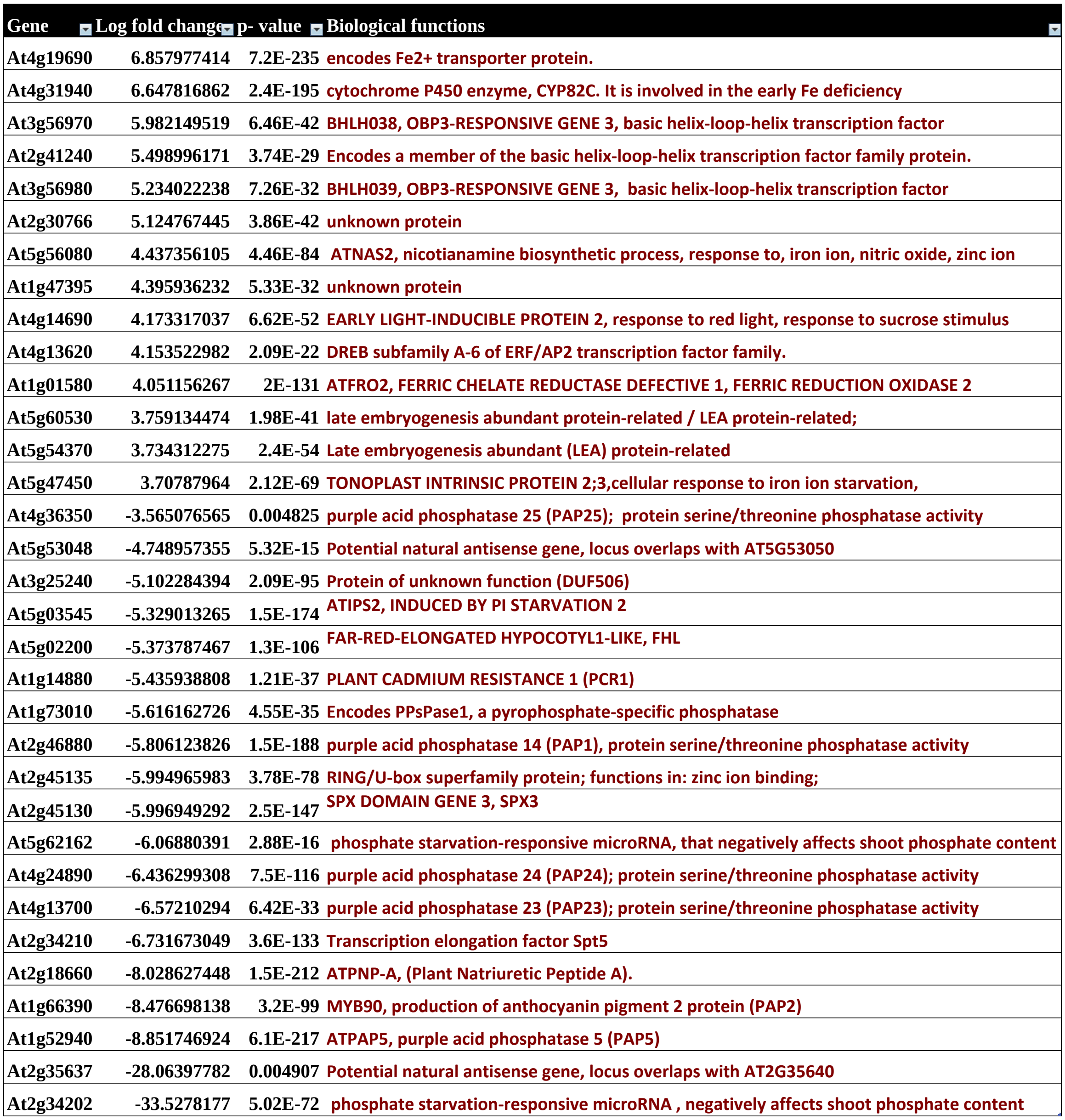
Figure 4.13c: AT5G40550 shows an example of Alternate (Alt) 3' splice site event during splicing of intron 3. The exon in yellow shows the splicing change. The number of reads for that splice junction for isoform 2 is 10 in wt line and no reads for that junction in tm and this depicts a qualitative change. The depth of the read ranges from 0 to 35.

Functional insights into the roles of these SRs based on RNA-Seq data from two-week old seedlings

One way to gain some functional insights into the roles of these SRs from transcriptomics study is to find the functional categories of genes that are differentially expressed in the mutant. We have analyzed all differentially expressed genes using gene ontology (G0) terms for biological processes. This revealed over representation of genes involved in plant immunity as well as iron and phosphorous homeostasis. The level of gene expression for these genes was found to be significantly different in the triple mutant as compared to the wild type (Table 4.7 and 8a and b). Expression of some of these is verified experimentally by RT-PCR (Figure 4.12a and 4.12b).

The first category of genes that were up-regulated in the triple mutant compared to wt included genes categorized to be important for iron uptake, translocation, subcellular translocation, and regulation in response to iron deficiency in higher plants (Table 4.8a). Several of the up-regulated genes in the mutant are the same as those that are induced by iron-deficiency (Kobayashi and Nishizawa, 2012). The fold change for the following genes: *Ferric-chelate reductase (FRO), Ferrous ion transporter (IRT1), Fe translocators (ATNAS2)*, positive transcriptional regulators such as *FIT*, and a subgroup of *BHLH* Ib comprising the genes AtbHLH38, AtbHLH39 and AtbHLH100 are shown in Table 4.7. Iron functions in various important processes, including photosynthesis, respiration, and chlorophyll biosynthesis; and is a component in heme, the Fe-sulfur cluster, and other Fe-binding sites (Kobayashi and Nishizawa, 2012).

Table 4.7: List of genes that are up-regulated or down-regulated in tm compared to wt. These genes show significant fold change in gene- expression and were also confirmed by RT-PCR (Figure 4.12a & b)



Gene	Log fold change	p- value	Biological functions
At4g19690	6.857977414	7.2E-235	encodes Fe2+ transporter protein.
At4g31940	6.647816862	2.4E-195	cytochrome P450 enzyme, CYP82C. It is involved in the early Fe deficiency
At3g56970	5.982149519	6.46E-42	BHLH038, OBP3-RESPONSIVE GENE 3, basic helix-loop-helix transcription factor
At2g41240	5.498996171	3.74E-29	Encodes a member of the basic helix-loop-helix transcription factor family protein.
At3g56980	5.234022238	7.26E-32	BHLH039, OBP3-RESPONSIVE GENE 3, basic helix-loop-helix transcription factor
At2g30766	5.124767445	3.86E-42	unknown protein
At5g56080	4.437356105	4.46E-84	ATNAS2, nicotianamine biosynthetic process, response to, iron ion, nitric oxide, zinc ion
At1g47395	4.395936232	5.33E-32	unknown protein
At4g14690	4.173317037	6.62E-52	EARLY LIGHT-INDUCIBLE PROTEIN 2, response to red light, response to sucrose stimulus
At4g13620	4.153522982	2.09E-22	DREB subfamily A-6 of ERF/AP2 transcription factor family.
At1g01580	4.051156267	2E-131	ATFRO2, FERRIC CHELATE REDUCTASE DEFECTIVE 1, FERRIC REDUCTION OXIDASE 2
At5g60530	3.759134474	1.98E-41	late embryogenesis abundant protein-related / LEA protein-related;
At5g54370	3.734312275	2.4E-54	Late embryogenesis abundant (LEA) protein-related
At5g47450	3.70787964	2.12E-69	TONOPLAST INTRINSIC PROTEIN 2;3,cellular response to iron ion starvation,
At4g36350	-3.565076565	0.004825	purple acid phosphatase 25 (PAP25); protein serine/threonine phosphatase activity
At5g53048	-4.748957355	5.32E-15	Potential natural antisense gene, locus overlaps with AT5G53050
At3g25240	-5.102284394	2.09E-95	Protein of unknown function (DUF506)
At5g03545	-5.329013265	1.5E-174	ATIPS2, INDUCED BY PI STARVATION 2
At5g02200	-5.373787467	1.3E-106	FAR-RED-ELONGATED HYPOCOTYL1-LIKE, FHL
At1g14880	-5.435938808	1.21E-37	PLANT CADMIUM RESISTANCE 1 (PCR1)
At1g73010	-5.616162726	4.55E-35	Encodes PPsPase1, a pyrophosphate-specific phosphatase
At2g46880	-5.806123826	1.5E-188	purple acid phosphatase 14 (PAP1), protein serine/threonine phosphatase activity
At2g45135	-5.994965983	3.78E-78	RING/U-box superfamily protein; functions in: zinc ion binding;
At2g45130	-5.996949292	2.5E-147	SPX DOMAIN GENE 3, SPX3
At5g62162	-6.06880391	2.88E-16	phosphate starvation-responsive microRNA, that negatively affects shoot phosphate content
At4g24890	-6.436299308	7.5E-116	purple acid phosphatase 24 (PAP24); protein serine/threonine phosphatase activity
At4g13700	-6.57210294	6.42E-33	purple acid phosphatase 23 (PAP23); protein serine/threonine phosphatase activity
At2g34210	-6.731673049	3.6E-133	Transcription elongation factor Spt5
At2g18660	-8.028627448	1.5E-212	ATPNP-A, (Plant Natriuretic Peptide A).
At1g66390	-8.476698138	3.2E-99	MYB90, production of anthocyanin pigment 2 protein (PAP2)
At1g52940	-8.851746924	6.1E-217	ATPAP5, purple acid phosphatase 5 (PAP5)
At2g35637	-28.06397782	0.004907	Potential natural antisense gene, locus overlaps with AT2G35640
At2g34202	-33.5278177	5.02E-72	phosphate starvation-responsive microRNA , negatively affects shoot phosphate content

Table 4.8a: Gene Ontology (GO) enrichment analysis of up-regulated genes in tm. The green-shaded GO categories are involved in metal ion homeostasis and the p-value suggests high significance levels, selecting these genes for further physiological validations.

GO Term	P-value	Sample frequency
GO:0000041 transition metal ion transport	7.93E-33	41/340 (12.1%)
GO:0006826 iron ion transport	6.99E-24	27/340 (7.9%)
GO:0010106 cellular response to iron ion starvation	1.91E-22	25/340 (7.4%)
GO:0010167 response to nitrate	9.59E-21	29/340 (8.5%)
GO:0030001 metal ion transport	6.53E-20	44/340 (12.9%)
GO:0015706 nitrate transport	4.73E-19	28/340 (8.2%)
GO:0006820 anion transport	1.74E-17	33/340 (9.7%)
GO:0006812 cation transport	2.74E-16	46/340 (13.5%)
GO:0010054 trichoblast differentiation	8.31E-14	29/340 (8.5%)
GO:0010053 root epidermal cell differentiation	3.36E-13	29/340 (8.5%)
GO:0048364 root development	8.08E-13	36/340 (10.6%)
GO:0022622 root system development	8.49E-13	36/340 (10.6%)
GO:0048765 root hair cell differentiation	4.55E-12	26/340 (7.6%)
GO:0048764 trichoblast maturation	4.55E-12	26/340 (7.6%)
GO:0048469 cell maturation	4.55E-12	26/340 (7.6%)
GO:0010015 root morphogenesis	6.15E-12	30/340 (8.8%)
GO:0021700 developmental maturation	1.26E-11	26/340 (7.6%)
GO:0031669 cellular response to nutrient levels	3.12E-11	26/340 (7.6%)
GO:0031667 response to nutrient levels	8.44E-11	26/340 (7.6%)
GO:0009267 cellular response to starvation	9.51E-11	25/340 (7.4%)
GO:0042594 response to starvation	1.63E-10	25/340 (7.4%)
GO:0031668 cellular response extracellular stimulus	3.33E-10	26/340 (7.6%)
GO:0009991 response to extracellular stimulus	8.51E-10	26/340 (7.6%)
GO:0009913 epidermal cell differentiation	9.13E-10	29/340 (8.5%)
GO:0008544 epidermis development	1.01E-09	29/340 (8.5%)
GO:0071554 cell wall organization or biogenesis	2.76E-08	38/340 (11.2%)
GO:0007043 cell-cell junction assembly	2.05E-07	5/340 (1.5%)
GO:0048468 cell development	5.83E-07	30/340 (8.8%)
GO:0034329 cell junction assembly	1.22E-06	5/340 (1.5%)
GO:0045216 cell-cell junction organization	4.22E-06	5/340 (1.5%)
GO:0070882 cellular cell wall organization	6.11E-06	29/340 (8.5%)
GO:0034330 cell junction organization	1.12E-05	5/340 (1.5%)
GO:0071555 cell wall organization	3.37E-05	25/340 (7.4%)
GO:0070592 cell wall polysaccharide biosynthesis	3.41E-04	13/340 (3.8%)
GO:0070589 cellular component macromolecule biosynthesis	3.41E-04	13/340 (3.8%)
GO:0009605 response to external stimulus	3.51E-04	28/340 (8.2%)
GO:0010382 cellular cell wall macromolecule metabolic process	4.05E-04	15/340 (4.4%)
GO:0010410 hemicellulose metabolic process	4.32E-04	13/340 (3.8%)
GO:0010383 cell wall polysaccharide metabolic process	5.67E-04	14/340 (4.1%)
GO:0010413 glucuronoxylan metabolic process	1.30E-03	12/340 (3.5%)
GO:0045492 xylan biosynthetic process	1.30E-03	12/340 (3.5%)
GO:0071248 cellular response to metal ion	4.60E-03	6/340 (1.8%)
GO:0048767 root hair elongation	9.87E-03	11/340 (3.2%)

Table 4.8b: Gene Ontology (GO) enrichment analysis of tm down-regulated genes compared to wt. The frequency of genes in different GO categories is listed in the table, showing the GO categories with significantly high p-values. The shaded GO categories are related to defense and cellular processes due to phosphate starvation. These two categories will be further explored to study the interaction of tm under similar physiological conditions.

