universität wien

DISSERTATION

Titel der Dissertation

„Unraveling Transcriptome Complexity in Plants: The Alternative Splicing Landscape of *Arabidopsis thaliana*“

Verfasserin

Yamile Marquez Ortiz

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 094 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreut von:	ao. Univ.-Prof. Mag. Dr. Andrea Barta

Acknowledgements

I want to dedicate this thesis to all the people who have contributed to reach this important accomplishment of my life.

First of all, there are not enough words to express my gratitude to my beloved parents, Maria Guadalupe Ortiz Valero and Jose Maria Marquez Beltrán, they are the most amazing people I could ever imagined and wished as parents, without their constant support and love I definitely would not be achieving this. “Gracias por todo el amor y la sabiduría con los que han llenado cada uno de mis días y guiado mi camino. Son los padres más maravillosos, jamás hubiera logrado lo que hasta ahora sin ustedes. Éste éxito se los dedico con todo mi amor, que sin duda alguna no se compara con todo lo que me han dado. Son los pilares en los que he fundado mi vida, el ejemplo a seguir para ser una mejor persona y el elemento esencial de cada uno de mis logros. Con todo mi amor para ustedes¨.

To my sister, Aislinn, who always encourage and motivate me with all her love. For giving me a different perspective about life and all her multiple advises that in her way gave me the strength to stand up. ¨Te adoro nena, tu forma de ser tan llena de vida y alegre, me llenan de energía y me hacen ver que la vida es para disfrutarse y rodearse de amor, esto va para ti¨.

To Kuauhyama, Nathalie, Zyanya y Zelik, thanks for being always there showing your support. “Los quiero muchísimo, gracias por todo¨.

To my grandmother, who always motivate me to fly as high as possible “Abuelita, gracias por todo el apoyo que me haz brindado para poder llegar hasta aqui, te dedico esto con todo mi cariño¨.

To my advisors Andrea Barta and Mariya Kalyna, without any doubt, I could not manage a PhD without the right advisors standing beside me. I did not only finish a PhD, I enjoyed it doing it. I thank you for all the knowledge, support and patience in the course of this journey. I have learned so many things. I looked back when I started and I considered myself so lucky to have such remarkable people to guide

me. You showed me the nicest colors of doing science. I enjoyed every discussion and the fun behind developing new ideas. You gave me all the incentive and expertise for this realization. I respect you and admire you. Thanks for everything. I dedicate this to you.

To my Scottish collaborators, John Brown and Craig Simpson, which are not only extraordinary scientists, they are wonderful people. Thanks for all the contributions and knowledge; you add indispensable elements in the course of my PhD and I really appreciate the fortune of that.

To my former advisor Federico Sánchez, you inspired me and give me all the foundations to pursue my scientific career. Here I am finishing my PhD and in each of my achievements you were always and will be a part of them. You thought me the balance of life and science and that there is always room for being successful in both aspects. Thanks for all counseling and motivation, I dedicate this to you ¨Y esto merece un goya¨.

To my two wonderful friends, Jeannette and Lucía, more than 15 years of knowing each other that despite the circumstances and the different ways we took in the course of our lives, we remained unconditional. Thanks for being there, for encouraging me and for showing me the real meaning of being friend. ¨Las adoro nenas, las llevo siempre conmigo, esto va para ustedes¨.

To the delightful people, Maria Elena and her family, Claudia, Salvador and their children, that showed me that I do not need to be part of their family in order to make me feel as part of theirs. Thanks for providing me your enthusiasm and appreciation. “Gracias por todo, dedico esto a ustedes con todo mi cariño” .

To all the wonderful people I met inside and outside the lab, giving me the support and pleasant moments. It is always critical to have the right atmosphere in order to enjoy and be successful.

Special thanks to my amazing multi-national friends Zahra, Mansoureh, Petry, Manali, Sofiya and Gustavo, who were there in the toughest situations and cheering me up to keep going. Each of you, have thought me important life lessons giving me a new perspective that positively change my life. It is a privilege to meet such outstanding people like you. Thanks for every moment we shared, the good and the

bad ones. I dedicate this with all my appreciation to you, expecting that this is just the beginning of sharing my accomplishments in which you will be taking part of.

And recalling Nelson Mandela's quote: "What counts in life is not the mere fact that we have lived. It is what difference we have made to the lives of others that will determine the significance of the life we lead" …and all of you have made a difference in mine.

Thank you

“For believers, God is in the beginning, and for scientists He is at the end of all considerations…”

Max Planck

Abstrakt

(German version)

Alternatives Spleißen (AS) ist ein wichtiger posttranskriptioneller Prozess, der zur Regulation der Genexpression beiträgt und zur Transkript- und Proteomdiversifizierung in einem Organismus führt. Meine Dissertation analysiert AS in der Modelpflanze *Arabidopsis thaliana,* ein bisher noch wenig erforschtes Arbeitsgebiet. AS kann die Genexpression durch die Erzeugung von Transkripten, die zum Abbau durch den „Nonsense-mediated decay (NMD)"-Weg bestimmt sind, stark beeinflussen. Durch unsere Untersuchungen von NMD-beinträchtigen Pflanzen mittels eines hochauflösenden „RT-PCR AS Panel" konnten wir zeigen, dass 13-18% der multi-exonischen Arabidopsis Gene durch AS gekoppelt mit NMD reguliert sind. Genauere Analysen der NMD-sensitiven AS Transkripte zeigten, dass zusätzlich zu den bereits bekannten Merkmalen eines Transkripts, die NMD auslösen können, auch AS in der 5'-untranslatierten Region der mRNA NMD auslösen kann, und zwar abhängig von der Präsenz, Länge, und Position von uORFs („upstream open reading frames"). Bemerkenswert war jedoch, dass Transkripte, die ein ungespleißtes Intron (Intronretention, IR) und typische NMD-Merkmalen aufwiesen, dennoch nicht durch NMD abgebaut wurden, was eine Erklärung für das relativ hohe Vorkommen von IR in Pflanzen sein könnte. Diese Untersuchungen zeigten auch, dass weitaus mehr unterschiedliche Transkripte von den getesteten Genen produziert wurden, als bisher gezeigt werden konnte. Um festzustellen, wie oft und in welcher Form AS genomweit vorkommt, sequenzierten wir eine normalisierte cDNA Bibliothek von Wildtyp-*Arabidopsis* Pflanzen mittels Hochdurchsatz-Techniken (RNAseq). Damit konnten wir zeigen, dass ~61% der multi-exonischen Gene alternativ gespleißt werden, was die Bedeutung von AS für die Transkriptdiversifizierung veranschaulicht. Obwohl IR der am häufigsten vorkommende AS Typ ist, kommen IR Transkripte oft nur in geringen Mengen vor und tragen somit wenig zur mRNA Gesamtpopulation bei. Eine gründliche Analyse der IR Transkripte führte uns zur Entdeckung eines neuartigen Typs von AS, den wir als „Exitron-Spleißen" bezeichnen, wobei interne Regionen eines kodierenden Exons heraus gespleißt werden können. Exitron-Spleißen kommt auch beim Menschen vor, was zeigt, dass es ein genereller AS Typ ist, der das Repertoire der bereits bekannten Typen von AS ergänzt. Exitronsequenzen kodieren oft für unstrukturierte Proteinregionen, was vermuten lässt, dass ihr Herausspleißen die Interaktionsnetzwerke von Proteinen beeinflusst. Durch evolutionäre Analysen

konnten wir auch zeigen, dass Exitrone ursprünglich von „Ur-Exon-Sequenzen“ abstammen. Wir schlagen auch eine Hypothese für den Mechanismus der Exitronbildung vor, bei der durch den Verlust der benachbarten Introns und mit Hilfe zurückgebliebene Spleiß-Signale die Evolution von Exitron-Spleißen vorangetrieben wurde (Hypothese des Spleißgedächnisses).

Abstract

Alternative splicing (AS) is a major post-transcriptional process involved in regulating gene expression and generating transcript and proteome diversity. Here, I present my contributions to the relatively unexplored AS landscape in the model plant *Arabidopsis thaliana*. AS can regulate gene expression by producing transcripts that are targeted for degradation via nonsense-mediated decay (NMD). Analyzing plants impaired in NMD by the high-resolution RT-PCR AS panel, we found that 13-18% intron-containing genes are regulated by coupled AS/NMD. Inspection of the NMD-sensitive AS transcripts showed that, in addition to known features, AS in 5'UTRs affects the presence, size, and position of uORFs which can trigger NMD. Notably, intron retention (IR) transcripts presenting classical NMD features were NMD insensitive, explaining the high frequency of IR events in plants. This study has also revealed an unexpectedly high prevalence of novel AS transcripts. To assess the frequency of AS at a genome-wide level, we employed RNA-Seq of a normalized cDNA library from wild-type *Arabidopsis* plants. We found that ~61% of the multi-exonic genes undergo AS, setting AS as a major mechanism for transcript diversity in *Arabidopsis*. Despite IR being the most frequent type of AS, most of the IR transcripts are low abundant, indicating their minor contribution to the total mRNA population. Closer inspection of the IRs led to the discovery of a novel type of AS event which we named exitron (**ex**onic **i**ntr**on**) splicing, whereby internal parts of a coding exon can be alternatively spliced. Exitron splicing is also present in humans, therefore it is a general AS event complementing the repertoire of existing AS events. Exitron-encoded protein sequences are enriched in disordered regions, implying an impact of exitron splicing on remodeling of protein interaction networks. We show that exitrons originate from ancestral exons through a history of adjacent intron loss and propose the "splicing memory" hypothesis stating that remaining splicing regulatory signals promoted the evolution of exitron splicing.

Contents

List of abbreviations

Abbreviation	Definition
AS	Alternative splicing
bp	Base pairs
BPS	Branch point sequence
CBC	Cap binding complex
cDNA	Complementary deoxyribonucleic acid
CHX	Cycloheximide
CPA	Cleavage and polyadenylation
CTD	Carboxy-terminal domain
DNA	Deoxyribonucleic acid
ds cDNA	Double stranded complementary deoxyribonucleic acid
EJC	Exon junction complex
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
EST	Expressed sequence tags
GRP	Glycine-rich protein
hnRNPs	Heterogeneous nuclear ribonucleoproteins
HTS	High-throughput sequencing
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
mRNA	Messenger ribonucleic acid
NMD	Nonsense-mediated decay
nt	Nucleotides
PABP	Poly(A)-binding protein
PPT	Polypyrimidine tract
pre-mRNA	Precursor messenger ribonucleic acid
PTB	Polypirimidine tract binding protein
PTC	Premature termination codon
PTM	Post-translation modification
RNA	Ribonucleic acid
RNA pol II	Ribonucleic acid polymerase II
RNA-Seq	Ribonucleic acid high-throughput sequencing
RRM	Ribonucleic acid recognition motif
RT-PCR	Reverse transcription polymerase chain reaction
SLiM	Short linear motifs
snRNA	Small nuclear ribonucleic acid
snRNP	Small nuclear ribonucleic particles
SR protein	Serine-rich protein
TREX	Transcription-export complex
uORF	Upstream open reading frame

UTR	Untranslated region
WT	Wild type

Introduction

Introduction

Organism complexity, a new perspective in modern biology: lessons from the RNA

How is the genetic information processed? Where and how is the fate of a specific cell during the life span of a living organism determined? What determines the complexity of an organism? Such questions have intrigued biologists for centuries. In the last decades, it has become clearer that in order to understand the biological properties of the cell at molecular level and what contributes to organism phenotype, the simplistic view of the molecular biology paradigm “one gene, one protein” stating the relationship between genes and proteins is far from explaining organism complexity.

The genetic information of an organism is stored in the DNA (deoxyribonucleic acid) and then this is interpreted to a messenger ribonucleic acid (mRNA) molecule that finally can be translated into a protein. The process of copying a segment of DNA into a precursor mRNA (pre-mRNA) molecule is known as transcription and is being executed by the RNA polymerase II (RNA pol II). From transcription to the final mRNA molecule, multiple processes occur that can change the final information copied in this molecule. The sequencing of the complete genomes of human and other model genetic organisms hinted about the complexity in the interpretation of the information written in the DNA. The analyses of the different genomes led to the realization that contrary to what was expected, the human genome contains a similar number of protein coding genes (~25,000 genes, International Human Genome Sequencing (2004)) as other far simpler multicellular organisms such as the plant *Arabidopsis thaliana* (~25,000 genes, (Arabidopsis Genome, 2000)), the worm *Caenorhabditis elegans* (18,000 genes, (Consortium, 1998)), or the fruit fly *Drosophila melanogaster* (~14,000 genes, (M. D. Adams et al., 2000)). After such observation some questions arose, where in the genome the information of organism complexity is written? Are there differences between organisms in how the DNA is interpreted? Could the RNA give us some clues about it?

Nowadays it is widely recognized that the mechanisms of transcription and pre-mRNA processing are very dynamic and highly regulated that depend on external inputs and/or the cellular state. Several layers of regulation exist to decipher the

information written in the DNA that ultimately will be reflected in a finely tuned gene expression and protein production. The multiple processes involved in the extensive and intricate journey of an RNA molecule, starting from epigenetic modifications encountered in the DNA, like methylation and histone marks, during its transcription, to its processing by mechanisms as alternative splicing, RNA editing or alternative polyadenylation and finally to its export, translation and degradation, have influenced our vision in modern molecular biology about how the DNA is interpreted in the final mRNA that probably is reflected in the organism phenotype and ultimately its complexity.

During my dissertation, I will focus on alternative splicing as a central process involved in increasing transcriptome diversity by selecting the pieces that will compose the final mRNA. Moreover, I will discuss about the growing importance of alternative splicing in plants, with emphasis on the model plant *Arabidopsis thaliana*, and how my research has contributed to this field.

Basics of the splicing process

The removal of introns (intervening sequences) from the precursor mRNA (pre-mRNA) and joining of the exons (exported sequences) that will be part of the final mRNA molecule is known as pre-mRNA splicing (Sharp, 1994). The spliceosome is the molecular machinery in charge to perform the splicing reaction *via* the recognition of sequence elements in the exon-intron borders. The major spliceosome consists of the small nuclear RNAs (snRNAs) U1, U2, U4/U6 and U5 in conjunction with around 170 core accessorial proteins (Will & Luhrmann, 2011). The major spliceosome is involved in the splicing of ~98% of the introns (Will & Luhrmann, 2011). There is also a minor splicesome that excises ~1-2% of the introns, comprising of the snRNAs, U11, U12, U4atac/U6atac and U5, being U5 the only snRNA shared between the major and the minor spliceosomes (Will & Luhrmann, 2005). Notably, in human over 100 additional spliceosomal proteins have been identified in comparison to yeast, some of them having roles in alternative splicing (Will & Luhrmann, 2011). The assembly of the spliceosome occurs in a stepwise manner where each of the steps is subjected to regulation ultimately defining the spliced transcript.

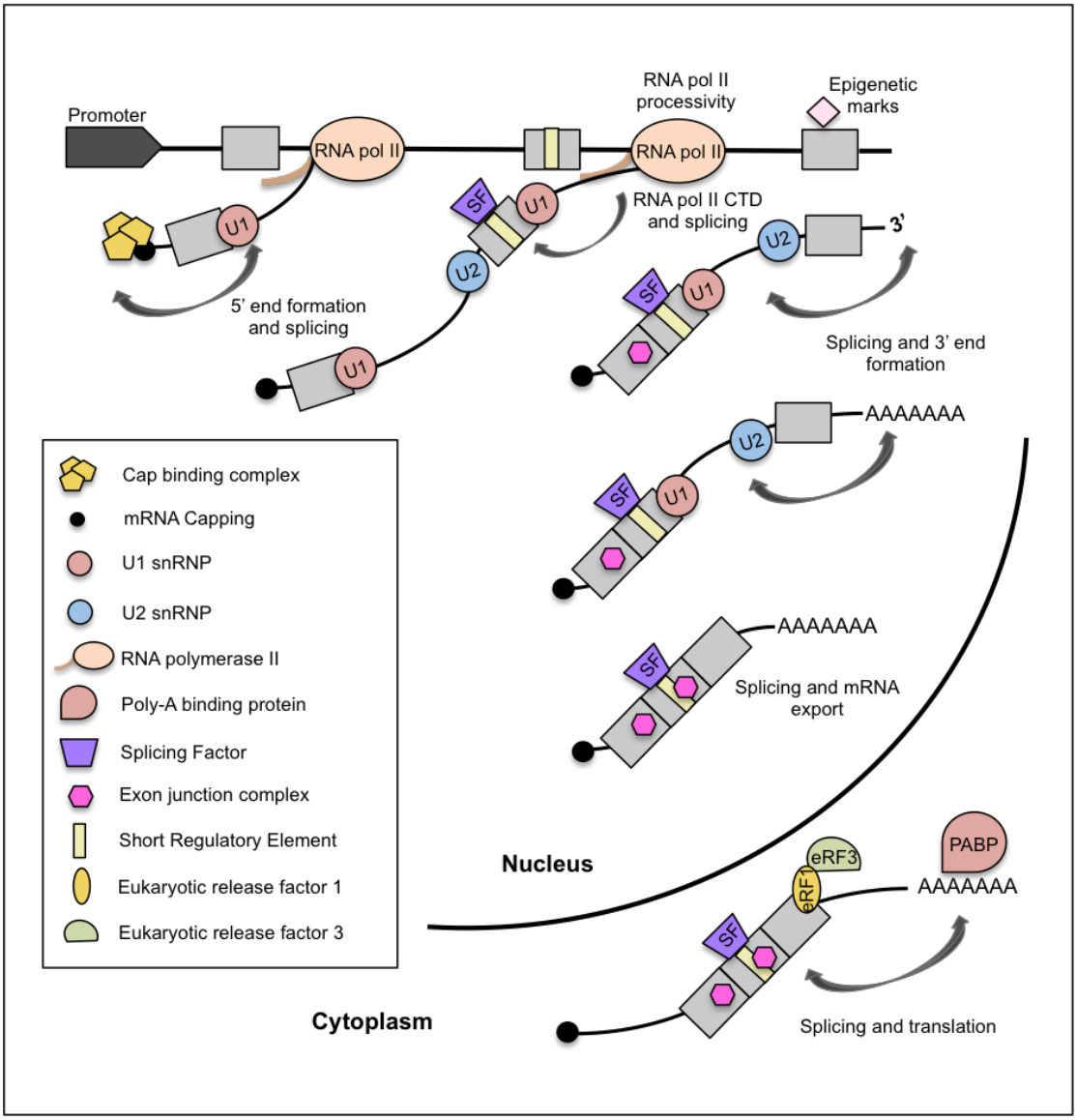


Figure 1. Cross talk between splicing and other RNA processing steps. During transcription different elements in the DNA, like the gene promoter and epigenetic marks, can play a role in the recruitment of splicing factors and RNA pol II processivity affecting the final splicing pattern in the mRNA. Moreover, proteins involved in 5' and 3' end formation can have an effect on splicing adding layers in the regulation of mRNA production. Finally, proteins that are deposited on the mRNA during splicing can influence mRNA export to the cytoplasm and mRNA translation (see text for details).

The splicing process occurs in the nucleus and increasing evidence indicates that co-occurs during transcription (Beyer & Osheim, 1988; Gornemann, Kotovic, Hujer, & Neugebauer, 2005; Kornblihtt, 2005; Kornblihtt, de la Mata, Fededa, Munoz, & Nogues, 2004), which leads to a cross talk of transcription, RNA splicing and other

mRNA processing and gene expression regulatory pathways (Figure 1). From early transcription, the capping of the 5' end promotes spliceosome assembly (Fresco & Buratowski, 1996; Izaurralde et al., 1994; J. D. Lewis & Izaurralde, 1997; J. D. Lewis, Izaurralde, Jarmolowski, McGuigan, & Mattaj, 1996; Schwer & Shuman, 1996) while the splicing of the last intron is coupled with mRNA polyadenylation (Figure 1, (Li, Chen, Wang, Baker, & Krug, 2001; Nesic & Maquat, 1994)). Furthermore exon junction complexes that are deposited on the mRNA during splicing facilitate export of the mRNA to the cytoplasm and therefore enhancing its translation (Figure 1, (Reed, 2003)). In addition, epigenetic marks such as methylation of DNA and covalent modification of histone proteins can influence the splicing process by affecting the speed of RNA pol II during transcription or modulating spliceosome assembly during exon-intron recognition (Figure 1, (Batsche, Yaniv, & Muchardt, 2006; Luco et al., 2010)).

The first step in the splicing reaction is the *bona-fide* identification of the exon-intron boundaries in the pre-mRNA. Such boundaries are defined by specific sequences known as splice sites. The donor site or 5' splice site is around nine nucleotides (nt) long sequence and it determines whether the major or the minor spliceosome would assemble and process the intron (Figure 2, (Sun & Chasin, 2000; M. Q. Zhang, 1998)). The 5' splice site sequence is bound by U1 snRNA (or U11 for the minor spliceosome) via sequence complementarity in the first step of the reaction (Sun & Chasin, 2000; M. Q. Zhang, 1998). In later steps, this sequence will be also bound by U5 and U6 snRNAs (or U6 atac for the minor spliceosome) (Sun & Chasin, 2000; M. Q. Zhang, 1998). The acceptor site or 3' splice site is defined by two sequence elements the actual intron-exon border and the intronic polypyrimidine tract (PPT) that is situated immediately upstream (Figure 2). The length and the amount of pyrimidines in the PPT can vary between introns playing a role in their recognition efficiency (Hertel, 2008). The PPT is a binding platform for U2AF protein that is an auxiliary factor of U2 snRNA, which recognizes another crucial intronic sequence during early steps of the splicing process, the branch point sequence (BPS) (Figure2, (Query, Moore, & Sharp, 1994; Reed, 1996)). Therefore, the basis of the splicing mechanism is the recognition and base pairing of the different snRNAs, components of small nuclear ribonucleic particles (snRNPs), to the nascent RNA molecule. The extent of complementarity of the different sequence elements to the snRNAs determines the strength of the splice site and consequently dictating how efficiently the introns are recognized and excised out (see alternative splicing below).

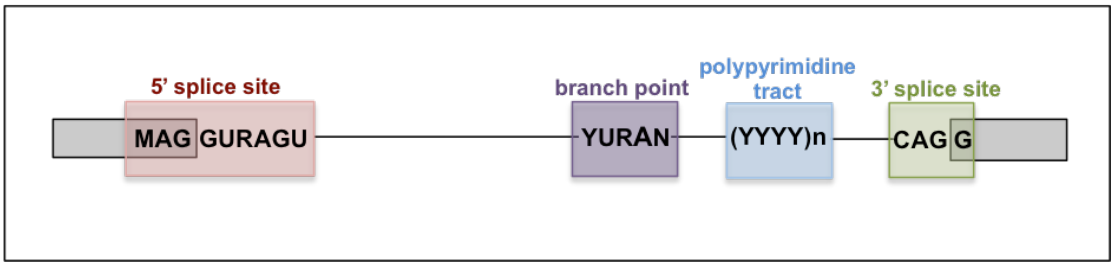


Figure 2. Core splicing signals in the pre-mRNA. Basic sequence elements in the pre-mRNA recognized by the spliceosome during the splicing reaction. Exons are depicted as gray boxes while the intron as a black line. For simplicity, only the consensus sequences for the major spliceosome are illustrated.

The core components of the spliceosome machinery are conserved from yeast to human; nevertheless the recognition of exons and introns can vary greatly. In yeast the exons are longer than introns (Ruby & Abelson, 1991), while in humans, exons are fairly small in comparison to the big sized introns that can be several kilobases long (Black, 2003). It has been observed that the length of exons and introns affects their recognition and spliceosome assembly (Berget, 1995). Therefore, variations in gene structures between species may account for the observed differences in the splicing process in terms of exon-intron recognition.

Two mechanisms of exon-intron recognition have been proposed, the intron definition and the exon definition models (Figure 3). In the intron definition model, applied for short introns, the excision of the intron occurs when the spliceosome components are assembled at the splice sites of the same intronic unit (Figure 3A). In the case for long introns, the exon definition model is proposed; where once the spliceosome components are assembled in the splice sites across the exon they become paired and activated for intron excision when the sequential exon (with its respective spliceosome components) is juxtaposed (Figure 3B). Both models of exon-intron definition have been supported by experimental evidence ((Berget, 1995; M. Guo, Lo, & Mount, 1993; Robberson, Cote, & Berget, 1990; Sterner, Carlo, & Berget, 1996; Talerico & Berget, 1990, 1994)).

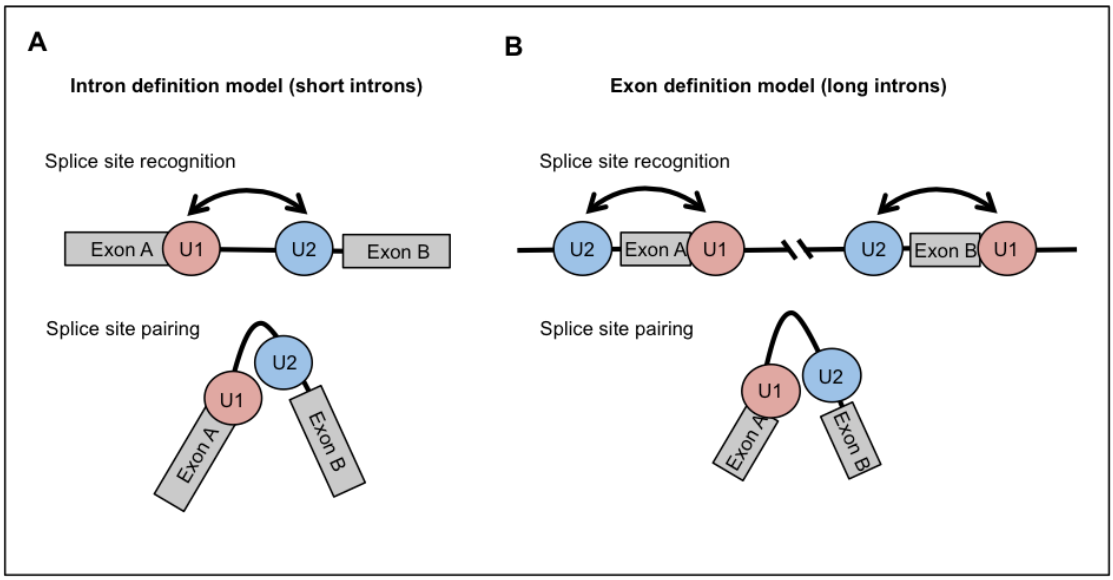


Figure 3. Intron and exon definition models of splicing. A) Intron definition model, the splice sites recognition occurs in the same intronic unit for subsequent intron removal. This model has been proposed for short introns. B) Exon definition model, the splice sites recognition occurs first across the exons and then splice site pairing occurs for intron excision. This model has been proposed for splicing of long introns.