GO Term	P-value	Sample frequency
GO:0009814 defense response, incompatible interaction	3.74E-30	55/362 (15.2%)
GO:0009627 systemic acquired resistance	1.08E-26	47/362 (13.0%)
GO:0045087 innate immune response	3.00E-26	62/362 (17.1%)
GO:0007243 intracellular protein kinase cascade	2.30E-24	34/362 (9.4%)
GO:0000165 MAPK cascade	3.11E-24	33/362 (9.1%)
GO:0080134 regulation of response to stress	1.06E-22	47/362 (13.0%)
GO:0031347 regulation of defense response	1.43E-22	46/362 (12.7%)
GO:0009751 response to salicylic acid stimulus	1.15E-21	43/362 (11.9%)
GO:0009697 salicylic acid biosynthetic process	2.15E-20	30/362 (8.3%)
GO:0031348 negative regulation of defense response	4.51E-20	33/362 (9.1%)
GO:0048583 regulation of response to stimulus	1.20E-19	50/362 (13.8%)
GO:0009696 salicylic acid metabolic process	1.30E-19	30/362 (8.3%)
GO:0009863 salicylic acid mediated signaling pathway	2.91E-19	36/362 (9.9%)
GO:0048585 negative regulation of response to stimulus	2.91E-18	35/362 (9.7%)
GO:0009862 systemic acquired resistance, salicylic acid signaling pathway	1.85E-17	30/362 (8.3%)
GO:0035556 intracellular signal transduction	2.33E-16	41/362 (11.3%)
GO:0045088 regulation of innate immune response	2.77E-16	35/362 (9.7%)
GO:0009753 response to jasmonic acid stimulus	3.14E-16	37/362 (10.2%)
GO:0002682 regulation of immune system process	3.76E-16	35/362 (9.7%)
GO:0009867 jasmonic acid mediated signaling pathway	2.35E-15	29/362 (8.0%)
GO:0010310 regulation of hydrogen peroxide metabolic process	8.29E-15	24/362 (6.6%)
GO:0009595 detection of biotic stimulus	4.13E-14	19/362 (5.2%)
GO:2000377 regulation of reactive oxygen species metabolic process	5.07E-14	24/362 (6.6%)
GO:0050832 defense response to fungus	4.11E-13	30/362 (8.3%)
GO:0008219 cell death	6.20E-13	34/362 (9.4%)
GO:0042742 defense response to bacterium	8.37E-13	30/362 (8.3%)
GO:0010363 regulation of plant-type hypersensitive response	2.41E-12	29/362 (8.0%)
GO:0009626 plant-type hypersensitive response	2.85E-12	30/362 (8.3%)
GO:0051606 detection of stimulus	4.00E-12	20/362 (5.5%)
GO:0080135 regulation of cellular response to stress	5.51E-12	29/362 (8.0%)
GO:0006612 protein targeting to membrane	1.07E-11	29/362 (8.0%)
GO:0012501 programmed cell death	1.20E-11	31/362 (8.6%)
GO:0043067 regulation of programmed cell death	1.58E-11	29/362 (8.0%)

……………(continued)

GO Term	P-value	Sample frequency
GO:0009620 response to fungus	6.23E-11	33/362 (9.1%)
GO:0002237 response to molecule of bacterial origin	1.22E-10	16/362 (4.4%)
GO:0042743 hydrogen peroxide metabolic process	4.06E-10	25/362 (6.9%)
GO:0072593 reactive oxygen species metabolic process	4.91E-10	26/362 (7.2%)
GO:0010200 response to chitin	2.32E-09	27/362 (7.5%)
GO:0009723 response to ethylene stimulus	2.48E-09	25/362 (6.9%)
GO:0002831 regulation of response to biotic stimulus	6.93E-09	15/362 (4.1%)
GO:0043900 regulation of multi-organism process	1.71E-08	15/362 (4.1%)
GO:0009581 detection of external stimulus	2.64E-08	12/362 (3.3%)
GO:0016036 cellular response to phosphate starvation	3.98E-08	17/362 (4.7%)
GO:0043069 negative regulation of programmed cell death	3.81E-07	16/362 (4.4%)
GO:0060548 negative regulation of cell death	4.53E-07	16/362 (4.4%)
GO:0009247 glycolipid biosynthetic process	8.59E-06	12/362 (3.3%)
GO:0071456 cellular response to hypoxia	1.83E-05	7/362 (1.9%)
GO:0006664 glycolipid metabolic process	2.17E-05	12/362 (3.3%)
GO:0071453 cellular response to oxygen levels	3.32E-05	7/362 (1.9%)
GO:0031669 cellular response to nutrient levels	6.08E-05	19/362 (5.2%)
GO:0031668 cellular response to extracellular stimulus	6.12E-05	20/362 (5.5%)
GO:0034976 response to endoplasmic reticulum stress	8.34E-05	19/362 (5.2%)
GO:0031667 response to nutrient levels	1.18E-04	19/362 (5.2%)
GO:0009991 response to extracellular stimulus	1.19E-04	20/362 (5.5%)
GO:0009267 cellular response to starvation	1.68E-04	18/362 (5.0%)
GO:0009743 response to carbohydrate stimulus	2.66E-04	29/362 (8.0%)
GO:0046467 membrane lipid biosynthetic process	3.52E-04	12/362 (3.3%)
GO:0009737 response to abscisic acid stimulus	6.16E-04	24/362 (6.6%)
GO:0006605 protein targeting	2.22E-03	29/362 (8.0%)
GO:0006643 membrane lipid metabolic process	2.23E-03	12/362 (3.3%)
GO:0016045 detection of bacterium	2.53E-03	5/362 (1.4%)
GO:0009873 ethylene mediated signaling pathway	4.38E-03	10/362 (2.8%)
GO:0001666 response to hypoxia	4.56E-03	9/362 (2.5%)
GO:0070482 response to oxygen levels	5.83E-03	9/362 (2.5%)

The *IRT1* gene encodes a transmembrane protein and produces two alternatively spliced isoforms. From the RT-PCR data (Figure 4.12a) the second isoform, the functional transcript is more abundant than isoform 1. Isoform 2 encodes a 347 aa protein with a ZIP zinc transporter (49-344aa) domain. Isoform 1 encodes a protein of 211 aa with a domain structure that has not been annotated. The expression of *IRT1* is known to be induced in Arabidopsis plants grown under iron deficiency. IRT1 is a metal transporter with a broad substrate range and helps in transport of mainly divalent cations (Korshunova et al., 1999). In *irt1* mutants, there is severe reduction in growth and fertility and the plants are chlorotic. The photosynthetic machinery is also perturbed in mutants with significant alteration in photosensitivity and chlorophyll fluorescence parameters. Overexpression of IRT1 accumulates higher levels of cadmium and zinc than wild-type plants under Fe-deficiency conditions, indicating that IRT1 is responsible for the uptake of these metals (Connolly et al., 2002).

The *FRO2* gene, encoding NADPH- dependent ferric reductase in the plasma membrane, is induced in the triple mutant and it was previously shown that low iron induces expression of this gene (Connolly et al., 2003). Similarly the *BHLH038, BHLH039, BHLH100* genes encoding basic helix-loop-helix transcription factors, that are induced under low iron, are up-regulated in the triple mutant (Wang et al., 2007b). The cytochrome P450 gene (CYP82C4) is highly expressed in the triple mutant and a recent study showed induction of CYP82C4 in Fe-deficient Arabidopsis seedlings through a FIT-dependent pathway (Murgia et al., 2011). *AtNAS2,* which was upregulated in the mutant, responds to zinc deficiency. Furthermore, co-overexpression of FIT with AtbHLH38 or AtbHLH39 enhanced the expression of NAS1 and NAS2, resulting in the accumulation of nicotinamide, a chelator for iron transport and homeostasis (Wu et al., 2012).

Up-regulation of several genes that are know to be induced by iron/metal deficiency in the triple mutant suggest a role for these three SRs in regulation of genes involved in iron homeostasis.

The second category of overrepresented differentially expressed genes in the mutant are involved in the regulation of phosphate uptake (Table 4.8b). These include MIR399c (At5g62162) and MIR399d (At2g34202) along with the transcription factor SPX3 (At2g45130). Unlike the genes involved in iron homeostasis, these genes are down-regulated in the mutant. Interestingly, one of the targets of these microRNAs, PHO2 (UBC24), a ubiquitin E2 conjugase gene, was upregulated in the mutant, which could be a consequence of down-regulation of the MIR399 microRNAs. This action could limit the activity of phosphate transporters at the post-translational level and, hence, influence its uptake. This regulation of Pi homeostasis by MiR399s has been well-established for regulating the expression of PHO2 and is mainly dependent on the availability of phosphate ions (Lin et al., 2008; Hsieh et al., 2009). Analysis of phenotypes under different metal and/or phosphate levels in the medium may provide insights into the roles of these SRs in iron and phosphate homeostasis.

Another major functional category of overrepresented differentially regulated genes in the triple mutant is involved in pathogenesis. These genes are mainly down-regulated in the triple mutant and are likely to modulate responses to plant pathogens (Table 4.8b). Several down-regulated genes in the mutant encode positive regulators of plant immunity (Mukhtar et al., 2011), suggesting that the triple mutant may show enhanced susceptibility to pathogens. Experiments are underway to test this prediction.

Mechanisms of gene regulation by SRs

The observed changes in gene expression in the triple mutants may be the result of multiple mechanisms (Reddy and Ali, 2011; Twyffels et al., 2011; Ausin et al., 2012; Risso et al., 2012). These include: i) lack of SRs changes alternative splicing of target genes directly thereby generating various isoforms that render them stable or make them unstable and so targeting them for degradation by nonsense mediated decay or other degradation pathways, ii) Pre-mRNAs of SRs are extensively alternatively spliced and there are well-documented examples of cross-regulation of splicing among *SRs*. Hence, the observed effect could be due to alteration of splicing patterns of other SRs, which in turn regulate splicing iii) Loss of SRs may change the splicing pattern of transcription factors and produce proteins with altered functions in transcription of downstream genes. In our study, expression of transcription factors such as the *BHLH* genes in Fe homeostasis and the *SPX3* in phosphate homeostasis are changed, thereby affecting the expression of downstream genes involved in these biological processes. iv) SRs may affect transcription of genes through changes in chromatin modification leading to epigenetic effects. For instance, an SR-like protein, SR45, in Arabidopsis was recently shown to be involved in DNA methylation (Ausin et al., 2012) and v) SRs may regulate biogenesis of microRNA or noncoding RNA thereby affecting regulation of target genes. A recent study suggests a role for an SR proteins in miRNA biogenesis (Wu et al., 2010). Interestingly, 27% of differentially expressed genes (i.e., 199 out 729) in our study do not have any introns. Hence, regulation of expression of these genes is indirect e.g., splicing of a transcription factors that regulate these genes is altered, or direct affects of SRs on transcription.

The observed changes in gene expression here are due to loss of three SR proteins. To identify direct and indirect targets of each SR and a combination of these SRs, RNA-Seq studies with all three single mutants, and three possible combinations of double mutants are needed. Such analyses will also help us identify synergistic, additive and/or antagonistic roles of SR in regulating gene expression. These studies coupled with global RNA binding studies using PAR-CLIP should shed light on direct and indirect targets of each of these SR proteins. Similar studies with all SRs in Arabidopsis will allow us to construct splicing regulatory networks of SR proteins in regulating gene expression.

Material and Methods

Generation of mutant genotypes of Arabidopsis SR genes

Generation and verification of the triple mutant is described in Chapter 2.

Generation of RNA-seq data

Two-week old seedlings of the triple mutant and wild-type for RNA-Seq analysis. Seedlings were grown on MS agar plates with 1% sucrose pH 5.7, in growth chambers under long-day conditions (16 hrs light and 8hrs dark; 70-80 μmol/m^2/s^2 light intensity, 22^{0}C). Two biological replicates were used for all RNA-Seq and RT-PCR experiments. Total RNA was extracted and purified using the RNAeasy Plant Mini kit (Qiagen). On-column DNase digestion was performed according to the manufacturer's protocol (Qiagen). The integrity and quality of the total RNA was checked by NanoDrop 1000 Spectrophotometer and formaldehyde-agarose gel electrophoresis. Poly (A) RNA was isolated from total RNA using oligo-dT beads, and randomly fragmented under elevated temperature (80^{0}C). The method we used in our study eliminates bias in reads toward the 3’ end of the transcript. Preparation of cDNA libraries and sequencing of cDNAs was done at Duke University using the Illumina TrueSeq RNA kit. First strand cDNA was synthesized using random primers and reverse transcriptase Superscript RTII.

Second strand cDNA synthesis was done using DNA Polymerase I and RNaseH, the cDNA fragments processed for end repair, and ligation of the adapters. These products were then purified and enriched by PCR to create the final cDNA library and sequenced on the Illumina Genome Analyser IIx to generate single end reads of 75 nt.

Processing of Illumina Reads

The RNA-Seq reads generated by Illumina GAIIx were initially processed to remove the adapter sequences and low quality bases at the 3' end. After preprocessing the RNA-Seq data, the quality of reads was checked by FASTQC. All the reads were then mapped to the Arabidopsis TAIR 10 genome and the pipelines used are described in Figures 4.5 and 4.6. SpliceGrapher was used to visualize isoforms (Rogers et al., 2012). The SAM (Sequence Alignment/Map) files generated by Tophat were provided as input to the software Cufflinks as shown in Figure 4.5. The class codes in the Cuffmerge output were used to identify novel isoforms and intergenic transcripts. For GO enrichment analysis, list of differentially expressed genes were analyzed using Gene Ontology Enrichment Analysis Software Toolkit (GOEAST; http://omicslab.genetics.ac.cn/GOEAST/).

Validation of differentially expressed genes by RT-PCR

One µg RNA from wt and triple mutant that was used for transcritpome analysis was used later to synthesize first-strand cDNA using Superscript II reverse transcriptase (Invitrogen, USA) and 1 µl of the first-strand cDNA was used for PCR in a reaction volume of 20 µl with gene-specific primers. Most gene-specific primers were made to the first and last exons.