Besides the splice sites recognition, other sequence elements in the pre-mRNA play a role in determining the splicing pattern in the mRNA molecule. These sequence elements are named exonic splicing enhancers/silencers (ESEs or ESSs) and intronic splicing enhancers/silencers (ISEs or ISSs), which are recognized by proteins that can recruit or inhibit the action of core spliceosome elements affecting exon-intron recognition. The ESEs are recognized mainly by the Serine/arginine (SR) rich protein family, which facilitates exon recognition by recruiting splicing components like U1 or U2 snRNPs via their RS domain (Black, 2003; Graveley, 2000; Hertel, Lynch, & Maniatis, 1997). Splicing silencers are less studied, however their impact on intron definition has been clearly recognized. Heterogeneous nuclear ribonucleoproteins (hnRNPs), like the polypyrimidine tract binding protein (PTB), bind these sequences interfering with the spliceosome assembly (Bothwell et al., 1991; Garcia-Blanco, Jamison, & Sharp, 1989; House & Lynch, 2006; Tange, Damgaard, Guth, Valcarcel, & Kjems, 2001) or by looping out the exon (Martinez-Contreras et al., 2006). The existence of splicing enhancers and silencers add complexity to the regulation of the splicing process, as both enhancers and silencers can coexist in the same pre-mRNA molecule and the final splicing output will ultimately depend on

competition and availability of the proteins that recognize them (Hertel, 2008; X. H. Zhang, Arias, Ke, & Chasin, 2009; J. Zhu, Mayeda, & Krainer, 2001). Additionally, the pre-mRNA can adopt secondary structures depending on its nucleotide composition or protein components that associate with it (Hiller, Zhang, Backofen, & Stamm, 2007; Shepard & Hertel, 2008). Such non-linear structures in the pre-mRNA can further influence spliceosome assembly as splice sites or regulatory sequences can be involved in the RNA fold making them unavailable for the RNA-protein complexes, resulting in an enhancing or inhibitory effect in the splicing reaction (Hiller et al., 2007; May, Olson, McManus, & Graveley, 2011; Shepard & Hertel, 2008).

Alternative splicing drives transcriptome diversity

The selection of splice sites during the splicing reaction can occur in a constitutive or alternative manner. In constitutive splicing, the same splice site pair of a given exon is selected each time splicing occurs. In the case of alternative splicing (AS), there is a differential selection of splice site pairs that yields the inclusion or exclusion of particular regions in the final mRNA (Figure 4A). Four main types of alternative AS have been described: 1) exon skipping/inclusion, 2) alternative 5' splice site, 3) alternative 3' splice site and 4) intron retention (Figure 4B). The frequency of each type of AS event can vary greatly between organisms, for example in mammals the most frequent event is exon skipping, while in plants is intron retention. Notably, several AS events can occur in the same pre-mRNA molecule increasing greatly the possibility of combination and consequently the spectrum of mRNA variants. The distinct mRNAs obtained from the same gene template can present different stability or generate proteins with diverse functional properties. The selection of the alternative splice sites is highly regulated allowing that only specific mRNA variants are produced in a given condition. A remarkable example of alternative splicing occurs in the *Dscam* gene of the fruit fly *D. melanoganster,* where the coordinated selection of different combinations of alternative splice sites can lead to the production of 38,016 mRNAs (Neves, Zucker, Daly, & Chess, 2004; Schmucker et al., 2000). The availability of the multiple isoforms allows that every neuron in the fly presents a unique set of proteins that is necessary for a proper neural circuit development (Hattori, Millard, Wojtowicz, & Zipursky, 2008). This example emphasizes the physiological relevance of the different protein variants generated by alternative splicing.

Recent studies using human RNA-Seq data have revealed that ~95% of the protein coding genes are alternatively spliced and the production of some isoforms can be regulated in a tissue-specific manner (Pan, Shai, Lee, Frey, & Blencowe, 2008; E. T. Wang et al., 2008) indicating a major role of alternative splicing in transcript diversity. Moreover, human transcriptome data from different tissues and developmental stages have allowed starting to understand the splicing code of alternative exons, the combination of *cis*-elements in the pre-mRNA that can determine the alternative exon patterns observed in a specific tissue or condition (Barash et al., 2010; Z. Wang & Burge, 2008). From these studies it became evident that the tissue-specific alternatively spliced transcripts are the result of a coordinated interplay of proteins that recognized pre-mRNA sequence elements and that an incorrect interpretation of such elements can lead to human disease (Barash et al., 2010; Z. Wang & Burge, 2008). Later on, analysis of the features in human tissue-specific alternative exons indicated that they have an important function in rewiring protein-protein interaction networks, highlighting their contribution to the observed tissue-specific phenotype (Buljan et al., 2012; Ellis et al., 2012; Weatheritt & Gibson, 2012).

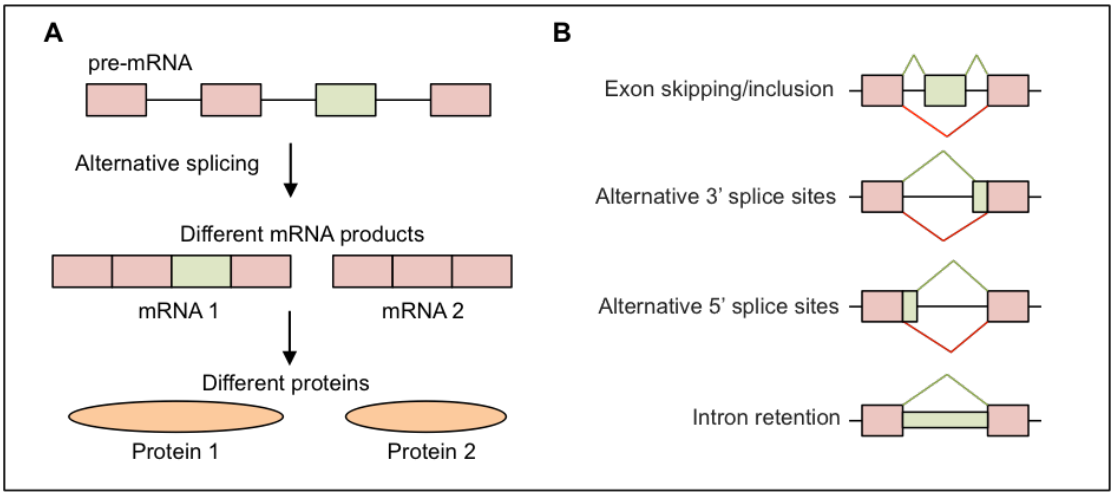


Figure 4. Alternative splicing. A) Differential processing of exons by alternative splicing (by selection of alternative splice sites) leads to the production of different mRNA products increasing the repertoire of transcript and protein variants from the same gene. B) Common types of alternative splicing events. Constitutive regions are depicted in red color, while alternative regions are shown in green.

Remarkably, the obtainment of the alternative splicing profiles of different tissues and organisms using RNA-Seq, led to the observation that an increase in the alternative

splicing landscape complexity correlates with tissue and species complexity, suggesting that alternative splicing played a role during speciation (Barbosa-Morais et al., 2012; Merkin, Russell, Chen, & Burge, 2012). The biological relevance of many of the isoforms that are produced by AS is still far to be understood, but how organisms with similar number of protein coding genes achieve different levels of complexity can be explained at least to some extent by the multiple transcripts that can be generated by alternative splicing.

Cross-talk between alternative splicing and other RNA processing pathways

In addition to the sequence elements in the pre-mRNA, other important factors are just starting to emerge in terms of alternative splicing regulation that undoubtedly brings more complexity to the AS picture. Recent evidence indicates that splicing occurs mainly co-transcriptionally (Figure 1), although some alternative introns can also be processed post-transcriptionally (J. Han, Xiong, Wang, & Fu, 2011). The concept of the co-transcriptional nature of splicing brought along ideas of its regulation at the DNA level. Two models have been proposed linking AS with gene promoter architecture and RNA transcription. The first model (recruitment model) states that distinct promoters can differentially recruit splicing factors to the transcriptional machinery and therefore affecting the final splicing pattern (Braunschweig, Gueroussov, Plocik, Graveley, & Blencowe, 2013; Kornblihtt, 2007). In the second model (kinetics model), variations in the protein complexes that assemble in different promoters affect the elongation rates of RNA Pol II, in a way that a slower kinetics of RNA Pol II will increase the chance of “weak” splicing signals to be recognized by the corresponding splicing factors (Braunschweig et al., 2013; Kornblihtt, 2007)

Characteristics of exons and introns seem also to play an important role during splicing. One of the first observed features that probably contribute to an efficient recognition of exons and introns is their distinct nucleotide composition, namely that exons are significantly GC richer than introns. This property was observed first in plants (Goodall & Filipowicz, 1989) and later also detected in humans (Amit et al., 2012). Moreover, some evidence suggests that the elongation rate of the RNA pol II is slower in presence of AT rich regions (Harada et al., 1999). Besides the differences in nucleotide composition between exons and introns, nucleosome occupancy also diverges between exons and introns, where nucleosomes with

specific histone modifications are generally enriched in exonic sequences (Andersson, Enroth, Rada-Iglesias, Wadelius, & Komorowski, 2009; Schwartz, Meshorer, & Ast, 2009; Tilgner et al., 2009). Two scenarios about the influence of nucleosome positioning on splicing have been suggested. The first one states that the RNA pol II "bumps" into the nucleosomes during the elongation process and this slows down the RNA pol II elongation rates allowing the spliceosome to recognize the splice sites (Hodges, Bintu, Lubkowska, Kashlev, & Bustamante, 2009). This scenario agrees with observations in humans where nucleosome occupancy correlates with the inclusion levels of an exon or an intron in the transcript. Introns have the lowest nucleosome occupancy whereas constitutive exons have the highest one (Schwartz et al., 2009). The question of whether indeed nucleosome enrichment in the exons facilitates their recognition by the splicing machinery is still under investigation. The relatively constant exon size observed in many organisms (between 125 bp and 165 bp) fits well with the DNA size that is wrapped around one nucleosome (close to 150 bp) suggesting that this might have influenced the evolution of gene structure in mammals, where introns have become considerably longer and the exon definition pathway is preferred (Keren, Lev-Maor, & Ast, 2010), or that it could have a more general impact on splicing evolution. The second scenario states that nucleosomes present on exons have particular histone modifications that favors the recruitment of splicing factors and facilitates exon recognition (Schwartz et al., 2009; Tilgner et al., 2009)). The direct link of histone modifications and splicing has been revealed when by altering histone modifications a change in the splicing pattern was observed (Luco et al., 2010).

Besides features of the DNA or RNA, the C-terminal domain (CTD) of the RNA pol II can also play a role in alternative splicing during transcription. The CTD consists of a series of seven amino acid repeats $Y_1S_2P_3T_4S_5P_6S_7$ where mainly serines can be dynamically phosphorylated. The phosphorylation pattern of the CTD not only has a crucial role in regulating the RNA pol II activity in the transcription process *per se* but, as evidenced by recent findings can affect also the alternative splicing outcome, perhaps by altering the association of splicing factors with the CTD (David, Boyne, Millhouse, & Manley, 2011; David & Manley, 2011; Hsin & Manley, 2012). Moreover *in vitro* studies have shown that a phosphorylated CTD stimulates splicing and spliceosome assembly when the splicing reaction occurs *via* exon definition, implying a mode of interaction between the CTD and the splicing factors that facilitates exon recognition (Zeng & Berget, 2000). Additional work on an *in vivo* system needs to be

done in order to confirm the influence of the CTD phosphorylation status on splice site recognition.

The CTD also serves as platform for the binding of protein factors involved in mRNA processing such as capping and 3' end formation. In yeast, it has been shown that deletion of the cap-binding complex (CBC) resulted in spliceosome formation defects and accumulation of unspliced transcripts (Fresco & Buratowski, 1996; Izaurralde et al., 1994; J. D. Lewis & Izaurralde, 1997; Schwer & Shuman, 1996). CBC interacts with U1 snRNP enhancing the recognition of the first intron and facilitating further spliceosome assembly during transcription, while the lack of mRNA capping reduces the splicing efficiency of the transcript (J. D. Lewis et al., 1996). Altogether, these observations evidenced a link between the CBC and the splicing machinery.