CONCLUSIONS

During the last decade post-transcriptional processing of pre-mRNAs has emerged as an important and pervasive mechanism of regulation of gene expression. Splicing of pre-mRNAs, one of the key steps in pre-mRNA processing, is dependent on hundreds of RNA-binding proteins that recognize sequence signals in pre-mRNA and regulate both constitutive and alternative splicing (Long and Caceres, 2009; Wahl et al., 2009). The serine/arginine (SR)-rich proteins are a family of RNA binding splicing factors that have essential functions during pre-mRNA splicing. These proteins bind to splicing regulatory sequences in pre-mRNAs and regulate splice site choice during both constitutive and alternative splicing. Recognition of weak splice sites that are important in regulated splicing is also accomplished by SR proteins. The functions of the SR family of proteins have been well studied in the mammalian system (Long and Caceres, 2009). This is primarily because of the availability of an *in vitro* splicing assay and the implications of altered splicing in many human diseases.

Differences in gene architecture and in the prevalence of types of splicing events exist between humans and plants, suggesting that some splicing regulatory mechanisms may be unique to plants. The fact that plant introns cannot be spliced accurately in animal systems also points to differences splicing regulation between plants animals (Schuler, 2008). The lack of a plant-derived *in vitro* splicing system hindered the progress in elucidating the roles of SR proteins in splicing. Consequently, we know very little about the roles of plant SRs in pre-mRNA splicing, and plant growth and development. As compared to mammalian systems, the SR family is considerably expanded with many plant-specific SRs in flowering plants (Richardson et al., 2011). This raises the question of whether the SRs in plants that arose because of gene-duplications have novel or redundant functions and if plant-specific SRs perform functions that are unique to plants. The pre-mRNAs of Arabidopsis SR genes undergo extensive alternative splicing giving rise to about 100 transcripts, thereby increasing the transcriptome complexity of SRs by about six fold (Palusa et al., 2007). Since SR genes regulate their own splicing and subsequently affect AS of other genes (Reddy, 2007), SRs are a good experimental model to dissect components of the splicing processes in plants. Thus, a study of the mechanism of AS controlled by SRs in plants is important to provide answers to the differences in splicing between plants and animals.

To assess the functional role of SRs, overexpression studies have been done, but these do not address the natural function of these genes in plants because pre-mRNA splicing is known to be affected by the relative amounts of SR proteins (Long and Caceres, 2009). In my research I have used comprehensive genetic, molecular and cell biological approaches using knockout mutants of *SR* genes to address the roles of three SR proteins in Arabidopsis. Since the SR family is expanded in plants with many paralogous genes with potential redundant functions, a strategy using a combination of multiple gene knockouts was designed to answer a number of biological questions about SR gene functions. This research focused on the functional analysis of three SR proteins, which were selected to address functional redundancy. SC35 is the sole member of the SC subfamily and two paralogous members of the SCL (SC35-like) family (SCL33 and SCL30a) are plant specific and share similarity to the RRM domain of SC35 (Barta et al., 2010).

This genetic strategy provided an insight into novel functions of SRs in controlling flowering time. Although lack of some SRs including SC35 was found to be lethal in mammals, in plants loss of SC35 alone or in combination with two related SCLs (30a and 33) did not lead to lethality. The loss of the individual paralogous genes, *SCL33* or *SCL30a,* resulted in opposite mutant phenotypes, namely late and early flowering. However, the combination of these two mutants showed enhanced late flowering signifying an epistatic effect of the *scl33* mutant over its paralog. This is an interesting phenomenon and suggests complex interactions between genes controlling flowering in different pathways. The loss of SC35, the sole member of the SC subclass, with early flowering phenotype, in combination with the loss of SCL30a also resulted in early flowering phenotype but no additive effect was observed suggesting that both these SRs function in the same pathway. The complexity of these interactions could be caused by the multiplicity of splice variants generated by autoregulation and/or cross-regulation of splicing of other genes including other SRs. The change in amounts of specific SRs in relation to each other could also alter the splicing patterns of different genes. In any specific cell type there would be different amounts and forms of SR proteins, and a mutation in any of the genes could change the balance of the SR proteins, resulting in a shift in the balance between the proteins. Triple knockout mutant (*scl33 scl30a sc35*) plants are viable, but with a more pronounced shift in flowering phenotype. The viability of triple mutant suggests that two other SCL members may have some functional overlap with these SRs. It appears that interactions among SRs allow plasticity in their function, where each mutant can affect the process of flowering through different pathways or regulatory mechanisms. Flowering time is an important adaptive process in plants, where the ability to flower at the right time and set seed is important for the survival of plants in response to environmental and endogenous cues (Amasino, 2010). Even though there are many genes that control the flowering pathways, the SRs whose splicing pattern is known to be affected by environmental factors also play a role in flowering time.

Based on the small size of introns in plants as compared to metazoans it is proposed that the pre-mRNA splicing occurs through the intron definition model in which *cis*-elements in introns play an important role in spliceosome assembly. In mammals, due to their large introns, the exon-definition model is thought be involve in pre-mRNA splicing. Like most plant SRs, *SCL33* has a large intron that produces numerous splice variants and became the subject of the study to address if this intron alone has the necessary signals for it to undergo AS, and if any of the gene products of *SCL33*, *SCL30a*, *SC35* alone or in combination are responsible for AS.

The *SCL33* intron was shown to have the necessary signals to undergo AS and produce 8 transcripts similar to the endogenous *SCL33* gene, suggesting that no other signals reside in other parts of the gene for AS. My results with the *SCL33* 3rd intron do support the intron definition model. Similar studies with several other genes could address the prevalence of intron definition in plants. I used the mutants that I generated together with biochemical and cell biological studies to address *trans*-acting SRs involved in regulating AS of this intron. The SCL33 protein binds to a 92bp segment of its own intron, which has four GAAG splicing regulatory elements and likely recruits U1SnRNP to the 5’ splice site for regulating splicing of three isoforms that share the same 5’ splice site. These three isoforms, which may form truncated proteins, have signals to undergo NMD and are most likely regulated at the post-transcriptional level rather than transcriptional shut down of its expression. The misregulation of these isoforms is also dependent on absence of the paralogous SCL33 and SCL30a proteins, suggesting a redundant role in maintaining optimal expression of the isoforms (Thomas et al., 2012). The novel *in vivo* splicing reporter assay we developed can be used to study alternative splicing of other genes and to identify other splicing regulatory elements and *trans*-acting splicing factors. My studies on the role of these SRs in flowering and alternative splicing *SCL33* 3rd intron highlight the functions of SRs and show that these can have novel and redundant functions depending on different biological pathways.

To understand the function of SRs as a master regulator of CS and AS, a global transcriptome analysis using RNA-Seq was conducted in the SR gene triple mutant genotype. This provided transcriptome-wide changes in gene expression and permitted analysis of AS of all gene expressed in seedlings with qualitative and quantitative differences in the mutant as compared to the wild type. Analysis of the GO categories of differentially expressed genes for overrepresented categories led to identification of biological processes that are likely affected in these mutants. Based on these results additional phenotypic screens can be designed.

The global transcriptome analysis revealed that the SR proteins are master regulators coordinating cascades of genes of different biological pathways. The triple mutant showed increased expression of genes involved in iron homeostasis and reduced expression of genes in phosphorus homeostasis, two pathways that are known to be functionally co-regulated in opposite directions. The set of differentially expressed/spliced genes that we identified here represents direct and indirect targets of these SR proteins. Similar studies with all other mutants will allow identification of direct and indirect targets of other SRs and determine if there any common targets for different SRs. Identification of direct targets of each of the SRs using methods such as PAR-CLIP (Hafner et al., 2010b) will help us not only to identify which of these are indirect targets but also pave the way to use computational tools to identify potential splicing regulatory elements in direct targets and formulate a mechanisms of intron recognition and AS by the SR proteins.

REFERENCES

Ai, X.Y., Lin, G., Sun, L.M., Hu, C.G., Guo, W.W., Deng, X.X., and Zhang, J.Z. (2012). A global view of gene activity at the flowering transition phase in precocious trifoliate orange and its wild-type [Poncirus trifoliata (L.) Raf.] by transcriptome and proteome analysis. Gene.

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

Ali, G.S., and Reddy, A.S. (2008a). Spatiotemporal organization of pre-mRNA splicing proteins in plants. Curr Top Microbiol Immunol **326,** 103-118.

Ali, G.S., and Reddy, A.S. (2008b). Regulation of alternative splicing of pre-mRNAs by stresses. Curr Top Microbiol Immunol **326,** 257-275.

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

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

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

Allo, M., Buggiano, V., Fededa, J.P., Petrillo, E., Schor, I., de la Mata, M., Agirre, E., Plass, M., Eyras, E., Elela, S.A., Klinck, R., Chabot, B., and Kornblihtt, A.R. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol **16,** 717-724.

Amasino, R. (2010). Seasonal and developmental timing of flowering. The Plant journal **61,** 1001-1013.

Ambrosone, A., Costa, A., Leone, A., and Grillo, S. (2012). Beyond transcription: RNA-binding proteins as emerging regulators of plant response to environmental constraints. Plant Sci **182,** 12-18.

Anko, M.L., Muller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome biology **13,** R17.

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

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

Baek, J.M., Han, P., Iandolino, A., and Cook, D.R. (2008). Characterization and comparison of intron structure and alternative splicing between Medicago truncatula, Populus trichocarpa, Arabidopsis and rice. Plant Molecular Biology **67,** 499-510.

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. PLoS Genet **2,** e106.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature **465,** 53-59.

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

Batey, R.T. (2012). Structure and mechanism of purine-binding riboswitches. Quarterly reviews of biophysics **45,** 345-381.

Beilharz, T.H., and Preiss, T. (2004). Translational profiling: the genome-wide measure of the nascent proteome. Briefings in functional genomics & proteomics **3,** 103-111.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J Biol Chem **270,** 2411-2414.

Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H., and Luhrmann, R. (2008). Isolation of an active step I spliceosome and composition of its RNP core. Nature **452,** 846-850.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem **72,** 291-336.

Blencowe, B.J., and Ouzounis, C.A. (1999). The PWI motif: a new protein domain in splicing factors. Trends Biochem Sci **24,** 179-180.

Blencowe, B.J., Issner, R., Nickerson, J.A., and Sharp, P.A. (1998). A coactivator of pre-mRNA splicing. Genes & Dev **12,** 996-1009.

Bocobza, S., Adato, A., Mandel, T., Shapira, M., Nudler, E., and Aharoni, A. (2007). Riboswitch-dependent gene regulation and its evolution in the plant kingdom. Genes and Development **21,** 2874-2879.

Boss, P.K., Bastow, R.M., Mylne, J.S., and Dean, C. (2004). Multiple Pathways in the Decision to Flower: Enabling, Promoting, and Resetting. Plant Cell **16,** S18-S31.

Bourgeois, C.F., Lejeune, F., and Stevenin, J. (2004). Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog Nucleic Acid Res Mol Biol **78,** 37-88.

Bove, J., Kim, C.Y., Gibson, C.A., and Assmann, S.M. (2008). Characterization of wound-responsive RNA-binding proteins and their splice variants in Arabidopsis. Plant molecular biology **67,** 71-88.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R., and Gorlach, J. (2001). Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell **13,** 1499-1510.

Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E., and Graveley, B.R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. Genome research **21,** 193-202.

Brown, J.W., Simpson, C.G., Thow, G., Clark, G.P., Jennings, S.N., Medina-Escobar, N., Haupt, S., Chapman, S.C., and Oparka, K.J. (2002). Splicing signals and factors in plant intron removal. Biochem Soc Trans **30,** 146-149.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Molecular and Cellular Biology **17,** 3194-3201.

Buratti, E., Stuani, C., De Prato, G., and Baralle, F.E. (2007). SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. Nucleic acids research **35,** 4359-4368.

Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdiscip Rev RNA **3,** 1-12.

Caceres, J.F., and Krainer, A.R. (1993). Functional analysis of pre-mRNA splicing factor SF2/ASF structural domains. Embo J **12,** 4715-4726.

Carle-Urioste, J.C., Brendel, V., and Walbot, V. (1997). A combinatorial role for exon, intron and splice site sequences in splicing in maize. Plant J. **11,** 1253-1263.

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.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stévenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA **5,** 468-483.

Chasin, L.A. (2007). Searching for splicing motifs. Advances in Experimental Medicine and Biology **623,** 85-106.

Cheah, M.T., Wachter, A., Sudarsan, N., and Breaker, R.R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature **447,** 497-500.

Chen, F.C., Wang, S.S., Chaw, S.M., Huang, Y.T., and Chuang, T.J. (2007). Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species. Plant Physiol **143,** 1086-1095.

Chiu, S.Y., Lejeune, F., Ranganathan, A.C., and Maquat, L.E. (2004). The pioneer translation initiation complex is functionally distinct from but structurally overlaps with the steady-state translation initiation complex. Genes & development **18,** 745-754.

Choi, J., Hyun, Y., Kang, M.J., In Yun, H., Yun, J.Y., Lister, C., Dean, C., Amasino, R.M., Noh, B., Noh, Y.S., and Choi, Y. (2009). Resetting and regulation of Flowering Locus C expression during Arabidopsis reproductive development. The Plant journal : for cell and molecular biology **57,** 918-931.

Chusainow, J., Ajuh, P.M., Trinkle-Mulcahy, L., Sleeman, J.E., Ellenberg, J., and Lamond, A.I. (2005). FRET analyses of the U2AF complex localize the U2AF35/U2AF65 interaction in vivo and reveal a novel self-interaction of U2AF35. RNA **11,** 1201-1214.

Connolly, E.L., Fett, J.P., and Guerinot, M.L. (2002). Expression of the IRT1 metal transporter is controlled by metals at the levels of transcript and protein accumulation. The Plant cell **14,** 1347-1357.

Connolly, E.L., Campbell, N.H., Grotz, N., Prichard, C.L., and Guerinot, M.L. (2003). Overexpression of the FRO2 ferric chelate reductase confers tolerance to growth on low iron and uncovers posttranscriptional control. Plant Physiol **133,** 1102-1110.

Coulter, L.R., Landree, M.A., and Cooper, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. Molecular and cellular biology **17,** 2143-2150.

Croft, M.T., Moulin, M., Webb, M.E., and Smith, A.G. (2007). Thiamine biosynthesis in algae is regulated by riboswitches. Proc Natl Acad Sci U S A **104,** 20770-20775.