The complete maturation of an mRNA is accomplished when the 3' end is processed by endonucleolytic cleavage followed by its polyadenylation. Splicing and 3' end formation are also tightly related, as sequences in the terminal exon are necessary for processing of the penultimate intron and binding of 3' end processing factors (Li et al., 2001; Nesic & Maquat, 1994). The proper association of splicing factors and 3' end processing factors at the 3' splice site of the last intron stimulate the cleavage and polyadenylation (CPA) of the pre-mRNA and its consequent release from the transcription site (Kyburz, Friedlein, Langen, & Keller, 2006; McCracken, Lambermon, & Blencowe, 2002; Millevoi et al., 2002; Rigo & Martinson, 2008, 2009; Vagner, Ruegsegger, Gunderson, Keller, & Mattaj, 2000; Vagner, Vagner, & Mattaj, 2000). Recently it has been shown that U1 snRNP, which is the most abundant snRNP, plays a crucial role in avoiding premature cleavage and polyadenylation. U1 snRNA binds to cryptic 5' splice sites blocking the association of 3' end processing factors to potential CPA sites and thereby preventing the production of premature terminating transcripts (Berg et al., 2012). Furthermore changing the levels of U1 snRNP also produces differences in the processing of the terminal exon, which is consistent with its crucial role on proper 3' end formation (Figure 1, (Berg et al., 2012)).

In addition to capping and 3' end processing, splicing has also been linked to mRNA export (Figure 1). During splicing, splicing factors associate with the exon-junction complex (EJC) positioned 20-24 nt upstream the exon-junction facilitating further steps of mRNA processing, such as mRNA export, and determining transcript stability ((Kataoka et al., 2000; Le Hir, Izaurralde, Maquat, & Moore, 2000), see NMD

pathway below). In Drosophila it has been shown that besides its role in splicing, the EJC is important for proper localization of specific transcripts during development, as mRNA export factors are also part of this complex (Ghosh, Marchand, Gaspar, & Ephrussi, 2012). Further evidence of coupling between splicing and mRNA export comes from the transcription/export (TREX) complex that is recruited to the mRNA by the splicing machinery (Masuda et al., 2005; Reed & Cheng, 2005). The TREX complex leads to the association of the TAP nuclear export receptor to finally export the mRNA from the nucleus to the cytoplasm (Stutz et al., 2000). Moreover, several splicing factors, for example the SR proteins SF2 and SRp20, are also known to be shuttling between the nucleus and the cytoplasm (Sanford, Gray, Beckmann, & Caceres, 2004; Swartz, Bor, Misawa, Rekosh, & Hammarskjold, 2007; Zhong, Wang, Han, Rosenfeld, & Fu, 2009) and probably influencing the nucleus-cytoplasm export and localization of their target transcripts. Altogether, the evidence suggests that the observed layers of regulation in alternative splicing and the cross-talk of alternative splicing with other steps of mRNA maturation seem to come from the multiple and dynamic roles of the RNA binding proteins involved in each of the RNA processing steps

Alternative splicing in regulating gene expression, the NMD pathway

In addition to producing protein variants, alternative splicing can play an important role in regulating the mRNA levels by generating transcripts containing specific features that targets them to RNA quality control pathways. The most studied pathway of RNA quality control is the nonsense-mediated decay (NMD) pathway that takes place in the cytoplasm after the first round of translation. The NMD pathway exerts its action by degrading a transcript when the translation termination process is perturbed due to the presence of specific protein factors on the mRNA molecule (Behm-Ansmant et al., 2007; Chang, Imam, & Wilkinson, 2007; Isken & Maquat, 2007; Nicholson et al., 2010; Rebbapragada & Lykke-Andersen, 2009; Shyu, Wilkinson, & van Hoof, 2008; Stalder & Muhlemann, 2008). The main proteins involved in the NMD degradation pathway are UPF1, UPF2 and UPF3, which eventually recruit other protein factors involved in mRNA decay (Behm-Ansmant et al., 2007; Chang et al., 2007; Isken & Maquat, 2007; Nicholson et al., 2010; Rebbapragada & Lykke-Andersen, 2009; Shyu et al., 2008; Stalder & Muhlemann, 2008).

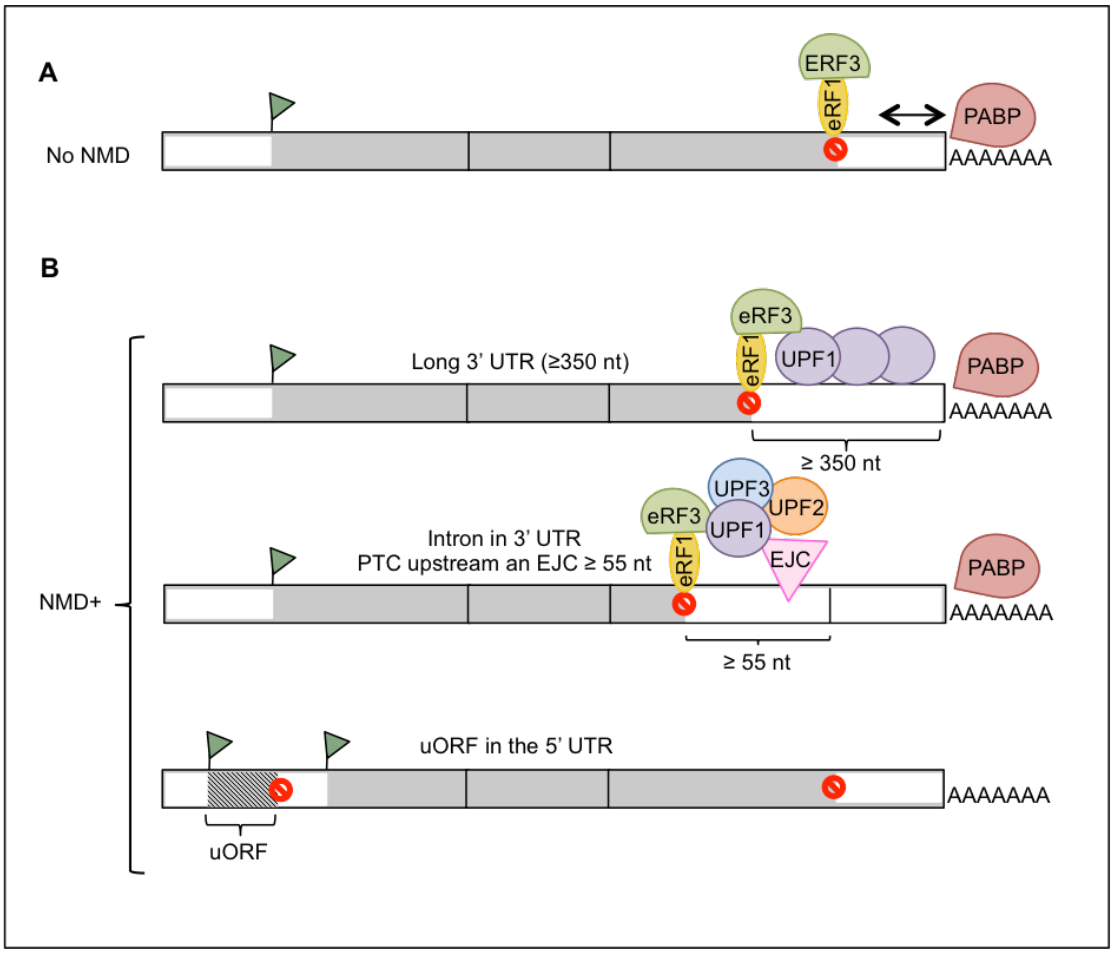


Figure 5. NMD eliciting features. A) In a canonical end of translation the close proximity of the stop codon and the poly-A tail in the mRNA allows the interaction between eRF3 and PABP leading to the peptide release. B) NMD eliciting features. A long 3' UTR (≥350 nt) permits the recruitment of UPF1 interfering with eRF3-PABP interaction and triggering NMD (upper). The presence of an exon junction complex (EJC) due to the presence of a splice junction over 55 nt downstream the stop codon leads to the recruitment of NMD factors, such as UPF1, UPF2 and UPF3, affecting the interaction of eRF3 with PABP and targeting the transcript for degradation (middle). Additionally the presence of upstream open reading frame (uORF) in the 5' untranslated region also triggers NMD (bottom, see text for detailed explanation). Coding sequence is colored in gray while untranslated regions in white. Start codons are represented in green flags, while stop codons are illustrated by red "forbidden" signals.

During translation termination, the release factors eRF1 and eRF3 recognize the stop codon. In a canonical translation termination, a close proximity of the stop codon with the poly A tail of the mRNA allows the interaction between eRF3 and the poly-A

binding protein (PABP) facilitating the release of the newly synthetized peptide (Cosson et al., 2002; Hoshino, Imai, Kobayashi, Uchida, & Katada, 1999). If the interaction between eRF3 and PABP does not occur, then eRF3 interacts with UPF1 instead, which results in the recruitment of NMD components leading to mRNA decay. A long 3' UTR in the mRNA leads usually to this scenario, as a distance of usually more than 350 bp between the stop codon and the poly-A tail site impairs the ERF3-PABP interaction (Figure 5B upper panel). Degradation by NMD due to a long 3' UTR is a common situation observed in invertebrates and yeast mRNAs (Kebaara & Atkin, 2009; Muhlrad & Parker, 1999). Therefore, alternative splicing events that result in transcripts that compromise the ERF3-PABP interaction by producing long 3' UTRs can be targeted to degradation by NMD. Another feature eliciting NMD is the presence of a premature termination codon (PTC) over 50 nucleotides upstream of an splice junction (and its loaded EJC), recruits UPF1 and other NMD related proteins, such as UPF2 and UPF3, targeting the transcript for degradation (Figure 5B middle panel, (Isken & Maquat, 2007)). In line with this NMD feature, occurrence of introns and AS events in the 3' UTRs of the mRNA have an important influence on NMD sensitivity, as AS events that impact the distance between the stop codon and a downstream splice junction in the transcript can affect its stability. In addition, the incidence of introns in the 5' UTR can also impact transcript stability, where AS events can generate or alter the size of upstream open reading frames (uORFs), that when longer of 35 amino acids results in mRNA degradation (Figure 5, (Johansson & Jacobson, 2010; Kalyna et al., 2012; Mendell, Sharifi, Meyers, Martinez-Murillo, & Dietz, 2004; Welch & Jacobson, 1999; Wittmann, Hol, & Jack, 2006; Yepiskoposyan, Aeschimann, Nilsson, Okoniewski, & Muhlemann, 2011)). In general, the stop codons of the uORFs are recognized as PTCs compromising transcript stability. Moreover, the presence of downstream introns from the uORF stop codon can position a splice junction over 50 nt or create a long 3'UTR triggering NMD. In plants, an additional NMD feature was observed, where uORFs that overlap with the authentic start codon results in NMD sensitivity (Kalyna et al., 2012).

As NMD pathway is an mRNA decay mechanism, the detection of NMD sensitive transcripts in a wild type background can be very challenging. Therefore, the use of mutant lines, particularly in the UPF proteins, has allowed investigating the features, including those introduced by alternative splicing, which are present in the transcripts that triggers NMD. Studies in different organisms (Drechsel et al., 2013; Guan et al., 2006; He et al., 2003; Johansson, He, Spatrick, Li, & Jacobson, 2007; Kalyna et al., 2012; Mendell et al., 2004; Rehwinkel, Letunic, Raes, Bork, & Izaurralde, 2005;

Viegas, Gehring, Breit, Hentze, & Kulozik, 2007; Wittmann et al., 2006; Yepiskoposyan et al., 2011), have revealed that not all the transcripts presenting NMD triggering features are necessarily affected, questioning the presence of additional rules in this pathway. Furthermore, in Drosophila it has been observed that not always the excision of an intron results in the deposition of an EJC complex, suggesting that only the splicing of specific introns can lead to NMD. In line with these observations, our lab found that the abundance of intron retention containing transcripts, despite possessing NMD features, did not increase in the *upf* mutants (see chapter and discussion, (Kalyna et al., 2012)). Current estimates are that up to 20-30% of the alternative splicing transcripts can be subjected to decay by the NMD machinery (Baek & Green, 2005; Drechsel et al., 2013; Hillman, Green, & Brenner, 2004; Kalyna et al., 2012; Lareau, Green, Bhatnagar, & Brenner, 2004; B. P. Lewis, Green, & Brenner, 2003). Besides the NMD features introduced in a transcript due to AS events, the cross talk between alternative splicing and NMD is also evidenced through the regulatory feedback loops of important alternative splicing factors, like PTB and the glycine-rich protein 7 (GRP7) in plants where binding of these proteins to their own mRNAs results in production of isoforms that are degraded by NMD (see below).

As multiple alternative splicing events can co-occur in the same transcript it is necessary to analyze the consequences of all the AS events in the transcript to establish its sensitivity for NMD. The rules of how the NMD pathway recognizes transcripts for degradation and the impact of NMD on the final transcript levels and transcriptome complexity are just starting to emerge. The availability of a good reference transcript annotation and the use of NMD impaired mutants will considerably contribute to the knowledge of the link between AS and NMD in regulating gene expression.

Challenges to establish a global map of alternative splicing by high-throughput RNA sequencing

The impact of alternative splicing in increasing transcript diversity in different organisms has been changing dramatically over the past few years. The technological advances for studying transcriptomes have established alternative splicing as a major mechanism for increasing the number of mRNA variants to produce different protein isoforms and for regulating gene expression.