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

Das, D., Clark, T.A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J.E., and Conboy, J.G. (2007). A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic acids research **35,** 4845-4857.

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

Deng, X.-W., Matsui, M., Wei, N., Wagner, D., Chu, A.M., Feldmann, A., and Quail, P. (1992). COPI, an arabidopsis regulatory gene encodes a protein with both a zinc binding motiff and a G-beta homologous domain. Cell **71,** 791-801.

Ding, J.H., Xu, X., Yang, D., Chu, P.H., Dalton, N.D., Ye, Z., Yeakley, J.M., Cheng, H., Xiao, R.P., Ross, J., Chen, J., and Fu, X.D. (2004). Dilated cardiomyopathy caused by tissue-specific ablation of SC35 in the heart. The EMBO journal **23,** 885-896.

Dreumont, N., Hardy, S., Behm-Ansmant, I., Kister, L., Branlant, C., Stevenin, J., and Bourgeois, C.F. (2010). Antagonistic factors control the unproductive splicing of SC35 terminal intron. Nucleic acids research **38,** 1353-1366.

Du, L., Ali, G.S., Simons, K.A., Hou, J., Yang, T., Reddy, A.S., and Poovaiah, B.W. (2009). Ca(2+)/calmodulin regulates salicylic-acid-mediated plant immunity. Nature **457,** 1154-1158.

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

Eissmann, M., Gutschner, T., Hammerle, M., Gunther, S., Caudron-Herger, M., Gross, M., Schirmacher, P., Rippe, K., Braun, T., Zornig, M., and Diederichs, S. (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA biology **9**.

Ellis, J.D., Lleres, D., Denegri, M., Lamond, A.I., and Caceres, J.F. (2008). Spatial mapping of splicing factor complexes involved in exon and intron definition. J. Cell Biol. **181,** 921-934.

Eshar, S., Allemand, E., Sebag, A., Glaser, F., Muchardt, C., Mandel-Gutfreund, Y., Karni, R., and Dzikowski, R. (2012). A novel Plasmodium falciparum SR protein is an alternative splicing factor required for the parasites' proliferation in human erythrocytes. Nucleic acids research.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science **297,** 1007-1013.

Felber, B.K., Orkin, S.H., and Hamer, D.H. (1982). Abnormal RNA splicing causes one form of alpha thalassemia. Cell **29,** 895-902.

Felsenfeld, G. (1992). Chromatin as an essential part of the transcriptional mechanism. Nature **355,** 219-224.

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.

Filipowicz, W., Gniadkowski, M., Klahre, U., and Liu, H.-X. (1995). Pre-mRNA splicing in plants **4,** 65-77.

Fischer, D.C., Noack, K., Runnebaum, I.B., Watermann, D.O., Kieback, D.G., Stamm, S., and Stickeler, E. (2004). Expression of splicing factors in human ovarian cancer. Oncology reports **11,** 1085-1090.

Fox, S., Filichkin, S., and Mockler, T.C. (2009). Applications of ultra-high-throughput sequencing. Methods in molecular biology **553,** 79-108.

Fukumura, K., Kato, A., Jin, Y., Ideue, T., Hirose, T., Kataoka, N., Fujiwara, T., Sakamoto, H., and Inoue, K. (2007). Tissue-specific splicing regulator Fox-1 induces exon skipping by interfering E complex formation on the downstream intron of human F1gamma gene. Nucleic acids research **35,** 5303-5311.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Ratsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature **477,** 419-423.

Garcia-Blanco, M.A., Baraniak, A.P., and Lasda, E.L. (2004). Alternative splicing in disease and therapy. Nat Biotechnol **22,** 535-546.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. Nat Rev Mol Cell Biol **8,** 113-126.

Gassmann, W. (2008). Alternative splicing in plant defense. In Nuclear pre-mRNA processing in plants, A.S.N.R.a.M. Golovkin, ed (Heidelberg: Springer-Verlag), pp. 219-233.

Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N., Park, C.C., Wen, P.Z., Brewer, H., Weitz, K., Camp, D.G., 2nd, Pan, C., Yordanova, R., Neuhaus, I., Tilford, C., Siemers, N., Gargalovic, P., Eskin, E., Kirchgessner, T., Smith, D.J., Smith, R.D., and Lusis, A.J. (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS genetics **7,** e1001393.

Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett **582,** 1977-1986.

Golovkin, M., and Reddy, A.S. (1996). Structure and expression of a plant U1 snRNP 70K gene: alternative splicing of U1 snRNP 70K pre-mRNAs produces two different transcripts. Plant Cell **8,** 1421-1435.

Golovkin, M., and Reddy, A.S. (1998). The plant U1 small nuclear ribonucleoprotein particle 70K protein interacts with two novel serine/arginine-rich proteins. Plant Cell **10,** 1637-1648.

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.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (1998). A systematic analysis of the factors that determine the strength of pre- mRNA splicing enhancers. EMBO J **17,** 6747-6756.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., Brown, J.B., Cherbas, L., Davis, C.A., Dobin, A., Li, R., Lin, W., Malone, J.H., Mattiuzzo, N.R., Miller, D., Sturgill, D., Tuch, B.B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R.E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J.E., Wan, K.H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P.J., Brenner, S.E., Brent, M.R., Cherbas, P., Gingeras, T.R., Hoskins, R.A., Kaufman, T.C., Oliver, B., and Celniker, S.E. (2011). The developmental transcriptome of Drosophila melanogaster. Nature **471,** 473-479.

Gudikote, J.P., Imam, J.S., Garcia, R.F., and Wilkinson, M.F. (2005). RNA splicing promotes translation and RNA surveillance. Nature structural & molecular biology **12,** 801-809.

Gulledge, A.A., Roberts, A.D., Vora, H., Patel, K., and Loraine, A.E. (2012). Mining Arabidopsis thaliana RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. Am J Bot **99,** 219-231.

Gullerova, M., Barta, A., and Lorkovic, Z.J. (2006). AtCyp59 is a multidomain cyclophilin from Arabidopsis thaliana that interacts with SR proteins and the C-terminal domain of the RNA polymerase II. RNA **12,** 631-643.

Gupta, S., Zink, D., Korn, B., Vingron, M., and Haas, S.A. (2004). Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics **20,** 2579-2585.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., and Tuschl, T. (2010a). PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp **41,** pii: 2034.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., and Tuschl, T. (2010b). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell **141,** 129-141.

Hatakeyama, S., Sugihara, K., Nakayama, J., Akama, T.O., Wong, S.M., Kawashima, H., Zhang, J., Smith, D.F., Ohyama, C., Fukuda, M., and Fukuda, M.N. (2009). Identification of mRNA splicing factors as the endothelial receptor for carbohydrate-dependent lung colonization of cancer cells. Proceedings of the National Academy of Sciences of the United States of America **106,** 3095-3100.

Hayden, E.C. (2010). Life is complicated. Natuer **464,** 664-667.

Holcik, M., and Pestova, T.V. (2007). Translation mechanism and regulation: old players, new concepts. Meeting on translational control and non-coding RNA. EMBO Rep **8,** 639-643.

Hsieh, L.C., Lin, S.I., Shih, A.C., Chen, J.W., Lin, W.Y., Tseng, C.Y., Li, W.H., and Chiou, T.J. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol **151,** 2120-2132.

Huang, Y., and Steitz, J.A. (2005). SRprises along a messenger's journey. Mol Cell **17,** 613-615.

Hugouvieux, V., Kwak, J.M., and Schroeder, J.I. (2001). An mRNA cap binding protein, ABH1, modulates early abscisic acid signal transduction in Arabidopsis. Cell **106,** 477-487.

Hui, J., Hung, L.H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. The EMBO journal **24,** 1988-1998.

Hull, J., Campino, S., Rowlands, K., Chan, M.S., Copley, R.R., Taylor, M.S., Rockett, K., Elvidge, G., Keating, B., Knight, J., and Kwiatkowski, D. (2007). Identification of common genetic variation that modulates alternative splicing. PLoS genetics **3,** e99.

Iida, K., and Go, M. (2006). Survey of Conserved Alternative Splicing Events of mRNAs Encoding SR Proteins in Land Plants. Mol Biol Evol **23,** 1085-1094.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2004). Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. Nucleic Acids Res **32,** 5096-5103.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell **147,** 789-802.

Isaacs, F.J., Dwyer, D.J., and Collins, J.J. (2006). RNA synthetic biology. Nat Biotechnol **24,** 545-554.

Isken, O., Kim, Y.K., Hosoda, N., Mayeur, G.L., Hershey, J.W., and Maquat, L.E. (2008). Upf1 phosphorylation triggers translational repression during nonsense-mediated mRNA decay. Cell **133,** 314-327.

Isshiki, M., Tsumoto, A., and Shimamoto, K. (2006). The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. Plant Cell **18,** 146-158.

Izquierdo, J.M., Majos, N., Bonnal, S., Martinez, C., Castelo, R., Guigo, R., Bilbao, D., and Valcarcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Molecular cell **19,** 475-484.

James, A.B., Syed, N.H., Brown, J.W., and Nimmo, H.G. (2012). Thermoplasticity in the Plant Circadian Clock: How Plants tell the Time-perature. Plant signaling & behavior **7**.

Jelen, N., Ule, J., Zivin, M., and Darnell, R.B. (2007). Evolution of Nova-dependent splicing regulation in the brain. PLoS genetics **3,** 1838-1847.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302,** 2141-2144.

Jumaa, H., and Nielsen, P.J. (1997). The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. EMBO J. **16,** 5077-5085.

Jumaa, H., Guenet, J.L., and Nielsen, P.J. (1997). Regulated expression and RNA processing of transcripts from the Srp20 splicing factor gene during the cell cycle. Mol Cell Biol **17,** 3116-3124.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell **12,** 5-14.

Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. RNA **8,** 426-439.

Kaldis, A., Tsementzi, D., Tanriverdi, O., and Vlachonasios, K.E. (2011). Arabidopsis thaliana transcriptional co-activators ADA2b and SGF29a are implicated in salt stress responses. Planta **233,** 749-762.

Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet **12,** 715-729.

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. Mol. Biol. Cell **14,** 3565-3577.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., Dinh, H.Q., Barta, A., and Brown, J.W. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. Nucleic Acids Res **40,** 2454-2469.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2012). Function of alternative splicing. Gene.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature **467,** 103-107.

Kim, C.Y., Bove, J., and Assmann, S.M. (2008). Overexpression of wound-responsive RNA-binding proteins induces leaf senescence and hypersensitive-like cell death. The New phytologist **180,** 57-70.

Kobayashi, T., and Nishizawa, N.K. (2012). Iron uptake, translocation, and regulation in higher plants. Annu Rev Plant Biol **63,** 131-152.

Koncz, C., Dejong, F., Villacorta, N., Szakonyi, D., and Koncz, Z. (2012). The spliceosome-activating complex: molecular mechanisms underlying the function of a pleiotropic regulator. Front Plant Sci **3,** 9.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology **17,** 909-915.

Koornneef, M., Hanhart, C.J., and van der Veen, J.H. (1991). A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. Mol Gen Genet **229,** 57-66.

Kopelman, N.M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet **37,** 588-589.

Korshunova, Y.O., Eide, D., Clark, W.G., Guerinot, M.L., and Pakrasi, H.B. (1999). The IRT1 protein from Arabidopsis thaliana is a metal transporter with a broad substrate range. Plant molecular biology **40,** 37-44.

Kramer, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu. Rev. Biochem. **65,** 367-409.

Kuhn, J.M., Hugouvieux, V., and Schroeder, J.I. (2008). mRNA cap binding proteins: effects on abscisic acid signal transduction, mRNA processing, and microarray analyses. Current topics in microbiology and immunology **326,** 139-150.

Labadorf, A., Link, A., Rogers, M.F., Thomas, J., Reddy, A.S., and Ben-Hur, A. (2010). Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. BMC Genomics **11,** 114.

Lam, B.J., and Hertel, K.J. (2002). A general role for splicing enhancers in exon definition. RNA **8,** 1233-1241.

Lambermon, M.H., Fu, Y., Kirk, D.A., Dupasquier, M., Filipowicz, W., and Lorkovic, Z.J. (2002). UBA1 and UBA2, two proteins that interact with UBP1, a multifunctional effector of pre-mRNA maturation in plants. Mol Cell Biol **22,** 4346-4357.

Lareau, L.F., Brooks, A.N., Soergel, D.A., Meng, Q., and Brenner, S.E. (2007a). The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv Exp Med Biol **623,** 190-211.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007b). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature **446,** 926-929.

Le Guiner, C., Lejeune, F., Galiana, D., Kister, L., Breathnach, R., Stevenin, J., and Del Gatto-Konczak, F. (2001). TIA-1 and TIAR activate splicing of alternative exons with weak 5' splice sites followed by a U-rich stretch on their own pre-mRNAs. The Journal of biological chemistry **276,** 40638-40646.

Le Hir, H., and Andersen, G.R. (2008). Structural insights into the exon junction complex. Current opinion in structural biology **18,** 112-119.

Le Hir, H., Moore, M.J., and Maquat, L.E. (2000). Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev **14,** 1098-1108.

Lemaire, R., Prasad, J., Kashima, T., Gustafson, J., Manley, J.L., and Lafyatis, R. (2002). Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. Genes Dev. **16,** 594-607.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A **100,** 189-192.

Lewis, J.D., and Izaurralde, E. (1997). The role of the cap structure in RNA processing and nuclear export. European journal of biochemistry / FEBS **247,** 461-469.

Li, J., Kinoshita, T., Pandey, S., Ng, C.K., Gygi, S.P., Shimazaki, K., and Assmann, S.M. (2002). Modulation of an RNA-binding protein by abscisic-acid-activated protein kinase. Nature **418,** 793-797.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature **456,** 464-469.

Lin, S.I., Chiang, S.F., Lin, W.Y., Chen, J.W., Tseng, C.Y., Wu, P.C., and Chiou, T.J. (2008). Regulatory network of microRNA399 and PHO2 by systemic signaling. Plant Physiol **147,** 732-746.

Lockhart, S.R., and Rymond, B.C. (1994). Commitment of yeast pre-mRNA to the splicing pathway requires a novel U1 small nuclear ribonucleoprotein polypeptide, Prp39p. Molecular and Cellular Biology **14,** 3623-3633.

Long, J.C., and Caceres, J.F. (2009). The SR protein family of splicing factors: master regulators of gene expression. Biochem. J. **417,** 15-27.

Longman, D., Johnstone, I.L., and Caceres, J.F. (2000). Functional characterization of SR and SR-related genes in *Caenorhabditis elegans*. EMBO J. **19,** 1625-1637.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J., and Barta, A. (1999a). A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. Plant Molecular Biology **39,** 761-773.

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

Lopato, S., Borisjuk, L., Milligan, A.S., Shirley, N., Bazanova, N., Parsley, K., and Langridge, P. (2006). Systematic identification of factors involved in post-transcriptional processes in wheat grain. Plant Molecular Biology **62,** 637-653.

Lopato, S., Forstner, C., Kalyna, M., Hilscher, J., Langhammer, U., Indrapichate, K., Lorkovic, Z.J., and Barta, A. (2002). Network of interactions of a novel plant-specific Arg/Ser-rich protein, atRSZ33, with atSC35-like splicing factors. J. Biol. Chem. **277,** 39989-39998.

Lorkovic, Z.J. (2009). Role of plant RNA-binding proteins in development, stress response and genome organization. Trends Plant Sci **14,** 229-236.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Huang, X., and Han, B. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Research **20,** 1238-1249.

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.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS biology **8,** e1000506.

Mai, Y.X., Wang, L., and Yang, H.Q. (2011). A gain-of-function mutation in IAA7/AXR2 confers late flowering under short-day light in Arabidopsis. J Integr Plant Biol **53,** 480-492.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature **416,** 499-506.

Manley, J.L., and Tacke, R. (1996). SR proteins and splicing control. Genes Dev. **10,** 1569-1579.

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.

Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol **5,** 89-99.

Mariman, E.C., van Beek-Reinders, R.J., and van Venrooij, W.J. (1983). Alternative splicing pathways exist in the formation of adenoviral late messenger RNAs. Journal of molecular biology **163,** 239-256.

Maris, C., Dominguez, C., and Allain, F.H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J **272,** 2118-2131.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research **22,** 1184-1195.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat Rev Genet **12,** 671-682.

Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS biology **4,** e21.

Martinez-Zapater, J.M., and Somerville, C.R. (1990). Effect of Light Quality and Vernalization on Late-Flowering Mutants of Arabidopsis thaliana. Plant Physiol **92,** 770-776.

Matlin, A.J., and Moore, M.J. (2007). Spliceosome assembly and composition. Advances in experimental medicine and biology **623,** 14-35.

May, G.E., Olson, S., McManus, C.J., and Graveley, B.R. (2011). Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. RNA **17,** 222-229.

McCullough, A.J., and Berget, S.M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol **17,** 4562-4571.

McCullough, A.J., and Schuler, M.A. (1997). Intronic and exonic sequences modulate 5' splice site selection in plant nuclei. Nucleic Acids Res **25,** 1071-1077.

McCullough, A.J., Lou, H., and Schuler, M.A. (1991). In vivo analysis of plant pre-mRNA splicing using an autonomously replicating vector. Nucleic Acids Res. **19,** 3001-3009.

McGlincy, N.J., Valomon, A., Chesham, J.E., Maywood, E.S., Hastings, M.H., and Ule, J. (2012). Regulation of alternative splicing by the circadian clock and food related cues. Genome biology **13,** R54.

McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. Current opinion in genetics & development **21,** 373-379.

Michaels, S.D., and Amasino, R.M. (1999). FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell **11,** 949-956.

Michaels, S.D., and Amasino, R.M. (2001). Loss of FLOWERING LOCUS C activity eliminates the late-flowering phenotype of FRIGIDA and autonomous pathway mutations but not responsiveness to vernalization. Plant Cell **13,** 935-941.

Michaels, S.D., Himelblau, E., Kim, S.Y., Schomburg, F.M., and Amasino, R.M. (2005). Integration of flowering signals in winter-annual Arabidopsis. Plant Physiol **137,** 149-156.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. (1999). Frequent alternative splicing of human genes. Genome Res **9,** 1288-1293.

Misteli, T., Caceres, J.F., and Spector, D.L. (1997). The dynamics of a pre-mRNA splicing factor in living cells. Nature **387,** 523-527.

Mitrovich, Q.M., and Anderson, P. (2000). Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes & development **14,** 2173-2184.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet **30,** 13-19.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic acids research **29,** 2850-2859.

Mola, G., Vela, E., Fernandez-Figueras, M.T., Isamat, M., and Munoz-Marmol, A.M. (2007). Exonization of Alu-generated splice variants in the survivin gene of human and non-human primates. Journal of molecular biology **366,** 1055-1063.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell **136,** 688-700.

Mukhtar, M.S., Carvunis, A.R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M.T., Pevzner, S.J., Donovan, S.E., Ghamsari, L., Santhanam, B., Romero, V., Poulin, M.M., Gebreab, F., Gutierrez, B.J., Tam, S., Monachello, D., Boxem, M., Harbort, C.J., McDonald, N., Gai, L., Chen, H., He, Y., Vandenhaute, J., Roth, F.P., Hill, D.E., Ecker, J.R., Vidal, M., Beynon, J., Braun, P., and Dangl, J.L. (2011). Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science **333,** 596-601.

Murgia, I., Tarantino, D., Soave, C., and Morandini, P. (2011). Arabidopsis CYP82C4 expression is dependent on Fe availability and circadian rhythm, and correlates with genes involved in the early Fe deficiency response. J Plant Physiol **168,** 894-902.

Muro, A.F., Caputi, M., Pariyarath, R., Pagani, F., Buratti, E., and Baralle, F.E. (1999). Regulation of fibronectin EDA exon alternative splicing: possible role of RNA secondary structure for enhancer display. Molecular and cellular biology **19,** 2657-2671.

Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R., and Fluhr, R. (2004). Intron retention is a major phenomenon in alternative splicing in Arabidopsis. Plant J **39,** 877-885.

Niedojadlo, J., Mikulski, Z., Delenko, K., Szmidt-Jaworska, A., Smolinski, D.J., and Epstein, A.L. (2012). The perichromatin region of the plant cell nucleus is the area with the strongest co-localisation of snRNA and SR proteins. Planta **236,** 715-726.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., and Barron, A.E. (2011). Landscape of next-generation sequencing technologies. Analytical chemistry **83,** 4327-4341.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., and Allain, F.H. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science **309,** 2054-2057.

Olson, S., Blanchette, M., Park, J., Savva, Y., Yeo, G.W., Yeakley, J.M., Rio, D.C., and Graveley, B.R. (2007). A regulator of Dscam mutually exclusive splicing fidelity. Nature structural & molecular biology **14,** 1134-1140.

Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B., Lander, E.S., Armstrong, S.A., and Daley, G.Q. (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature **483,** 598-602.

Ong, C.T., and Corces, V.G. (2012). Enhancers: emerging roles in cell fate specification. EMBO Rep **13,** 423-430.

Palusa, S.G., and Reddy, A.S. (2010). Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. New Phytol **185,** 83-89.

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

Palusa, S.G., Golovkin, M., Shin, S.-B., Richardson, D., and Reddy, A.S., N. (2007b). Organ-specific, developmental, hormonal and stress regulation of expression of putative pectate lyase genes in Arabidopsis New Phytol.**,** In press.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. **40,** 1413-1415.

Papp, I., Mur, L.A., Dalmadi, A., Dulai, S., and Koncz, C. (2004). A mutation in the Cap Binding Protein 20 gene confers drought tolerance to Arabidopsis. Plant molecular biology **55,** 679-686.

Peng, J., Carol, P., Richards, D.E., King, K.E., Cowling, R.J., Murphy, G.P., and Harberd, N.P. (1997). The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. Genes and Development **11,** 3194-3205.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20,** 3551-3567.

Pertea, M., Mount, S.M., and Salzberg, S.L. (2007). A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana. BMC Bioinformatics **8,** 159.

Power, K.A., McRedmond, J.P., de Stefani, A., Gallagher, W.M., and Gaora, P.O. (2009). High-throughput proteomics detection of novel splice isoforms in human platelets. PloS one **4,** e5001.

Qian, W., Iqbal, K., Grundke-Iqbal, I., Gong, C.X., and Liu, F. (2011). Splicing factor SC35 promotes tau expression through stabilization of its mRNA. FEBS letters **585,** 875-880.

Quesada, V., Dean, C., and Simpson, G.G. (2005). Regulated RNA processing in the control of Arabidopsis flowering. Int J Dev Biol **49,** 773-780.

Quesada, V., Macknight, R., Dean, C., and Simpson, G.G. (2003). Autoregulation of FCA pre-mRNA processing controls Arabidopsis flowering time. EMBO J. **22,** 3142-3152.

Rao, N., Nguyen, S., Ngo, K., and Fung-Leung, W.P. (2005). A novel splice variant of interleukin-1 receptor (IL-1R)-associated kinase 1 plays a negative regulatory role in Toll/IL-1R-induced inflammatory signaling. Molecular and cellular biology **25,** 6521-6532.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res. **12,** 1231-1245.

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

Reddy, A.S., Day, I.S., Gohring, J., and Barta, A. (2012a). Localization and dynamics of nuclear speckles in plants. Plant Physiol **158,** 67-77.

Reddy, A.S., 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. Frontiers in plant science **3,** 18.

Reddy, A.S., N. (2001a). Nuclear Pre-mRNA Splicing in Plants. Crit. Rev, Plant Sci. **20,** 523-571.

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., and Ali, G.S. (2011). Plant SR proteins: Roles in pre-mRNA splicing, plant development and stress responses. . WIREs RNA **2,** 875-889.

Reddy, A.S.N., Ali, G.S., and Golovkin, M. (2004). Arabidopsis U1 snRNP 70K protein and its interacting proteins: Nuclear localization and in vivo dynamics of a novel plant-specific serine/arginine-rich protein. The Nuclear Envelope.

Reed, R. (1996). Initial splice-site recognition and pairing during pre-mRNA splicing. Current opinion in genetics & development **6,** 215-220.

Reed, R. (2003). Coupling transcription, splicing and mRNA export. Curr Opin Cell Biol **15,** 326-331.

Richardson, D.N., Rogers, M.F., Labadorf, A., Ben-Hur, A., Guo, H., Paterson, A.H., and Reddy, A.S. (2011). Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing. PLoS One **6,** e24542.

Risso, G., Pelisch, F., Quaglino, A., Pozzi, B., and Srebrow, A. (2012). Regulating the regulators: Serine/arginine-rich proteins under scrutiny. IUBMB Life **64,** 809-816.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics **27,** 2325-2329.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26,** 139-140.

Rogers, M.F., Ben-Hur, A., and Reddy, A.S.N. (2010). SpliceGrapher: Predicting Splice Graphs from short read seqeunces from nextgen sequencing. In 8 th Annual Rocky Mountain Bioinformatics Conference, Snowmass/Aspen, Colorado

Rogers, M.F., Thomas, J., Reddy, A.S., and Ben-Hur, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biol **13,** R4.

Rose, D., Hiller, M., Schutt, K., Hackermuller, J., Backofen, R., and Stadler, P.F. (2011). Computational discovery of human coding and non-coding transcripts with conserved splice sites. Bioinformatics **27,** 1894-1900.

Ru, Y., Wang, B.B., and Brendel, V. (2008). Spliceosomal proteins in plants. Curr. Top. Microbiol. Immunol. **326,** 1-15.

Rubio, V., Shen, Y., Saijo, Y., Liu, Y., Gusmaroli, G., Dinesh-Kumar, S.P., and Deng, X.W. (2005). An alternative tandem affinity purification strategy applied to Arabidopsis protein complex isolation. Plant J. **41,** 767-778.

Sanford, J.R., Longman, D., and Caceres, J.F. (2003). Multiple roles of the SR protein family in splicing regulation. In Regulation of alternative splicing, P. Jeanteur, ed (New York: Springer), pp. 33-58.

Sanford, J.R., Gray, N.K., Beckmann, K., and Caceres, J.F. (2004). A novel role for shuttling SR proteins in mRNA translation. Genes Dev. **18,** 755-768.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Research **19,** 381-394.

Sapra, A.K., Anko, M.L., Grishina, I., Lorenz, M., Pabis, M., Poser, I., Rollins, J., Weiland, E.M., and Neugebauer, K.M. (2009). SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. Mol Cell **34,** 179-190.

Sasaki, T., Song, J., Koga-Ban, Y., Matsui, E., Fang, F., Higo, H., Nagasaki, H., Hori, M., Miya, M., Murayama-Kayano, E., and et al. (1994). Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J **6,** 615-624.

Schaal, T.D., and Maniatis, T. (1999a). Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol Cell Biol **19,** 261-273.

Schaal, T.D., and Maniatis, T. (1999b). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol **19,** 1705-1719.

Schindler, S., Szafranski, K., Hiller, M., Ali, G.S., Palusa, S.G., Backofen, R., Platzer, M., and Reddy, A.S. (2008). Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes. BMC Genomics **9,** 159.

Schoning, J.C., and Staiger, D. (2009). RNA-protein interaction mediating post-transcriptional regulation in the circadian system. Methods in Molecular Biology **479,** 337-351.

Schoning, J.C., Streitner, C., Meyer, I.M., Gao, Y., and Staiger, D. (2008). Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. Nucleic Acids Res **36,** 6977-6987.

Schoning, J.C., Streitner, C., Page, D.R., Hennig, S., Uchida, K., Wolf, E., Furuya, M., and Staiger, D. (2007). Auto-regulation of the circadian slave oscillator component AtGRP7 and regulation of its targets is impaired by a single RNA recognition motif point mutation. Plant J **52,** 1119-1130.

Schor, I.E., Allo, M., and Kornblihtt, A.R. (2010). Intragenic chromatin modifications: A new layer in alternative splicing regulation. Epigenetics **5,** 174-179.

Schuler, M.A. (2008). Splice site requirements and switches in plants. Current Topics in Microbiology and Immunology **326,** 39-59.