The first estimations for genes undergoing alternative splicing were done mainly by sequencing expressed sequence tags (ESTs) and complementary DNA (cDNA) libraries, but more recently the use of splice junction arrays, tilling arrays and high-throughput sequencing (HTS) technologies are allowing a better resolution of the transcriptome at a global scale. These new technologies have brought considerable advantages to analyze transcriptomes of different tissues and conditions and to enable to explore alternative splicing regulation. The most popular technologies for studying transcriptome profiles have been the use of hybridization arrays and HTS. Both technologies have advantages and disadvantages for exploring alternative splicing. In the case of the arrays prior knowledge of the genome and transcripts is necessary, while HTS allows *de novo* identification of splicing events at a single base resolution but at the cost of generating a large amount of data and posing computational challenges. Nowadays the relatively low cost of HTS and its high resolution without previous knowledge of the isoforms, has established this technology as the preferred one to obtain alternative splicing profiles. The use of high-throughput RNA sequencing (RNA-Seq) led to estimate that the majority (~95%) of the protein coding genes in human is alternatively spliced. This indicates that despite the unexpected low number of protein coding genes in the human genome, the final number of proteins can be increased considerably ((Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Pan et al., 2008; Sultan et al., 2008; E. T. Wang et al., 2008). Importantly, the use of RNA-Seq in plants revealed that alternate splicing is also a major mechanism for generating transcript diversity (Filichkin et al., 2010; S. Guo et al., 2010; Walters, Lum, Sablok, & Min, 2013; Zenoni et al., 2010; G. Zhang et al., 2010). The current estimate for the model plant *Arabidopsis thaliana* is that ~60% of the intron-containing genes undergo AS ((Marquez, Brown, Simpson, Barta, & Kalyna, 2012), see chapter).

A typical RNA-seq experiment starts with RNA extraction from the sample of interest followed by the synthesis of a double stranded cDNA (ds cDNA) library using oligo-dT or random primers that is then prepared for its sequencing (Figure 6). Millions of sequencing reads in the range of 35 to 400 bp can result from a single run and they can be derived from one (single end reads) or both ends (paired end reads) of the cDNA fragment. The number of reads generated by HTS correlates well with the representation of the specific cDNA molecule in the sample, which allows the quantification of the respective transcripts, even when they are present in relatively low amounts (Figure 6, (Garber, Grabherr, Guttman, & Trapnell, 2011)). Notably, the

probability of representation of low abundant transcripts increases with a higher sequencing depth. Two challenges of HTS technology are obvious in terms of data processing, namely the number of reads that is generated and their small size. The small size of the reads has an important impact on distinguishing similar transcripts and their quantification, as short reads can be shared by alternatively spliced isoforms or by transcripts arising from recently duplicated paralogues. Several tools have been developed in order to cope with these problems but a reliable pipeline for transcript assembly and quantification, especially in terms of determining novel AS variants, is still under development (see below).

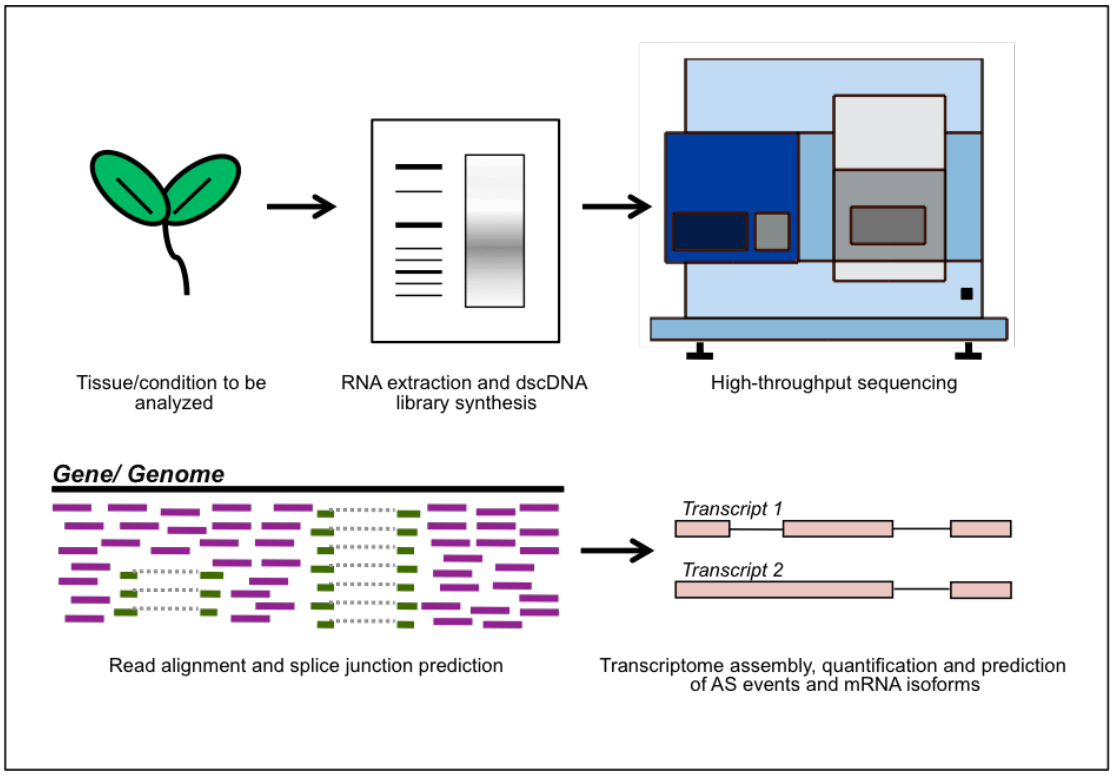


Figure 6. Basic pipeline of an RNA-Seq experiment. RNA extraction and ds cDNA library synthesis is performed for the tissue/condition of interest. Afterwards, the ds cDNA library is prepared for HTS. The resulting sequencing reads are aligned to the genome (genome based strategy, see text) followed by assembly of transcripts and their quantification. Figure adapted from (Reddy, Marquez, Kalyna, & Barta, 2013).

One of the crucial steps for transcriptome profiling of any given sample using RNA-Seq is the correct prediction of splicing events and eventually isoforms, where a gapped alignment that can detect exon-intron boundaries is necessary. Due to the

fact that splice sites have degenerate sequences, the use of a reference genome as scaffold is preferred for splice junction read alignments. Tools as; QPALMA (Jean, Kahles, Sreedharan, De Bona, & Ratsch, 2010), Tophat (Trapnell, Pachter, & Salzberg, 2009), MapSplice (K. Wang et al., 2010), and Splicemap (Au, Jiang, Lin, Xing, & Wong, 2010) use this strategy, although other tools that do not use any reference genome are also available, like Trinity (Haas et al., 2013), TransABySS (Robertson et al., 2010) and Rnnotator (Martin et al., 2010). The genome-based aligners are based on the assumption that reads that cannot completely align to the genome can span splice junction regions. The splitting of the read in several pieces and subsequent alignment represents a computational challenge, as a smaller read size increases its probability to match multiple times to the genome and therefore compromises the reliability of the alignment. Therefore, the use of longer reads is always preferable to improve the quality of the alignment, especially when dealing with highly repetitive genomes or complex transcriptomes. The introduction of paired end reads in the sequencing platforms has helped to enhance the performance of short read aligners. During sample preparation for sequencing, a fractionation step at the level of RNA or cDNA is necessary, where the size distribution of the resulting fragments is known, hence the distance between the sequenced paired end reads can be calculated and used as information for the alignment.

The ultimate goal of an RNA-Seq experiment is to determine the genes expressed in a given sample and ideally also the expression profile of the transcript variants of a given gene. Most of the initial programs for gene expression profiling using HTS were based on transcript annotation, which limited the discovery of novel genes and isoforms. Nowadays some transcriptome assemblers are available using HTS, and two main strategies have been adopted for this task. The first one builds transcripts using known isoforms as a scaffold, while the second strategy consists in a completely *de novo* assembly without using any previous knowledge (Garber et al., 2011). The trancriptome assembly employing a known annotation results in a better accuracy, but for newly sequenced organisms this approach is not always a possibility (Garber et al., 2011). Nevertheless all the assemblers still face the problems inherent to HTS, being the small size of the reads the main challenge to determine from which transcript each read is coming. In the case of alternatively spliced transcripts most of the isoforms have minor sequence differences so the accuracy of transcript determination can also impact its quantification. The improvement of the sequencing technologies, especially in terms of read length and error rates will alleviate considerably the challenge of transcriptome profiling.

Undoubtedly, an accurate transcript expression profile will contribute to elucidate the principles governing the regulation of gene and isoform expression in a given condition.

The splicing process in plants and the impact of alternative splicing in the plant transcriptome.

Most of the knowledge about the splicing process and its regulation has been obtained from studies in yeast and mammalian systems. In this chapter, I will briefly introduce the specificities of splicing in plants, with an emphasis on the model plant Arabidopsis, and our contributions to this field.

Most of the genes in Arabidopsis (~80%) have at least one intron. Arabidopsis introns are considerably short (average size around 160 nt) when compared to mammals and they have a tendency of being uridine-adenosine (UA) rich (Iwata & Gotoh, 2011; Marquez et al., 2012). Importantly, this UA richness, and in particular the presence of UA or U stretches, was found to be an important determinant on intron splicing efficiency, whereby introns with higher UA content are more effectively spliced (Goodall & Filipowicz, 1989). The lack of an *in vitro* plant splicing system has limited the possibility to study the splicing reaction in plants. Nevertheless, some hints about the plant splicing machinery came when plant introns were shown to be spliced in HeLa system (Brown, Feix, & Frendewey, 1986; Hartmuth & Barta, 1986), but human introns were not spliced *in planta* (Barta et al., 1986). Therefore, although the core splicing signals, as the 5'/3' splice sites and BP are comparable in plants and animals (Iwata & Gotoh, 2011), some specific features in plant introns must exist when compared to animal introns that might explain the inability of plant systems to efficiently splice animal introns. Nevertheless, the finding that plant introns are successfully excised in mammalians systems implied that the core components of the spliceosome are present in plants.

The release of the complete genome sequence of *A. thaliana* made possible to identify the complete set of the orthologous genes of the spliceosomal components. Interestingly, more splicing related genes were found in Arabidopsis genome when compared to the human genome (Koncz, Dejong, Villacorta, Szakonyi, & Koncz, 2012; Reddy et al., 2013; Silverman, Li, & Gregory, 2013; B. B. Wang & Brendel, 2004). The vast majority of these genes are sharing a high percent of homology,

suggesting that they very likely originated in the course of Arabidopsis genome duplications (Adams & Wendel, 2005; Cui et al., 2006; Kalyna & Barta, 2004). The presence of paralogues of splicing factors has raised the question of whether they have overlapping functions or whether they have acquired specific ones. The functional roles of the identified putative splicing related genes in plants still require further investigation, especially for those factors that are lacking an animal counterpart. In particular, the research on plant splicing has been focused on determining the functional properties of hnRNPs and SR proteins.

hnRNPs are a heterogeneous group of proteins that have affinity to RNA through their RRMs, while their additional motifs, like the KH homology domain or the glycine-rich motif give them binding specificities (S. P. Han, Tang, & Smith, 2010; Martinez-Contreras et al., 2007; Wachter, Ruhl, & Stauffer, 2012). Most of the hnRNPs are localized in the nucleus, but some of them also possess nuclear-cytoplasm shuttling activity indicating a more extensive role in RNA processing (Wachter et al., 2012). The first hnRNPs-like proteins characterized in plants were UBP1, RBP45 and RBP47, where only UBP1 stimulates intron splicing (Lambermon et al., 2000; Lorkovic, Wieczorek Kirk, Klahre, Hemmings Mieszczak, & Filipowicz, 2000). UBP1 seems to associate with the 3' end of the pre-mRNA and protect it from degradation, but the mechanism by which it impacts splicing is still not clear. The proteins UBA1a (UBP1-associated protein 1a) and UBA2a (UBP1-associated protein 2a) have been identified to associate with UBP1 and they preferentially bind to U-rich sequences *in vitro* but they themselves do not enhance splicing (Lambermon et al., 2000). The best characterized hnRNP-like proteins in plants are the Glycine-rich proteins (GRPs), GRP7 and GRP8 and the polypyrimidine binding proteins PTB1, PTB2 and PTB3.

GRP7 and GRP8 are hnRNPs whose expression is regulated by the circadian clock (Carpenter, Kreps, & Simon, 1994; Heintzen et al., 1994; Staiger, 2001; Staiger, Zecca, Wieczorek Kirk, Apel, & Eckstein, 2003). These RNA binding proteins regulate the production of their own alternatively spliced isoforms by binding to their own pre-mRNAs and targeting these transcripts for degradation by the NMD pathway (Staiger, Zecca, Wieczorek Kirk, Apel, & Eckstein, 2003). A recent study employing the high-resolution RT-PCR panel showed that GRP7 affects the splicing patterns of at least 59 genes in Arabidopsis (Streitner et al., 2012). RNA immunoprecipitation experiments confirmed the binding of GRP7 to seven of its putative targets via its RNA recognition motif (RRM) (Streitner et al., 2012). Remarkably, the observed AS

changes were not exclusively affected by GRP7, but some of them were shared with the cap binding complex proteins (CBC) and various SR proteins, suggesting a cross-talk of different splicing and pre-mRNA processing factors to determine the final AS outcome (Streitner et al., 2012).