Searle, I., He, Y., Turck, F., Vincent, C., Fornara, F., Krober, S., Amasino, R.A., and Coupland, G. (2006). The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in Arabidopsis. Genes & development **20,** 898-912.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics **19 Suppl 1,** i273-282.

Seo, P.J., Hong, S.Y., Kim, S.G., and Park, C.M. (2011a). Competitive inhibition of transcription factors by small interfering peptides. Trends in plant science **16,** 541-549.

Seo, P.J., Kim, M.J., Ryu, J.Y., Jeong, E.Y., and Park, C.M. (2011b). Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nature communications **2,** 303.

Seo, P.J., Park, M.J., Lim, M.H., Kim, S.G., Lee, M., Baldwin, I.T., and Park, C.M. (2012). A Self-Regulatory Circuit of CIRCADIAN CLOCK-ASSOCIATED1 Underlies the Circadian Clock Regulation of Temperature Responses in Arabidopsis. Plant Cell **24,** 2427-2442.

Sharp, P.A. (1994). Split genes and RNA splicing. Cell **77,** 805-815.

Shen, H., and Green, M.R. (2004). A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Mol Cell **16,** 363-373.

Shikata, H., Nakashima, M., Matsuoka, K., and Matsushita, T. (2012). Deletion of the RS domain of RRC1 impairs phytochrome B signaling in Arabidopsis. Plant Signal Behav **7**.

Shukla, G.C., and Padgett, R.A. (1999). Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA **5,** 525-538.

Simon, R., and Starlinger, P. (1987). Transposable element Ds2 of Zea mays influences polyadenylation and splice site selection. Molecular & general genetics : MGG **209,** 198-199.

Simpson, C.G., and Brown, J.W. (2008). U12-dependent intron splicing in plants. Current Topics in Microbiology and Immunology **326,** 61-82.

Simpson, C.G., Jennings, S.N., Clark, G.P., Thow, G., and Brown, J.W. (2004). Dual functionality of a plant U-rich intronic sequence element. Plant J **37,** 82-91.

Simpson, G.G., and Dean, C. (2002). Arabidopsis, the Rosetta stone of flowering time? Science **296,** 285-289.

Simpson, G.G., Dijkwel, P.P., Quesada, V., Henderson, I., and Dean, C. (2003). FY is an RNA 3' end-processing factor that interacts with FCA to control the Arabidopsis floral transition. Cell **113,** 777-787.

Slotte, T., Huang, H.R., Holm, K., Ceplitis, A., Onge, K.S., Chen, J., Lagercrantz, U., and Lascoux, M. (2009). Splicing variation at a FLOWERING LOCUS C homeolog is associated with flowering time variation in the tetraploid Capsella bursa-pastoris. Genetics **183,** 337-345.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb. Perspect. Biol. **3,** doi: 10.1101/cshperspect.a000646.

Srikanth, A., and Schmid, M. (2011). Regulation of flowering time: all roads lead to Rome. Cellular and Molecular Life Sciences **68,** 2013-2037.

Staiger, D., and Green, R. (2011). RNA-based regulation in the plant circadian clock. Trends Plant Sci **16,** 517-523.

Staiger, D., and Koster, T. (2011). Spotlight on post-transcriptional control in the circadian system. Cellular and Molecular Life Sciences **68,** 71-83.

Staiger, D., Zecca, L., Wieczorek Kirk, D.A., Apel, K., and Eckstein, L. (2003). The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA. Plant J **33,** 361-371.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. Gene **344,** 1-20.

Stark, H., Dube, P., Luhrmann, R., and Kastner, B. (2001). Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. Nature **409,** 539-542.

Staudt, A.C., and Wenkel, S. (2011). Regulation of protein function by 'microProteins'. EMBO reports **12,** 35-42.

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). Architectural limits on split genes. Proc Natl Acad Sci U S A **93,** 15081-15085.

Stoltzfus, C.M., and Madsen, J.M. (2006). Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing. Current HIV research **4,** 43-55.

Streitner, C., Danisman, S., Wehrle, F., Schoning, J.C., Alfano, J.R., and Staiger, D. (2008). The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana. Plant Journal **56,** 239-250.

Suganuma, T., and Workman, J.L. (2011). Signals and combinatorial functions of histone modifications. Annual review of biochemistry **80,** 473-499.

Sureau, A., Gattoni, R., Dooghe, Y., Stevenin, J., and Soret, J. (2001). SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. Embo J **20,** 1785-1796.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W. (2012). Alternative splicing in plants - coming of age. Trends Plant Sci.

Tacke, R., and Manley, J.L. (1995). The human splicing factor ASF/SF2 and SC35 possess different, functionally significant RNA binding specificities. EMBO J. **14,** 3540-3551.

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

Terzi, L.C., and Simpson, G.G. (2008). Regulation of flowering time by RNA processing. Current topics in microbiology and immunology **326,** 201-218.

Thomas, J., and Reddy, A.S.N. (2012). Opposing roles of members of serine/arginine (SR)-rich protein family in regulating flowering time. In preparation.

Thomas, J., Palusaa, S.G., Prasada, K.V.S.K., Ali, G.S., Surabhi, G.-K., Ben-Hur, A., Abdel-Ghany, S.E., and Reddy, A.S.N. (2012). Identification of an intronic splicing regulatory element involved in autoregulation of alternative splicing of the *SCL33* pre-mRNA. Plant J.**,** in press.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., and Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol **16,** 996-1001.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28,** 511-515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols **7,** 562-578.

Tripathi, V., Song, D.Y., Zong, X., Shevtsov, S.P., Hearn, S., Fu, X.D., Dundr, M., and Prasanth, K.V. (2012). SRSF1 modulates the organization of splicing factors in nuclear speckles and regulates transcription. Molecular biology of the cell.

Turek, C., and Gold, L. (1990). Systemic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249,** 505-510.

Twyffels, L., Gueydan, C., and Kruys, V. (2011). Shuttling SR proteins: more than splicing factors. Febs J.

Ueda, M., Fujimoto, M., Arimura, S., Murata, J., Tsutsumi, N., and Kadowaki, K. (2007). Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene **402,** 51-56.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science **302,** 1212-1215.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B.R., Kane, D., Weinstein, J.N., Blume, J., and Darnell, R.B. (2005). Nova regulates brain-specific splicing to shape the synapse. Nat Genet **37,** 844-852.

Underwood, J.G., Boutz, P.L., Dougherty, J.D., Stoilov, P., and Black, D.L. (2005). Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. Molecular and cellular biology **25,** 10005-10016.

Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nature methods **7,** 995-1001.

Valadkhan, S., and Jaladat, Y. (2010). The spliceosomal proteome: At the heart of the largest cellular ribonucleoprotein machine. Proteomics.

Wachter, A. (2010). Riboswitch-mediated control of gene expression in eukaryotes. RNA Biol **7,** 67-76.

Wachter, A., Ruhl, C., and Stauffer, E. (2012). The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. Frontiers in plant science **3,** 81.

Wachter, A., Tunc-Ozdemir, M., Grove, B.C., Green, P.J., Shintani, D.K., and Breaker, R.R. (2007). Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. Plant Cell **19,** 3437-3450.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell **136,** 701-718.

Wang, B.B., and Brendel, V. (2004). The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. Genome Biol **5,** R102.

Wang, B.B., and Brendel, V. (2006a). Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci U S A **103,** 7175-7180.

Wang, B.B., and Brendel, V. (2006b). Molecular characterization and phylogeny of U2AF35 homologs in plants. Plant Physiol. **140,** 624-636.

Wang, C., Tian, Q., Hou, Z., Mucha, M., Aukerman, M., and Olsen, O.A. (2007a). The Arabidopsis thaliana AT PRP39-1 gene, encoding a tetratricopeptide repeat protein with similarity to the yeast pre-mRNA processing protein PRP39, affects flowering time. Plant cell reports **26,** 1357-1366.

Wang, C.J., Vlajkovic, S.M., Housley, G.D., Braun, N., Zimmermann, H., Robson, S.C., Sevigny, J., Soeller, C., and Thorne, P.R. (2005). C-terminal splicing of NTPDase2 provides distinctive catalytic properties, cellular distribution and enzyme regulation. The Biochemical journal **385,** 729-736.

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.

Wang, H.Y., Klatte, M., Jakoby, M., Baumlein, H., Weisshaar, B., and Bauer, P. (2007b). Iron deficiency-mediated stress regulation of four subgroup Ib BHLH genes in Arabidopsis thaliana. Planta **226,** 897-908.

Wang, J., Takagaki, Y., and Manley, J.L. (1996). Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability. Genes Dev **10,** 2588-2599.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., MacLeod, J.N., Chiang, D.Y., Prins, J.F., and Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research **38,** e178.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA **14,** 802-813.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10,** 57-63.

Watanuki, T., Funato, H., Uchida, S., Matsubara, T., Kobayashi, A., Wakabayashi, Y., Otsuki, K., Nishida, A., and Watanabe, Y. (2008). Increased expression of splicing factor SRp20 mRNA in bipolar disorder patients. Journal of affective disorders **110,** 62-69.

Werneke, J.M., Chatfield, J.M., and Ogren, W.L. (1989). Alternative mRNA splicing generates the two polypeptides in spinach and Arabidopsis. Plant Cell **1,** 815-825.

Wessler, S.R. (1991). The maize transposable Ds1 element is alternatively spliced from exon sequences. Molecular and cellular biology **11,** 6192-6196.

Wilson, R.N., Heckman, J.W., and Somerville, C.R. (1992). Gibberellin is required for flowering in *Arabidopsis thaliana* under short days

. Plant Physiol **100,** 403-408.

Wilusz, J., and Shenk, T. (1988). A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. Cell **52,** 221-228.

Wirth, B., Brichta, L., and Hahnen, E. (2006). Spinal muscular atrophy: from gene to therapy. Seminars in pediatric neurology **13,** 121-131.

Wolffe, A.P. (1997). Chromatin remodeling regulated by steroid and nuclear receptors. Cell Research **7,** 127-142.

Wolffe, A.P. (2001). Chromatin remodeling: why it is important in cancer. Oncogene **20,** 2988-2990.

Wu, C. (1997). Chromatin remodeling and the control of gene expression. The Journal of biological chemistry **272,** 28171-28174.

Wu, H., Chen, C., Du, J., Liu, H., Cui, Y., Zhang, Y., He, Y., Wang, Y., Chu, C., Feng, Z., Li, J., and Ling, H.Q. (2012). Co-overexpression FIT with AtbHLH38 or AtbHLH39 in Arabidopsis-enhanced cadmium tolerance via increased cadmium sequestration in roots and improved iron homeostasis of shoots. Plant Physiol **158,** 790-800.

Wu, H., ., Sun, S., Tu, K., Gao, Y., Xie, Krainer, A.R., and Zhu, J. (2010). A splicing-independent function of SF2/ASF in microRNA processing. Mol. Cell **38,** 67-77.

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q.Q., and Hunt, A.G. (2011). Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proceedings of the National Academy of Sciences of the United States of America **108,** 12533-12538.

Xiao, R., Tang, P., Yang, B., Huang, J., Zhou, Y., Shao, C., Li, H., Sun, H., Zhang, Y., and Fu, X.D. (2012). Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation. Molecular cell **45,** 656-668.

Xiao, X., Wang, Z., Jang, M., and Burge, C.B. (2007). Coevolutionary networks of splicing cis-regulatory elements. Proceedings of the National Academy of Sciences of the United States of America **104,** 18583-18588.

Xing, Y., and Lee, C. (2007). Relating alternative splicing to proteome complexity and genome evolution. Advances in experimental medicine and biology **623,** 36-49.

Xu, S., Zhang, Z., Jing, B., Gannon, P., Ding, J., Xu, F., Li, X., and Zhang, Y. (2011). Transportin-SR Is Required for Proper Splicing of Resistance Genes and Plant Immunity. PLoS Genet **7,** e1002159.

Yoo, S.D., Cho, Y.H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nature Protocols **2,** 1565-1572.

Yoshimura, K., Yabuta, Y., Ishikawa, T., and Shigeoka, S. (2002). Identification of a *cis* element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. J. Biol. Chem. **277,** 40623-40632.

Yuan, Y.X., Wu, J., Sun, R.F., Zhang, X.W., Xu, D.H., Bonnema, G., and Wang, X.W. (2009). A naturally occurring splicing site mutation in the Brassica rapa FLC1 gene is associated with variation in flowering time. J Exp Bot **60,** 1299-1308.

Zagotta, M.T., Shannon, S., Jacobs, C., and Meeks-Wagner, D.R. (1992). Early-flowering mutants of *Arabidopsis thaliana*. Aust. J Plant Physiol **19,** 411-418.

Zahler, A.M., Lane, W.S., Stolk, J.A., and Roth, M.B. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. Genes Dev. **6,** 837-847.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome research **20,** 646-654.

Zhang, W.-J., and Wu, J.Y. (1998). Sip1, a novel RS domain-containing protein essential for pre-mRNA splicing. Mol. Cell. Biol. **18,** 676-684.

Zhang, X.H., Kangsamaksin, T., Chao, M.S., Banerjee, J.K., and Chasin, L.A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. Molecular and Cellular Biology **25,** 7323-7332.

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.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Molecular cell **8,** 1351-1361.

Zhu, J., Dong, C.H., and Zhu, J.K. (2007). Interplay between cold-responsive gene regulation, metabolism and RNA processing during plant cold acclimation. Curr Opin Plant Biol **10,** 290-295.

Appendix I

IDENTIFICATION OF SCL33 INTERACTING PROTEINS FROM ARABIDOPSIS

SUMMARY

The SR protein family members interact with RNA and other proteins to regulate pre-mRNA splicing. To study the interaction of AtSCL33 with other proteins in Arabidopsis, an SCL33 TAP-tagged construct was introduced into the *scl33* knockout mutant background. More than 100 transgenic lines were analyzed by immunoblotting using a myc-tag antibody. Only one plant showed expression of the TAP-tagged protein. When the seeds from that line were germinated seedlings growth was arrested. To rescue this line, callus was generated from the SCL33 TAP-tagged line and used for isolation of SCL33 interacting proteins by tandem affinity purification and analyzed by LC-MS-MS to identify proteins. Peptide sequences specific to the SCL33-TAP line revealed the identify of SCL33 interacting proteins. Interestingly, a set of RNA binding proteins including UBA2c, AtGRP7, AtGRP8 and SWAP were found to associate with SCL33. This provides the evidence of SCL33 interaction with other RNA binding proteins as part of an interaction complex involved in regulation of splicing. This analysis is a first step towards the creation of a comprehensive protein-protein network map of SCL33 to gain insights into the function of SCL33.

INTRODUCTION

RNA binding proteins (RBPs) have been shown to be involved in regulation of post-transcriptional processing events including pre-mRNA splicing, polyadenylation, RNA stability and RNA export (Lorkovic, 2009). RBPs contain one or more RNA-binding domains (Glisovic et al., 2008) and other auxiliary domains such as glycine rich, arginine-rich, arginine–glycine (RGG) or SR (serine-arginine) domains (Ambrosone et al., 2012). Biochemical and structural studies have demonstrated that RRM motifs are involved in RNA recognition and in protein–protein interactions, leading to the formation of heterogeneous complexes (Maris et al., 2005).

The Arabidopsis genome codes for more than 200 RBPs that are found associated with cellular RNA in form of RNP complexes. Pre-mRNA splicing takes place in a large protein complex consisting of small ribonucleoprotein particles (snRNPs) and many non-snRNPs proteins (Sharp, 1994). The major class of non-snRNP proteins that regulate both constitutive and alternative splicing are serine/arginine-rich (SR) proteins. The SR proteins, which were first discovered in 1990, are conserved between plants and animals (Kalyna and Barta, 2004; Reddy et al., 2004). During the last 20 years, extensive studies have been carried out on metazoan SR proteins (Long and Caceres, 2009), but there has been a major gap in understanding the roles of SRs in plants. All SR proteins are phosphoproteins with typical characteristic features that include (1-2) N-terminal RRMs followed by a downstream arginine/serine-rich (RS) region of at least 50 amino acids with RS or SR dipeptides. In-vivo and in-vitro protein-protein interaction experiments have revealed a complex network of direct interactions among SR proteins and with other spliceosomal proteins (Reddy, 2007). The plant specific SRs (SCL33, RSZ21, RSZ22) and SR45 were isolated using the arginine-rich domain of plant specific U1-70K in yeast two hybrid screens (Golovkin and Reddy, 1996, 1998).

There are 18 SR proteins in Arabidopsis, 22 in rice, 12 in humans and 7 in *C. elegans* (Barta et al., 2010; Manley and Krainer, 2010). The RRM domain binds to specific regulatory sequences in pre-mRNA, and the RS domain facilitates protein-protein and protein-RNA in the splicing machinery. Plant SR proteins have been known for about 15 years but their biochemical analysis has been hampered due to the lack of plant-derived *in-vitro* splicing extracts.

In higher eukaryotes the sequences around the splice sites are less conserved compared to yeast. Therefore, for accurate and efficient recognition of splice sites additional regulatory sequences called splicing enhancers/repressors are required. These *cis*-elements are recognized by splicing regulators such as SR and hnRNP proteins (Manley and Tacke, 1996; Long and Caceres, 2009). The recruitment of U1snRNP to the 5' splice site and other snRNPs to assemble in the spliceosome is facilitated by members of the SR family splicing (Long and Caceres, 2009). SR proteins are also involved in bridging 5′ and 3′ splice sites by interacting with U1-70K and U2AF^{35} and enable incorporation of the tri-snRNP complex (U4/U6.U5 tri snRNP) into the spliceosome (Kramer, 1996).

SR proteins also function in selecting alternative weak splice sites. The protein-protein interactions between U1-70K or U11-35K with SRs have helped to unravel some crucial and specific roles of SRs, including early stages of spliceosomal complex formation and maintenance of SR functions/activity by interacting with protein kinases (e.g., AFC2, CypRS64/92, CypRS and PK12) (Reddy et al., 2004). Also, the interaction of the largest subunit of RNA polymerase II (Cyp59) with SRs connects RNA synthesis to the splicing process (Gullerova et al., 2006). Recent studies in animals show that SR proteins perform additional roles, which include export of mRNA to cytoplasm, mRNA stability, translation, genome maintenance, microRNA biogenesis and oncogene transformation (Huang and Steitz, 2005; Long and Caceres, 2009; Wu et al., 2010).

To identify SCL interacting proteins, transgenic lines were generated expressing tagged SCL33 in the *sr33* knockout background. The tandem affinity purification (TAP) strategy employed here is an efficient approach for both protein complex purification and mRNA binding studies. The TAP system has allowed efficient isolation of a multiprotein complexes in plants (Rubio et al., 2005). SR proteins are known to interact with a number of other SRs and spliceosomal proteins. However, untill today most of the splicing factor interactions in plants have been studied by yeast two hybrid or other *in-vitro* assays. The earlier systems might not have identified some of the interactions of SRs because it is known that SR proteins interact primarily with RNA in living cells and that they are recruited to chromatin primarily via interactions with nascent mRNA (Sapra et al., 2009). Therefore, a strategy to unravel protein-protein interactions *in vivo* should provide more meaningful information. The experiments and data described here is the first *in-vivo* purification approach to identify proteins which associate with a splicing factor in plants. The SCL33-TAP tagged lines are used here to isolate the SCL33 protein complexes as a first step to develop methods towards the identification of all the interacting proteins, which will shed light on protein interaction networks, as well as provide novel insights into other splicing regulators.

MATERIAL AND METHODS

Generation of SR33CTAP and SR33NTAP constructs

The SCL33 gene was used to make TAP-tag fusions in the C-terminal TAPa T-DNA vector (pC-TAPa) and N-terminal TAPa T-DNA vector (pN-TAPa) for affinity purification of the protein from plants (Figure A.1). The vectors have 6x His-tags, 9 myc-tags, 3C protease cleavage site and 2 copies of the IgG binding domain driven by the 35S promoter double enhancer and a TMVU1 leader. The development of these vectors is described previously (Deng et al., 1992). The SCL33 gene was PCR amplified with attB sequence primers and used to clone in the Gateway cloning kit (Invitrogen). The full-length sequence of SCl33 was cloned into the pDONR201 plasmid with the attB1 and attR1 cloning sites for the BP reaction, using two sets of forward primers:

5’-GGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAGG

5’-AGGCTTCGAAGGAGAGATAGAAACCATGAGGGGAAGGAG,

and two sets of reverse primers:

5’-CAAGAAAGCTGGGTCCTGGCTTGGTGAACGGTCTTC

5’-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTGGCT.

The transfer of genes from the pDONR 201 plasmid to the corresponding TAPa vector was performed using the LR reaction (Gateway, Invitrogen).

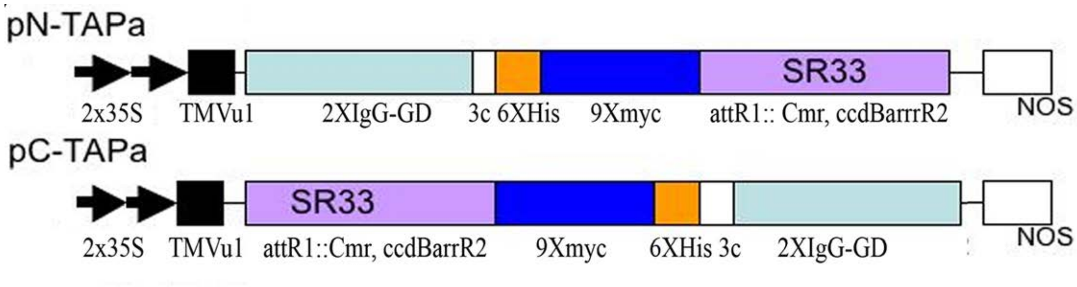


Figure A.1: Schematic representation of the pNTAP and pCTAP vectors expressing SCL33 in plants. Both pNTAPa and pCTAPa vectors allow translational fusion of tags to either N-terminus or C-terminus of the desired protein, respectively. The expression is driven by two copies of tobacco mosaic virus (2x35S) and a tobacco mosaic virus (TMV) U1Ω translational enhancer. The TAPa tag consists of two copies of the IgG binding domains (2XigG-GD), an eight amino acid sequence corresponding to the 3C protease cleavage sites (3C), six histidine stretch 6xHis, and nine repeats of the myc epitope (9xmyc). Both pNTAPa and pCTAPa vectors contain a Gateway cloning site (attR1::Cm^{r}::ccdB::attR2). The Nos terminator is located downstream of each expression cassette.

Plant Material and growth conditions

The TAPa constructs were transformed into T-DNA insertion mutant *scl33* line from the Salk collection *(SALK_058566)* using the floral dip method with GV3101 strain of Agrobacterium. The To seeds were plated on MS plates with 3% sucrose, gentamycin (75 ug/ul) and carbenicillin (200 ug/ul) to obtain T1 plants. The callus tissue from the T1 plants was generated on MS plates with 0.5 mg/L 2,4-dichlorophenoxy acetic acid (2,4-D). The media with Benzylaminopurine (BA at 1mg/L) and Naphthalene acetic acid (NAA at 0.01 mg/L) hormone combination was used to generate differentiated green callus tissue.

SR33CTAP purification protocol

The subcultured SR33CTAP callus line was grown for 15 days under 16 hrs light : 8 hrs dark at 22^0C. For isolation of protein as shown in Figure A.2, the tissue (20g fresh weight) was ground in liquid nitrogen and thawed in 2 volumes of extraction buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% Nonidet P-40, 1 mM PMSF), and 1x complete protease inhibitor cocktail (Roche). The mix was sonicated for 10-15 minutes at 4^0C and filtered through four layers of cheesecloth, and centrifuged at 12000g for 10 min at 4^0C. The supernatant was pre-cleared with uncharged His beads, and then passed through a column of 1 ml His beads coated with charged 1X NiSO4 buffer (Qiagen, Valencia, CA, USA). The Ni- coated His beads were washed three times with 10 ml of washing buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% Nonidet P-40). The elution was performed using 5 ml of imidazole containing buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% Nonidet P-40, 0.05 M imidazole). All the steps in the purification procedure were carried out at 4^0C with one protease inhibitor tablet in each 10 ml of wash or elution buffer. The 5 ml His eluate was incubated with 100ul c-myc beads for 15 minutes and then washed 2 times with wash buffer. The final elution was performed with three aliquots of 200ul of low pH elution buffer (0.1M citric acid) and neutralized with 30ul of neutralization buffer (2M Tris base). The three separate eluted samples were pooled together, dried in a speedvac, and dissolved in 30ul 8M urea and trypsin digested in a total volume of 120ul (2M urea) for LC-MS/MS.

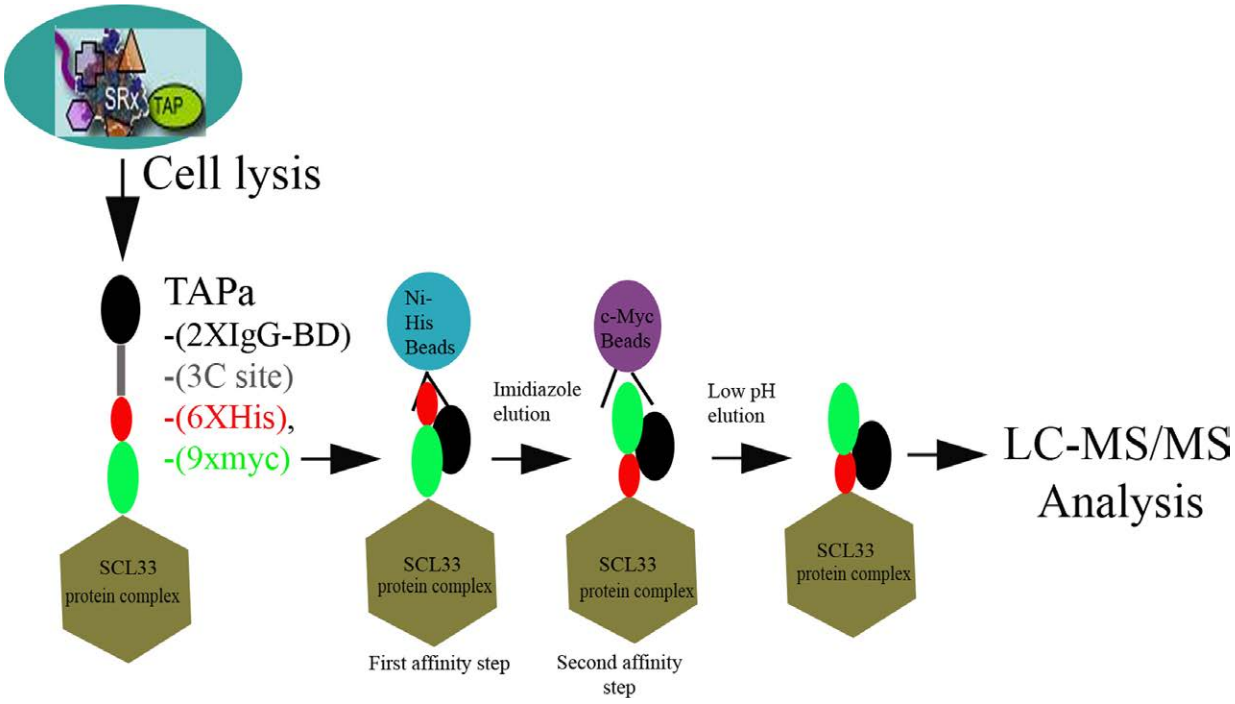


Figure A.2: Workflow of affinity purification of the SCL33 protein interaction complex proteins. Schematic representation showing SCL33-protein complexes. For the first affinity purification callus tissue expressing the SCL33-TAG protein is lysed in buffer and passed through the (Ni-His beads) column. The next affinity purification step consists of the incubation with c-myc beads. The complex of SCL33 is eluted at low pH and used for LC-MS/MS to study the interacting proteins.