Another group of hnRNP-like proteins that have been studied in plants are the three genes in Arabidopsis that code for the PTB proteins. As in the case for GRP7 and GRP8, they also present a negative auto-regulation by binding to their own pre-mRNAs and target the resulting transcripts for NMD mediated degradation (Stauffer, Westermann, Wagner, & Wachter, 2010), a negative feedback regulatory loop that is also observed for the PTB proteins in mammals (Grabowski, 2007). A genome-wide study using overexpression and artificial micro RNA lines of the three PTB genes in Arabidopsis showed that PTB1 and PTB2 affect the splicing profile of around 300 genes, whereas for PTB3 no major effect was observed, implying that PTB3 has a divergent role from PTB1 and PTB2. In contrast to human PTB where exon skipping was the main type of AS affected, PTB1 and PTB2 in Arabidopsis affect all types of AS (Ruhl et al., 2012), but still their mechanism of action for determining the observed AS landscape remains to be elucidated.

SR proteins are proteins that have a particular protein domain composition, they contain one or two RRMs, followed by an argine-serine (RS) rich region and in some cases they also possess zinc knuckles (Barta, Kalyna, & Reddy, 2010). SR proteins are involved mainly in the regulation of constitutive and alternative splicing by binding to specific sequences in the pre-mRNA and their action is regulated in a phosphorylation dependent manner (Long & Caceres, 2009). In animals it has been shown that these proteins also have other functions in mRNA export, mRNA stability, mRNA translation, genome stability and chromatin binding (Long & Caceres, 2009; Risso, Pelisch, Quaglino, Pozzi, & Srebrow, 2012; Shepard & Hertel, 2009). In Arabidopsis there are 18 SR proteins (compared to only 12 in human) where many of them are located in duplicated genomic regions, but presenting differences in their spatial-temporal expression (M Golovkin & Reddy, 1999; M. Golovkin & Reddy, 1998; Kalyna & Barta, 2004; Kalyna, Lopato, & Barta, 2003; Lazar & Goodman, 2000; Lopato et al., 2002; Lopato, Gattoni, Fabini, Stevenin, & Barta, 1999; Lopato, Kalyna, et al., 1999; Lopato, Waigmann, & Barta, 1996) implying specialized functions. In plants, SR proteins are classified in six subfamilies based on their protein domain composition, where three subfamilies (SF2/ASF, 9G8 and SC35) are having members with homology to human SR proteins, and three subfamilies being

plant specific (SCL, RS and RS2Z) (Barta et al., 2010). As in the case of hnRNPs, the genes coding for SR proteins can be also alternatively spliced and remarkably their splicing patterns are conserved even in very ancient plants like the mosses (Kalyna & Barta, 2004; Kalyna, Lopato, Voronin, & Barta, 2006), implying that alternative splicing has a critical role in their regulation. Moreover, it has been found that some SR proteins can interact with each other and with many other components of the spliceosome (M Golovkin & Reddy, 1999; M. Golovkin & Reddy, 1998 Lopato et al., 2002;(Barta, Kalyna, & Lorkovic, 2008; Gullerova, Barta, & Lorkovic, 2006; Reddy, 2007) adding complexity to the mechanism by which they regulate splicing. The availability of RNA-Seq and RIP-Seq of the different SR proteins will help in the identification of their targets and add valuable information about their roles in shaping plant transcriptomes.

In plants, the impact of alternative splicing on the transcriptome has just started to being elucidated. Not many years ago, it was suggested that plants did not use alternative splicing as a major generator of transcript diversity and that such diversity was probably achieved by the presence of gene families with multiple members having different degrees of homology (Barbazuk, Fu, & McGinnis, 2008; Brett, Pospisil, Valcarcel, Reich, & Bork, 2002; Moore & Purugganan, 2005; Ober, 2005). Nevertheless, the presence of all the spliceosome components and plant specific splicing regulators seemed to contradict those implications. In Arabidopsis, the first estimation stated that only 1.2% of the genes undergo alternative splicing based on ESTs libraries (W. Zhu, Schlueter, & Brendel, 2003). In the following years, the increased amount of ESTs led to the identification of more alternatively spliced variants, raising the estimate to up to 30% of alternatively spliced genes (Barbazuk et al., 2008; Campbell, Haas, Hamilton, Mount, & Buell, 2006; Iida et al., 2004). The clear dependence of a comprehensive transcriptome data for observing splicing variants hinted that the low frequency of alternative splicing in plants was a data issue. However, the importance of AS in plants became evident from several independent studies, which showed that genes with important functions in development, response to stress and response to environmental cues are affected by alternative splicing (Reddy et al., 2013; Staiger & Brown, 2013). The advent of HTS provided considerable advance in determining the global impact of AS in plants. In 2010 the first RNA-Seq analyses of Arabidopsis and rice, revealed that ~42% and ~33% of genes, respectively, showed evidence of AS (Filichkin et al., 2010; G. Zhang et al., 2010), implying that AS is a more prevalent mechanism in plants than previously thought.

In this work, I will present the contribution of my colleagues and me in determining the alternative splicing landscape in Arabidopsis. During my PhD studies I was involved in determining the cross talk of AS and NMD in regulating gene expression (chapter), the extent of AS in Arabidopsis at genome-wide level under normal growth conditions (chapter) and finally the discovery and the characterization of a new AS event (chapter).

Publications

Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis.

Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., . . . Brown, J. W. (2012).

Nucleic Acids Res, 40(6), 2454-2469.

Author contribution: Validation of AS variants detected by the HR RT-PCR panel using RNA-Seq

2454–2469 Nucleic Acids Research, 2012, Vol. 40, No. 6 Published online 29 November 2011
doi:10.1093/nar/gkr932

Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*

Maria Kalyna[1], Craig G. Simpson[2], Naeem H. Syed[3], Dominika Lewandowska[2], Yamile Marquez[1], Branislav Kusenda[1], Jacqueline Marshall[3], John Fuller[2], Linda Cardle[2], Jim McNicol[4], Huy Q. Dinh[5], Andrea Barta[1,*] and John W. S. Brown[2,3,*]

[1]Max F. Perutz Laboratories, Medical University of Vienna, Dr. Bohr-Gasse 9/3, A-1030 Vienna, Austria, [2]Cell and Molecular Sciences, The James Hutton Institute, [3]Division of Plant Sciences, University of Dundee at JHI, [4]Biomathematics and Statistics Scotland at JHI, Invergowrie, Dundee DD2 5DA, Scotland, UK and [5]Center for Integrative Bioinformatics, Max F. Perutz Laboratories, Medical University of Vienna, University of Vienna, University of Veterinary Medicine Vienna, Dr. Bohr-Gasse 9/3, A-1030 Vienna, Austria

Received July 26, 2011; Revised September 22, 2011; Accepted October 10, 2011

ABSTRACT

Alternative splicing (AS) coupled to nonsense-mediated decay (NMD) is a post-transcriptional mechanism for regulating gene expression. We have used a high-resolution AS RT–PCR panel to identify endogenous AS isoforms which increase in abundance when NMD is impaired in the *Arabidopsis* NMD factor mutants, *upf1-5* and *upf3-1*. Of 270 AS genes (950 transcripts) on the panel, 102 transcripts from 97 genes (32%) were identified as NMD targets. Extrapolating from these data around 13% of intron-containing genes in the *Arabidopsis* genome are potentially regulated by AS/NMD. This cohort of naturally occurring NMD-sensitive AS transcripts also allowed the analysis of the signals for NMD in plants. We show the importance of AS in introns in 5′ or 3′UTRs in modulating NMD-sensitivity of mRNA transcripts. In particular, we identified upstream open reading frames overlapping the main start codon as a new trigger for NMD in plants and determined that NMD is induced if 3′-UTRs were >350 nt. Unexpectedly, although many intron retention transcripts possess NMD features, they are not sensitive to NMD. Finally, we have shown that AS/NMD regulates the abundance of transcripts of many genes important for plant development and adaptation including transcription factors, RNA processing factors and stress response genes.

INTRODUCTION

Alternative splicing (AS) is an important mechanism to control gene expression and increase the proteome complexity of higher eukaryotes (1–3). Regulated AS drives developmental pathways and responses to environmental pressures. Following transcription, splicing of the exons requires removal of introns by assembling a large RNP complex, the spliceosome, with five snRNPs and about 180 proteins (4). Splice site selection has to be precise but consensus sequences defining splice sites are degenerate and how a splice site is selected from many similar sites within a transcript remains a major question. In many cases, specific splice sites are used in all transcripts (constitutive splicing) while in alternative splicing, other splice sites are used to various extents giving rise to alternate transcripts with variable sequences. It is now well established that in addition to splice sites, sequence elements within exons and introns, termed either splicing enhancers or silencers are binding sites for splicing factors which either enhance or repress splicing depending on their activities (5,6). These splicing regulators are, for example, SR and hnRNP protein families, and other cell-, stage- or tissue-specific proteins involved in constitutive and alternative splicing which establish the splicing code and determine which splice site is selected (7–10). The regulation of alternative splicing is brought about by the relative levels of the RNA-binding proteins determining how efficiently different splice sites are used to generate more than one spliced mRNA from one gene.

Alternatively spliced mRNA variants can produce functionally different protein isoforms with altered amino acid

*To whom correspondence should be addressed. Tel: +43 1 4277 61640; Fax: +43 1 4277 9616; Email: andrea.barta@meduniwien.ac.at
Correspondence may also be addressed to John W. S. Brown. Tel.: +44 1382 568533; Fax: +44 1382 562426; Email: j.w.s.brown@dundee.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

sequences and protein domains resulting in changes in activity, localization, interaction partners or post-translational modifications (1,11). In addition, alternative splicing can regulate mRNA levels through the targeted degradation of specific AS isoforms by nonsense-mediated decay (NMD) (see below). In particular, alternative splicing can result in mRNAs with premature termination codons (PTCs) which could give rise to truncated proteins which are detrimental to cell survival and energy costly for the cell. RNA quality control mechanisms have evolved at all levels of gene expression to identify and remove aberrant RNA transcripts. One of the best investigated mRNA quality control mechanisms is NMD which degrades mRNAs which possess a premature termination codon (PTC+) and other physiological mRNAs without a PTC such as transcripts with long 3′-UTRs [reviewed in (12–18)]. Despite great advances in understanding of the NMD pathway, it is apparent that not every PTC triggers NMD and that this pathway controls the abundance of certain mRNAs which do not contain known NMD features, arguing that not all the factors inducing NMD have been identified yet.

Several features of NMD-sensitive, PTC^+ transcripts have been elucidated and have led to models of how PTCs are recognized and degradation triggered. In the current model for mammals, NMD initiates the rapid decay of a transcript if translation termination is perturbed [reviewed in (12–18)]. Efficient translation termination of the ribosome is proposed to involve the interaction of the release factor, eRF3, and poly(A) binding proteins (PABP) on the poly(A) tail of the mRNA. If this interaction is prevented or impaired by, for example, an unusually long 3′-UTR, the eRF3 on the ribosome will bind UPF1 which then recruit UPF2 and UPF3, all core NMD proteins. This functional NMD complex (which includes many other proteins) then elicits the phosphorylation of UPF1 and rapid degradation of the transcript. This 'long 3′UTR' mechanism is characteristic for transcripts in invertebrates and yeast. In mammals, the NMD response triggered by a ribosome terminating at a PTC is stimulated by UPF3 associated with a downstream exon-junction complex (EJC) which is deposited on the mRNA 20–25 nt upstream of a spliced exon–exon junction (19,20). In the course of splicing the EJC complex binds the NMD factors UPF2/UPF3 which can then associate with a ribosome terminating at a PTC upstream of the EJC which has recruited UPF1 in the SURF complex (SMG1-UPF1-eRF1-eRF3) (21). On a normal, non-PTC-containing mRNA, the EJC is removed in the first round of translation (22) except when the EJC is located in the 3′-UTR. This is consistent with the observation that introns in the 3′-UTR may significantly enhance NMD. Thus, in mammals, both the length of the 3′-UTR and the presence of an EJC complex downstream of a PTC can trigger NMD. However, the recent demonstration in *Drosophila* that EJCs are not deposited at each splice junction such that only some introns are able to trigger intron-dependent NMD (23) suggests that NMD may rely more widely on long 3′UTR signals.

The NMD pathway in plants is as yet not well characterized (24,25). Plants possess orthologues of the key eukaryotic NMD proteins, UPF1, UPF2, UPF3 and SMG-7 (but not SMG-1, SMG-5 or SMG-6) and these have been shown to be involved in degrading mRNAs with PTCs (26–31). Efforts to determine the rules for NMD substrates suggest that like mammals, plants are able to recognize different types of PTC-containing transcripts (Figure 1c). Firstly, it was shown that both long 3′-UTRs and introns located in 3′-UTRs are signals for efficient NMD (30,32–34). This indicates that like in invertebrates and yeast, the distance between a stop codon and the PABP on the poly(A) sequence is important and that translation termination is also likely to require the interaction of the release factor-containing ribosome with PABP. Secondly, EJC components, which are required for the intron-based NMD mechanism proposed for mammalian NMD, have been demonstrated to be similarly important in plants (28). Furthermore, upstream open reading frames (uORFs) of more than 35 amino acids can trigger NMD in plants (35).