RESULTS

Generation of *SCL33 TAP* tagged gene constructs in Arabidopsis

To isolate an SR protein interaction complex from plants the SCL33 gene was used to make gene fusions for tandem affinity purification (TAP). The two vectors used to make TAP tag fusions with the SCL33 gene are pCTAPa and pNTAPa (Figure A.1). The vectors have 6x His-tags, 9 myc-tags, 3C protease cleavage site and 2 copies of an IgG binding domain driven by a double 35S promoter and a TMVU1 leader for efficient translation. The main purpose of different affinity purification tags is to remove most of the nonspecific binding.

The TAPa constructs were transformed into mutant *sr33* background to avoid competition of endogenous SCL33 protein. The To seeds were plated on gentamycin and generated about 100 T1 CTAP lines. These ~100 T1 plants were tested for protein expression, using c-myc antibody. Out of 100 T1 plants screened for protein expression using c-myc antibody, only one line was identified that strongly expressed the SCL33CTAP protein (83 kda). This plant grew normally but set only a few seed. Around 10 seed were plated on MS plates, and surprisingly all of them produced small hypocotyls and stopped growing.

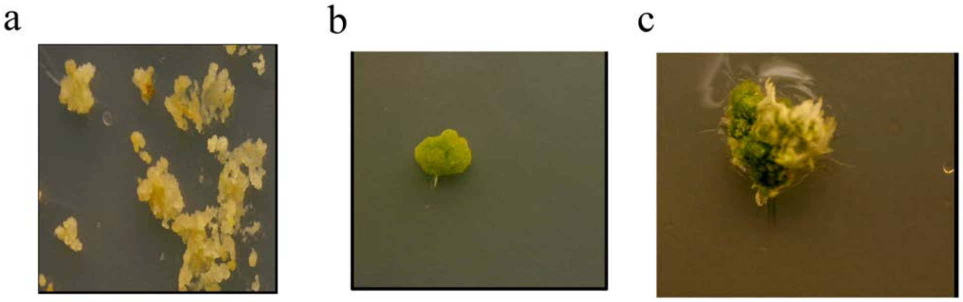


Figure A.3: Callus tissue from SCL33C-TAP expressing lines (a) white-undifferentiated callus (b) green callus (c) green callus differentiating into roots & shoots.

To rescue this SCL33-CTAP line, tissue culture was initiated on the germinated seedlings (Figure A.3). For this, varying concentrations of BA, NAA and 2-4D were used to test and rescue the genotypes, and the regenerated calli were tested for protein expression. The objective was to regenerate multiple plants out of these calli and use the plant material for protein purification purpose. However, as soon the undifferentiated callus turned green (differentiated) no expression of the protein was found (Figure A.4). A possible explanation could be that as soon as the tissue turns green, there is silencing at the post-transcriptional or translational level. This could be another reason why no expression was found in almost 100 T1 plants. As a control, callus from an *sr33* line was also made to use for further analysis.

The callus tissue is an aggregate of single cells and is a more homogenous cell system and is probably a better material for isolating the protein interaction complex than whole plants, since it excludes the highly expressed photosynthesis related proteins. However, this limits our studies to a single undifferentiated tissue. The TAP tagged lines are however necessary since antibodies specific to SRs proteins are unavailable, and it is difficult to design antigens that distinguish between paralogs, so the endogenous protein can't be used for mRNA targets.

Isolation of SCL33 protein complexes

For isolation of TAP-tagged protein complexes, 10 plates were used to grow callus tissue. Harvested tissue was frozen in liquid nitrogen and stored at -70^{0}C. For the first test extraction 15g of white callus tissue (SCL33-CTAP) and control (*sr33*) callus were used to extract proteins. Crude extracts were incubated with IgG beads, washed, and treated with cleavage buffer (Figure A.2). However, the eluate did not show any signal on western blots. The protein could have been folded in such a way that the IgG domain was masked. Therefore, the Ni-NTA resin column was used purify protein with His.bind agarose.

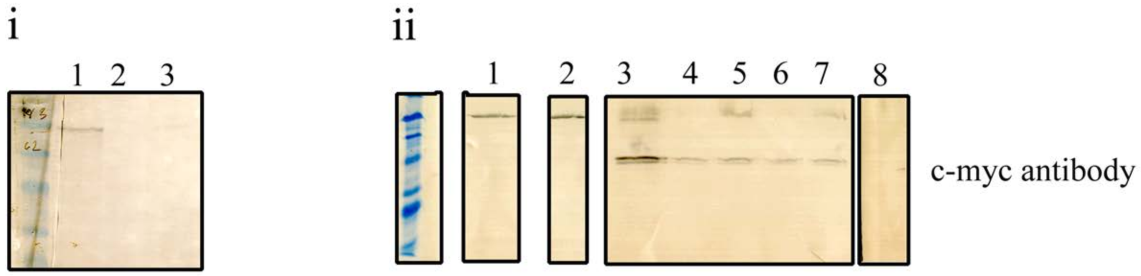


Figure A.4: Study of stability of SCL33 protein in different callus types and during two purification steps of SCL33 protein complex . Western blots probed with c-myc antibody. (i) Lane1-white callus expresses SCL33-CTAP (83kDa) protein. Lane2- green callus and Lane3- the differentiated calli do not show expression of protein. (ii) Lane1-input SCL33C-TAP protein, Lane 2-elution from Ni-column, Lanes 3-7 are 1st - 5th elutions from c-myc beads. The protein gets degraded during the low pH elution from the c-myc beads (Lane 3-7). Lane 8 is the control with just *sr33* mutant callus showing no expression of the SCL33 protein and with no background of unspecific proteins.

The eluate was tested on western-blot, Coomassie, and silver staining gels. The western blot gave a distinct signal with c-myc antibody. The Coomassie stained gel and silver stained gel gave multiple bands for both control and SCL33CTAP lines, suggesting that His beads eluate is enriched in SCL33 complexes but contain contaminants. To overcome this problem the crude extract was passed through His beads and later through c-myc beads. Following this procedure the silver stained gel showed no bands in the control *sr33* callus but the SCL33CTAP line showed a few distinct bands. As seen in the western blot there is degradation of the protein after elution from c-myc beads (Figure A.4). Various time points were tested to find where the degradation could have happened. From the experiments, it appears that the protein complex was intact on the beads. The degradation happened during the elution process. Since the degradation happened during elution, it was decided to purify the complex using myc beads and directly perform proteomic analysis. From the initial callus harvest, multiple samples were ground and treated with binding buffer, purified protein through myc-beads and sent to Yale University for proteomics analysis.

Identification of SCL33 interacting proteins

This analysis of peptides from the LC-MS-MS analysis (Yale, School of Medicine) revealed multiple peptides by the MASCOT search tool (Perkins et al., 1999) that are presented only in SCL33-CTAP samples but not in controls. The peptides and Gene-IDs that were found consistently in the SCL33-CTAP line samples and not in the *scl33* mutant line were considered further for analysis and are shown in Table A.1. The proteins that are present only in SCL33-CTAP include several that have RNA-binding motifs and were examined further.

The analysis of the SCL33 TAP-tagged complex was done from eight samples of callus compared to control tissue of the *scl33* mutant genotype. The peptides identified from different samples by LC-MS/MS analysis (Yale, Proteomics facility) were screened for their representation in multiple samples and level of significance using the calculated protein scores and e-value scores (Perkins et al., 1999); http://www.matrixscience.com/help/interpretation_help.html). After eliminating peptides such as keratin that are often present as contaminants (http://www.thegpm.org/crap/index.html), peptides present in control and those that could be carried through by IgG binding (e.g. lectins) a set of proteins were identified that were predicted to be RNA binding in function and were examined for their role in interaction with SCL33 protein. The analysis and identity of these proteins is described further in the Discussion.

Table A.1: Proteins interacting with SCL33 in vivo

RNA binding proteins	Function	# Hits	Expectation e-value	Score	Expectation e-value	Peptide Sequence
AT3G15010	RNA recognition motif (RRM)-containing protein	2	2.3E-85	95.47	1.1E-8	R.SLFSSYGDLEEAIVILDK.V
				84.85	9.3E-8	R.FTTDQLLDLLQEAIVR.H
				79.08	5.4E-7	K.TAEGAQAALADPVK.V
				77.75	1.0E-6	R.GLAADTTTEGLR.S
				74.31	2.2E-6	R.LTADSDISQR.K
				65.49	1.2E-5	R.VTVTQLAASGNQGTGSQIADISMR.K
				63.81	2.4E-5	K.IYVANVPFDMPADR.L
				53.69	1.9E-4	K.TAEGAQAALADPVK.V
AT4G31200	SWAP (Suppressor-of-White-APricot)/ surp domain-containing protein	1	5.1E-3	70.86	2.4E-6	R.SPFAPALAEALR.D
AT2G21660	CCR2/AtGrp7 Encodes a small glycine-rich	2	2.5E-5	79.21	7.2E-7	R.SITVNEAQSR.G
				44.26	1.5E-3	R.ALETAFAQYGDVIDSK.I
AT4G39260	AtGRP8 encodes a glycine-rich protein with	2	8.4E-5	69.18	6.3E-6	R.VITVNEAQSR.G
		1	1.3E-3	45.06	1.1E-3	R.TFSQFGDVIDSK.I
				76.8	9.9E-7	R.VITVNEAQSR.G

# of hits refer to the number of times the peptide is found in different replicates.

DISCUSSION

The objective of this work was to identify proteins bound to the SCL33 protein *in vivo* that function together with SCL33 in its biological roles. To get information on SCL33 interacting proteins *in vivo*, an approach was taken to directly isolate the SCL33 interacting protein complex. To facilitate this strategy the SCL33 gene was fused to a TAP-tag domain for affinity purification and the strong CaMV35S promoter controlling the TAP-tagged SCL33 construct was introduced into the *scl33* knockout mutant background to avoid competition with the native protein. The high level expression of the CaMV35S promoter restricted the recovery of transgenic plants that were photosynthetically active, since the selected CaMV35S SCL33 expressing seedlings did not grow further after initiation of greening on media. However, it was possible to generate transgenic callus tissue expressing the SCL33 TAP-tagged protein and this tissue was used for characterization of the SCL33 interactome.

The recovery of the SCL33 TAP-tag complex from callus turns out to have some advantages over the whole plant system. Leaf tissue has a predominance of photosynthesis related proteins such as RUBISCO that can mask other less prevalent proteins. Callus tissue also displays the most diversity in expressed genes and has often been used to isolate cDNAs for EST analysis (Sasaki et al., 1994). The SCL33CTAP callus tissue thus turns out to be a suitable experimental system to examine the diversity of proteins bound and interacting with the SR protein SCL33.

The most significant proteins bound to the SCL33 TAP-tagged protein, with eight peptides identified, is the Arabidopsis At3g15010 gene product, which is predicted to be an RNA recognition motif (RRM)-containing protein with functions in RNA and nucleotide binding. The corresponding gene, also referred to as UBA2c encodes heterogeneous nuclear ribonucleoprotein (hnRNP)-type RNA-binding protein. The UBA2c/At3g15010 gene shares significant sequence homology to the other two members of the same UBP family, UBA2a (At3g56860) and UBA2b and *to Vicia faba AKIP1*. The *Vicia faba* gene *AKIPl* is involved in the abscisic acid (ABA) signaling pathway and regulates stomatal closure (Li et al., 2002)**.** UBA2a has been previously characterized for its ability to interact with UBP1, an hnRNP-like protein involved in both mRNA splicing and stability, and increased splicing efficiency of suboptimal introns (Lambermon et al., 2002). Th*e* UBA2 genes are conditioned on alternative splicing affecting only the 3'-untranslated regions (UTRs), and different splice variants are differentially induced by wounding via the methyl jasmonate pathway (Bove et al., 2008). The gain of function mutants of each of the three UBA2 genes leads to a leaf yellowing/cell death-like phenotype in Arabidopsis plants and lethality (Kim et al., 2008). Like SR proteins, hnRNP proteins direct their influence on pre-mRNA splicing through site-specific binding with the target RNA. But unlike SR proteins, the mechanism through which hnRNPs affect splicing is by repressing spliceosomal assembly through blocking the recruitment of snRNPs along the exon/intron boundaries, although some reports also document a positive role for these proteins in generic splicing (Martinez-Contreras et al., 2006; Busch and Hertel, 2012). In agreement with the antagonistic nature of SR and hnRNP proteins, we propose that the SCL33 protein binds to the RGG domain of UBA2c protein to enhance the availability of weak splice sites for alternative splicing, or promotes/represses overall spliceosomal machinery for efficient splicing.

The At4g31200 protein also was found bound to the SCL33 TAP-tagged protein, identified by one peptide. This protein is predicted to contain a SWAP (Suppressor-of-White-APricot) domain, which has been shown in mammalian systems to act as a splicing factor and interacts with other SR proteins (Zhang and Wu, 1998). In Arabidopsis there are 18 SWAP domain containing genes, including At4g31200, out of which interesting results have recently been revealed for the At5g25060 gene (Shikata et al., 2012). The At5g25060 (also called RRC1) gene functions were revealed by a mutant phenotype showing developmental defects and involvement in photomorphogenesis. The corresponding *rrc1* mutant, which lacks the C-terminal RS (arginine/serine rich) domain displays reduced phyB signaling and causes aberrant alternative splicing of several SR protein genes. These results of a SWAP domain protein showing interactions with SR proteins, support our observation of protein interactions between the SCL33

SR protein and the At4g31200 SWAP domain protein, signifying a biological role of this interaction in alternative splicing.

The observed interaction of the AtGRP7 and AtGRP8 proteins with the SCL33 protein in the TAP-tagged protein interaction complex is supported by functional analysis of the AtGRP7 and AtGRP8 proteins and their role in alternative splicing and the circadian rhythm clock (Staiger et al., 2003; Streitner et al., 2008). This experimental evidence of *in vivo* interaction with the SR protein SCL33, brings these two protein families (GRP and SR proteins) together as part of the protein interaction complex involved in alternative splicing.

These analyses on SCL33 interactions with a number of RNA binding proteins validate the system used here, since many of the interactions are observed to occur by other independent means. The interaction discovered here may occur directly between the proteins, or include interaction with the pre-mRNA substrate for alternative splicing. The results are quite encouraging for the use of this TAP-Tag strategy to identify interacting protein partners involved in alternative splicing and in other biological roles of the SR protein family.