Alternative splicing in plants is an important regulatory process for plant development and for the response of plants to environmental factors. However, its frequency of occurrence has been grossly underestimated largely due to low depth of sequencing and relatively few available ESTs. The most recent estimate based on next generation sequencing is that about 42% of intron-containing genes undergo alternative splicing (36) and this is still likely to be a significant underestimate. In humans, over 95% of genes undergo alternative splicing and more importantly around 20–30% of alternatively spliced transcripts contain PTCs and are potentially turned over by NMD (37–40). The importance of the link between AS and NMD has been highlighted by the regulation of functional transcript levels of key splicing factors such as SR proteins and PTB through alternative splicing via conserved splice sites (41,42). Conservation of alternative splice sites to produce PTC-containing transcripts has also been demonstrated for SR protein genes in lower and higher plants (43) and more generally, NMD seems to play a regulatory role in gene expression using alternative spliced transcripts (27). The best characterized examples of gene regulation by AS/NMD in plants are GRP7 and GRP8, components of a slave oscillator coupled to the circadian clock. These glycine-rich RNA-binding proteins bind their own pre-mRNAs inducing AS which produces NMD-sensitive transcripts thereby auto- and cross-regulating their mRNA levels (44,45). Other examples of such regulation are SR protein genes (46), polypyrimidine tract binding protein (PTB) (47), possibly SUPPRESSOR OF OVEREXPRESSION OF CO1 (48) and riboswitch-regulated alternative splicing controlling NMD (49).

The definition of the rules of NMD in plants have mainly relied on mutations in a small number of model transcripts and it is necessary to examine how these features correspond to those in NMD-sensitive endogenous transcripts. As about 78% of alternative transcripts in *Arabidopsis* introduced in-frame PTCs more than 55 nt upstream of exon junction, it was speculated that NMD is a widespread mechanism for regulating gene expression (36), however this has not been experimentally addressed.

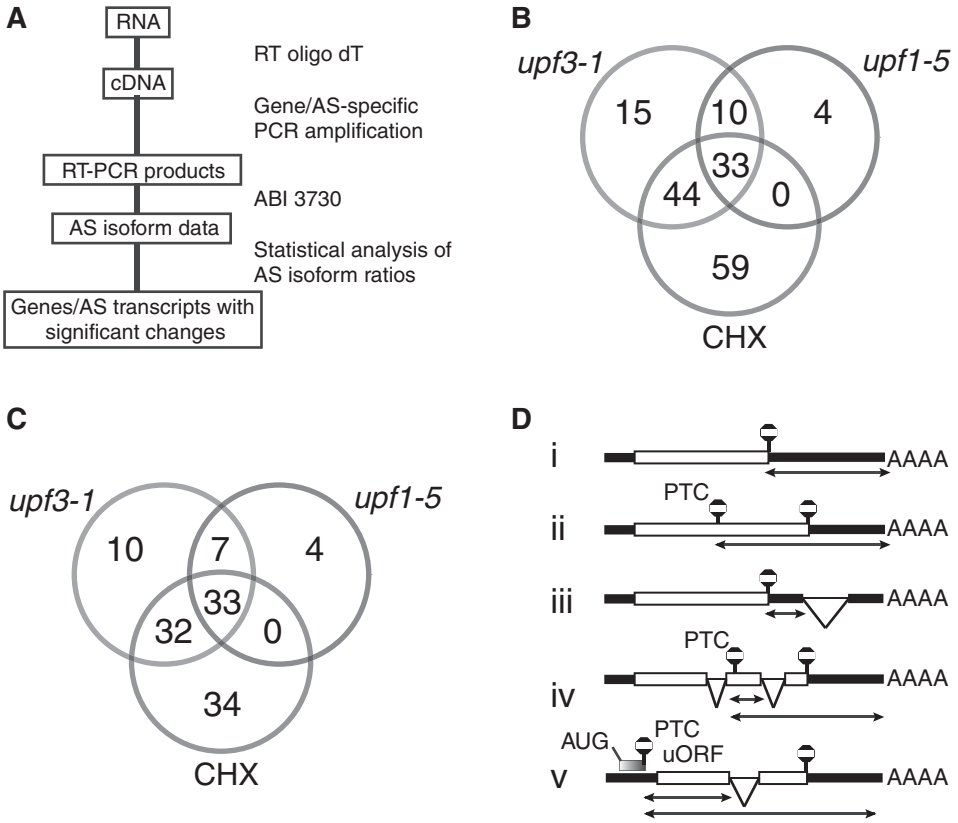


Figure 1. Analysis of alternatively spliced NMD substrates. (**A**) Schematic figure of RT–PCR panel analysis (see 'Materials and Methods' section). (**B**) Venn diagram of the number of transcripts which increase significantly in *upf* mutants and cycloheximide treatment. (**C**) Venn diagram of the number of genes with splice isoforms which increase significantly in *upf* mutants and cycloheximide treatment. (**D**) General features of transcripts which trigger NMD: (i) long 3′-UTR; (ii) PTC—long 3′-UTR; (iii) splice junction downstream of authentic stop codon (3′-UTR intron); (iv) PTC—downstream splice junctions and long 3′-UTR; (v) uORFs in 5′-UTR. Endogenous transcripts regulated by NMD contain long 3′-UTRs, introns in the 3′-UTR where the splice junction is >50–55 nt from the authentic stop or UORFs (i, iii and v, respectively). Transcripts which contain PTC in the coding region or 5′-UTR (uORF) also generate long 3′-UTRs with or without downstream splice junctions (ii, iv and v, respectively). Exons—open boxes; UTRs—black rectangles; thin lines—introns; diagonal lines—splicing events; stop sign—PTC or authentic termination codon.

Two previous studies performed genome-wide transcriptome profiling of NMD-defective plants using tiling and expression microarrays (31,50) and found that only about 1% of plant protein-coding genes were up-regulated in NMD-deficient plants. These arrays have limited ability to distinguish AS transcripts in contrast to the splicing-sensitive microarrays successfully used in animals (51–53), and it is therefore necessary to thoroughly investigate the fate and characteristics of endogenous plant AS transcripts turned over by NMD.

In the absence of splicing-sensitive microarrays for plants, we have used a high-resolution RT–PCR system (54) which is able to detect multiple AS transcript isoforms simultaneously and obtained isoform-level measurements from strong, but still viable mutant alleles (26,27) of the NMD protein genes, *UPF1* and *UPF3*, and cycloheximide (an inhibitor of translation and thus of NMD) treated plants. To address the link between AS and NMD we (i) investigated the effect of NMD impairment on a population (~950) of endogenous alternatively spliced transcripts, (ii) identified NMD-sensitive AS isoforms from significant changes in the ratio of AS isoforms, (iii) identified the characteristics of AS transcripts which trigger NMD, and (iv) identified transcripts which contain NMD features but which are insensitive to NMD. Our results demonstrate that alternative splicing and NMD affect a broad range of different genes in *Arabidopsis* and regulate expression of these genes via targeted degradation of specific AS transcripts using different mechanisms depending on the position of the AS event in the gene.

MATERIALS AND METHODS

Plant material, growth conditions, treatments and RNA isolation

Wild-type (ecotype Col-0) and *UPF* mutant *Arabidopsis* plants were used for the analysis. *UPF* mutants, *upf1-5* and *upf3-1*, (26) were a gift from Brendan Davies (Centre for Plant Sciences, University of Leeds, UK). Plants were grown *in vitro* on plates containing germination medium (55). Plants were maintained in 16-h light/8-h dark cycle at 22°C. Three week old plants were transferred into liquid half-strength Murashige and Skoog medium (56) and infiltrated with either 20 μM cycloheximide or the same volume of dimethylsulfoxide

as a control. Samples of cycloheximide-treated plants were collected after 5 h (27). RNA was isolated using RNeasy Plant Mini Kit (Qiagen).

High-resolution alternative splicing RT–PCR panel and data analysis

The original panel (54) was expanded to 289 primer pairs by identifying alternative splicing events which were either published, annotated in The *Arabidopsis* Information Resource (TAIR8—http://www.arabidopsis.org/) or in the Alternative Splicing in Plants database (ASIP—http://www.plantgdb.org/ASIP/). Primer pairs where one primer is fluorescently labelled were designed as described previously (54). Primer pairs used are listed in Supplementary Table S1. RT–PCR analysis was performed as described earlier (54). In brief, the reverse transcription reaction was carried out with total RNA using oligo-dT primers and the first-strand cDNA was aliquoted into microtitre plates, and PCR with the gene/alternative splicing event-specific primers performed using 24 cycles. We have previously shown that 24 cycles was still in the linear amplification range for various splicing substrates using [^{32}P]-labelling (57) and for a number of the AS primers used here to amplify transcripts of different abundance and size (54). The high-resolution RT–PCR system is capable of detecting multiple different AS transcripts from a gene, distinguishing alternative splicing events involving small size differences in transcripts (as few as 2–3 nt) and identifying small but significant changes in the ratios of alternatively spliced variants. The AS variants for each of the genes are amplified simultaneously by the same primers in the same reaction. The different AS isoforms usually have substantial common sequence which will reduce variation in amplification efficiency. In addition, if there are differences in amplification efficiency among particular AS isoforms, these differences will occur in the PCR reactions with wild-type, mutants and cycloheximide treatment. Electropherograms produced by the ABI 3730 genotyping software identified the exact size of the RT–PCR products for each primer pair. Peak areas for each RT–PCR product were extracted from the three reps, ratios of the different peaks were calculated generating a mean and standard error for each AS transcript as a percentage of the total transcript across the three reps.

Statistical analysis

The response to genotype, treatment and genotype by treatment interaction was assessed by analysis of variance (ANOVA). Each peak of each primer was analysed separately assuming a completely randomized design with three replicate values for each treatment combination. Response was measured as the percentage contribution of a particular isoform to the total transcripts measured and ANOVA was carried out after an angular transformation of the percentage values. In addition to assessing the significance of genotype and treatment main effects and their interaction, three specific comparisons (contrasts) were made: wild-type versus *upf1-5*, wild-type versus *upf3-1* and wild-type versus cycloheximide treatment. Residual plots were used to monitor the ANOVA assumptions of approximate normality and equality of variance. For the small number of cases where these assumptions did not hold either the response levels were all very low (or all very high) or the differences between treatments were so large as to render the ANOVA redundant.

In the analysis of the direct comparisons (above and Table 1) *P*-values were determined. In the subsequent analysis, we focussed on those transcripts which showed a significant percentage increase or decrease with a 3% difference between the means of wild-type plants and mutants/cycloheximide-treated plants. This level of difference was selected because we previously determined that when comparing variation in technical reps in the AS RT–PCR system, the majority of transcripts showed a standard error of the mean of <3% (54).

Sequencing analysis of AS RT–PCR products

Many RT–PCR products corresponded to unknown splice variants. To identify these products, RT–PCR reactions were purified using Agencourt AMPure beads (Beckman Coulter Genomics) and re-amplified prior to cloning into pGEM-T. Clones with differently sized inserts were identified by colony PCR and sequenced by standard procedures. Sequences were analysed either by using ClustalW or spliced alignments generated by GeneSeqer (http://www.plantgdb.org/tool/GeneSeqer/) (58). An in-house Perl script was used to parse the output from GeneSeqer for categorizing the annotations of alternative splicing events.

RESULTS

Analysis of alternative splicing coupled to nonsense-mediated decay using a high-resolution RT–PCR panel

To analyse the levels of alternatively spliced isoforms in mutants in the NMD protein genes, *UPF1* and *UPF3,* and to investigate the link between AS and NMD in *Arabidopsis thaliana*, we have exploited an alternative splicing RT–PCR panel. This unique high-resolution system is very sensitive and capable of detecting small but significant changes in alternative splicing at single nucleotide resolution (54) (Figure 1A). It detects different AS variants from the same gene/region simultaneously, variants containing more than one AS event and novel AS transcripts (Supplementary Figure S1). The AS RT–PCR panel used here consists of 289 primer pairs covering different alternative splicing events in 270 genes, and 3 control primer pairs (Supplementary Table S1). The AS events were selected from publications or from plant alternative splicing databases without prior knowledge of NMD-sensitivity with the exception of GRP7 and GRP8 (Figure 2). The AS events were mainly from genes encoding transcription factors, RNA-interacting proteins (including splicing factors) and stress-related proteins and included many important regulatory genes. The relative levels of AS isoforms were compared between wild-type, the two *upf* mutants and cycloheximide treatment (see Figures 2 and 3 and Table 1 for examples). The

Table 1. Selected AS transcripts which increase significantly in *upf* mutants and/or cycloheximide treatment

Primer pair	Gene ID	Name	Band size (bp)	AS event	NMD features	Mean ratio of transcripts			
						wt	*upf1-5*	*upf3-1*	CHX
7	At1g55310	At-SCL33	361	E4 Alt3′ss(+162)	PTC$^+$; ds SJ	0.15	0.19	**0.32**	**0.47**
7	At1g55310	At-SCL33	364	E4 Alt3′ss(+165) or AE (165)	PTC$^+$; ds SJ	0.02	0.03	**0.05**	**0.08**
7	At1g55310	At-SCL33	293	Unknown	Unknown	0.05	0.08	**0.09**	***0.07***
21	At2g37340	At-RS2Z33	341	E3 Alt3′ss(+218)	PTC$^+$; ds SJ	0.04	**0.07**	**0.13**	**0.20**
86	At4g16845	VRN2	396	IR2(170); E5 Alt3′ss(−4); E5 Alt5′ss(+22)	PTC$^+$; ds SJ	0.07	0.09	**0.12**	**0.13**
90	At4g39260	GRP8/CCR1	316	E1 Alt5′ss(+158)	PTC$^+$; ds SJ	0.03	0.05	**0.12**	**0.14**
109	At1g77080	MAF1	118	E4 Alt3′ss(−38)	PTC$^+$; ds SJ	0.14	**0.22**	**0.29**	0.19
118	At2g02960	Zinc finger (C3HC4) protein	233	E2 Alt3′ss(+37)	uORFs; uORF(26aa)	0.03	0.05	**0.07**	**0.07**
118	At2g02960	Zinc finger (C3HC4) protein	237	E2 Alt3′ss(+41)	uORFs; uORF(26aa)	0.04	0.05	**0.07**	**0.07**
118	At2g02960	Zinc finger (C3HC4) protein	240	E2 Alt3′ss(+44)	uORFs; uORF(26aa)	0.02	0.02	***0.03***	0.02
118	At2g02960	Zinc finger (C3HC4) protein	289	E2 Alt3′ss(+92)	uORFs; uORF(26aa)	0.25	0.30	0.33	**0.51**
125	At2g46790	APRR9/TL1	251	E2 Alt5′ss(+8)	PTC$^+$; ds SJ	0.22	**0.32**	**0.47**	**0.37**
131	At2g38880	NF-YB1/HAP3a	373	E6 Alt5′ss(+62)	3′UTR intron	0.21	0.26	**0.30**	**0.41**
194	At3g49430	At-SR34a	366	I1 AE(224)	uORFs; uORF(61aa)	0.07	**0.21**	**0.35**	**0.30**
195	At3g01150	At-PTB2a	156	E8 Alt5′ss(−47)	PTC$^+$; no ds SJ - possible long 3′UTR	0.10	0.12	**0.16**	**0.19**
196	At3g01150	At-PTB2a	268	I2 AE(102)	PTC$^+$; ds SJ	0.11	0.17	**0.25**	**0.37**
202	At3g13570	At-SCL30a	351	I3 AE(161)	PTC$^+$; ds SJ	0.18	0.30	**0.48**	**0.52**
204	At3g53500	At-RS2Z32	376	E3 Alt3′ss(+218)	PTC$^+$; ds SJ	0.09	0.11	**0.25**	**0.40**
205	At3g61860	At-RS31	556	I2 AE(393)	PTC$^+$; ds SJ	0.01	0.04	**0.10**	**0.27**
206	At2g21660	GRP7/CCR2	349	E1Alt5′′ss(+166)	PTC$^+$; long 3′UTR	0.02	**0.07**	**0.16**	**0.13**
213	At5g53180	At-PTB2b	198	I3 AE(58)	PTC$^+$; ds SJ	0.08	0.12	**0.19**	**0.31**
213	At5g53180	At-PTB2b	201	I3 AE(61)	PTC$^+$; ds SJ	0.06	0.09	**0.14**	**0.22**
217	At1g16610	SR45	175	FS	no reason for NMD	0.68	0.71	**0.72**	**0.77**
218	At2g30260	U2B′′	134	E2 Alt5′′ss(−35)	PTC$^+$; ds SJ	0.05	0.09	**0.19**	**0.17**
219	At4g25500	At-RS40	382	I2 AE(257)	PTC$^+$; ds SJ	0.14	**0.33**	**0.54**	**0.34**
220	At3g55460	At-SCL30	590	I3 AE(449)	PTC$^+$; ds SJ	0.01	**0.05**	**0.11**	**0.08**
220	At3g55460	At-SCL30	673	Unknown	Unknown	0.01	0.02	**0.05**	**0.13**
223	At2g29210	Splicing factor PWI domain-containing protein	202	E6 Alt3′ss(+50)	PTC$^+$; ds SJ	0.36	0.36	**0.59**	**0.63**
237	At1g07830	RPL29 family	123	E2 Alt3′ss(−70)	uORFs; uORF(31aa)	0.01	***0.02***	***0.03***	0.02
237	At1g07830	RPL29 family	273	E1 Alt5′ss, predicted	uORFs	0.02	**0.05**	**0.09**	**0.07**
241	At1g02090	FUS5/CSN7/COP15	119	E8 Alt3′ss(−17)	PTC$^+$; ds SJ	0.15	**0.27**	**0.29**	**0.30**
249	At1g72560	PSD/Exportin-t	174	E13 Alt5′ss(+25)	3′UTR intron	0.04	**0.15**	0.04	0.05
259	At3g17090	PP2C group D	214	E2 Alt3′ss(+17)	PTC$^+$; ds SJ	0.07	0.11	**0.23**	0.14
284	At4g33060	CYP57	290	I5 AE(87)	PTC$^+$; ds SJ	0.26	0.34	**0.49**	**0.71**
284	At4g33060	CYP57	295	I5 AE(92)	PTC$^+$; ds SJ	0.05	0.06	**0.08**	**0.11**
306	At2g46830	CCA1	218	FS	No reason for NMD from AS event—other AS/NMD events known	0.79	0.85	**0.87**	**0.89**
309	At5g65060	MAF3	223	E4 Alt3′ss(−38)	PTC$^+$; ds SJ	0.19	**0.35**	**0.42**	**0.38**
309	At5g65060	MAF3	219	E4 Alt3′ss(−38); E5 Alt3′ss(−4)	PTC$^+$; ds SJ	0.04	***0.06***	**0.08**	**0.07**
311	At5g65080	MAF5/AGL68/FCL1	462	E5 Alt3′ss(−4)	PTC$^+$; ds SJ	0.76	**0.82**	0.80	0.75
343	At3g29160	AKIN11	159	FS	No reason for NMD from AS event—other AS/NMD events known	0.27	0.27	**0.34**	**0.40**
344	At3g29160	AKIN11	195	E9 Alt3′ss(−5)	PTC$^+$; ds SJ	0.08	**0.12**	**0.17**	**0.15**
363	At5g24270	SOS3/CBL4	337	E5 Alt5′ss(+29)	PTC$^+$; ds SJ	0.06	0.08	**0.13**	**0.25**
370	At5g35410	SOS2/CIPK24	120	E9 Alt3′ss(+5)	PTC$^+$; ds SJ	0.42	**0.49**	**0.59**	**0.63**
374	At4g36960	RRM-containing protein	399	IR1(172)	uORFs; uORF(73aa)	0.05	**0.15**	**0.32**	0.09
375	At3g20270	lipid-binding serum glycoprotein family protein	195	FS	uORFs; uORF(22aa)	0.54	**0.62**	**0.64**	**0.59**
384	At4g02200	At-Di19-5	149	E2 Alt5′ss(+32)	PTC$^+$; ds SJ	0.06	**0.13**	**0.21**	**0.22**
384	At4g02200	At-Di19-5	110	Unknown	Unknown	0.01	0.01	0.02	**0.04**

Values are mean values from three biological reps and significance is $P < 0.1$ (see text). **En Alt5′ss** and **En Alt3′ss**—alternative 5′ splice site and alternative 3′ splice site in exon, where n is the exon number. The number in brackets indicates the number of nucleotides added (+) or removed (−) from the exonic sequence. **IRn**—retention of intron, where n is intron number [number in brackets indicates the size (nt) of the intron]. **In AE**—alternative (cryptic) exon created in an intron, where n is the intron number [number in brackets indicates the size(s) of the alternative exon(s)]. Exons and introns are numbered with respect to the TAIR reference splice variant. **FS**—fully spliced; **PTC$^+$**—transcript containing premature termination codon(s); **ds SJ**—presence of downstream splice junction(s). Numbers in bold—significant percentage difference >3%; numbers in bold italic—significant percentage difference <3%. Full data set is presented in the Supplementary Table S2. Note that *Arabidopsis* SR proteins are named according to the recently proposed nomenclature (77).

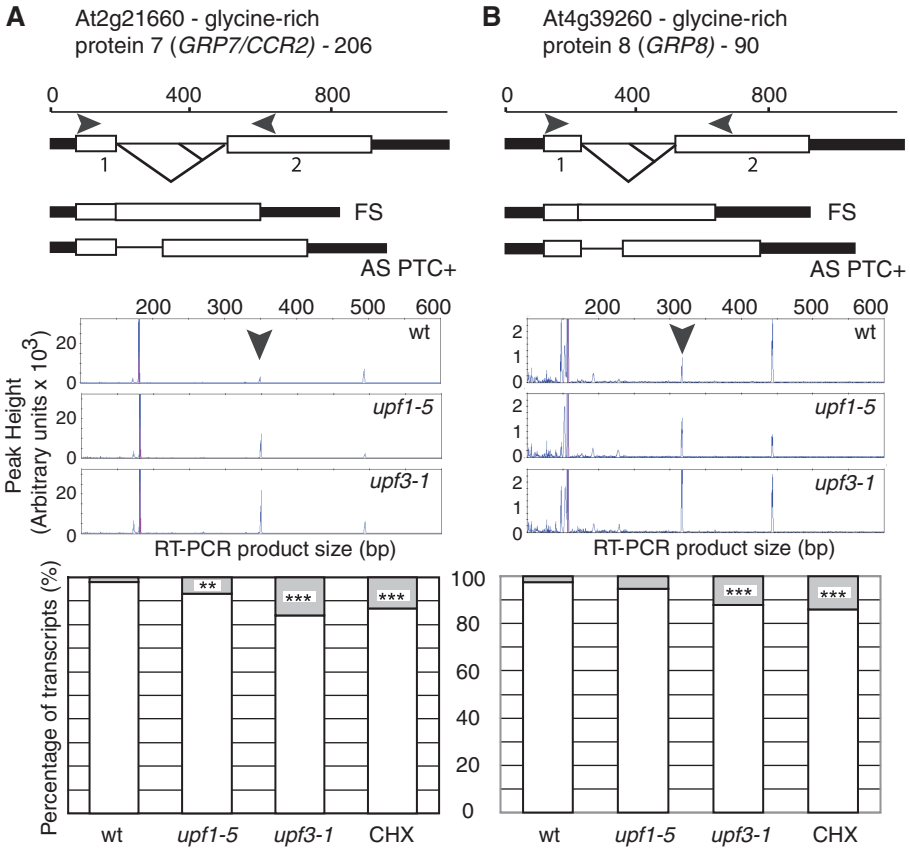


Figure 2. Regulation of *GRP7* and *GRP8* by alternative splicing and NMD. (**A**) *GRP7* and (**B**) *GRP8* are known to be regulated by AS/NMD. Figures show the *GRP7* and *GRP8* gene and transcript structures and the alternative splicing events in the introns: alternative 5′ splice sites generate AS isoforms which increase in abundance in the *upf1-5* and *upf3-1* mutants as illustrated on scans generated from the ABI 3730 data by GeneMapper (transcripts are arrowed). The significant increases in NMD-sensitive transcript abundance are shown in histograms of the ratio of normally spliced and alternatively spliced isoforms (shaded). Significance: ***$P < 0.01$; **$0.01 > P < 0.05$. For diagram key see legend to Figure 1.

upf1-5 and *upf3-1* mutants are impaired in NMD and have severe growth phenotypes but are viable (26,27). Translation is required for NMD (22,59) and the translation inhibitor, cycloheximide, leads to accumulation of NMD-sensitive transcripts in plants (27). Therefore a 5 hour cycloheximide treatment was used to inhibit translation and thereby NMD (27). Consequently, AS isoforms which are targets of NMD are expected to increase in their levels in the *upf* mutants and cycloheximide-treated plants. Three biological replicates were analysed for each, and significant changes in alternative splicing ratios were determined by statistical analysis (see 'Materials and Methods' section).

High frequency of novel alternatively spliced transcripts in regulatory plant genes

Based on publications or plant databases the majority of AS events were expected to generate two alternative transcripts. However, the number of observed RT–PCR products varied among the different amplified regions from a single product to as many as 15 different alternatively spliced products. Just over 950 RT–PCR products were observed using the 289 primer pairs and therefore approximately 350 new transcripts were discovered. To identify the nature of the novel AS transcripts, cloning and sequencing of RT–PCR products was carried out (for examples of analysis see Supplementary Figure S1). In addition, many were identified by RNA-Seq (our unpublished data). In general, our data shows an increase of AS frequency in our gene set by one third compared to presently annotated events. The identification of so many novel products illustrates that far more alternative splicing occurs in *Arabidopsis* than is currently known.

Identification of endogenous alternatively spliced targets of nonsense-mediated decay

Our high-resolution RT–PCR system allowed the determination of the ratio of AS transcript variants for each gene region amplified. Mean ratios of AS products obtained with the *upf* mutants and cycloheximide treatment were compared to the wild-type values to identify the particular AS isoforms which increased significantly when NMD is impaired. Significance was determined at the $P < 0.1$ level, although for the vast majority significant changes in AS isoform levels, the *P*-value was considerably smaller (Supplementary Table S2). Of the >950 transcripts in the study, 638 showed no significant change in the transcript ratios between wild-type, mutants or cycloheximide-treated plants. Of the 313 RT–PCR

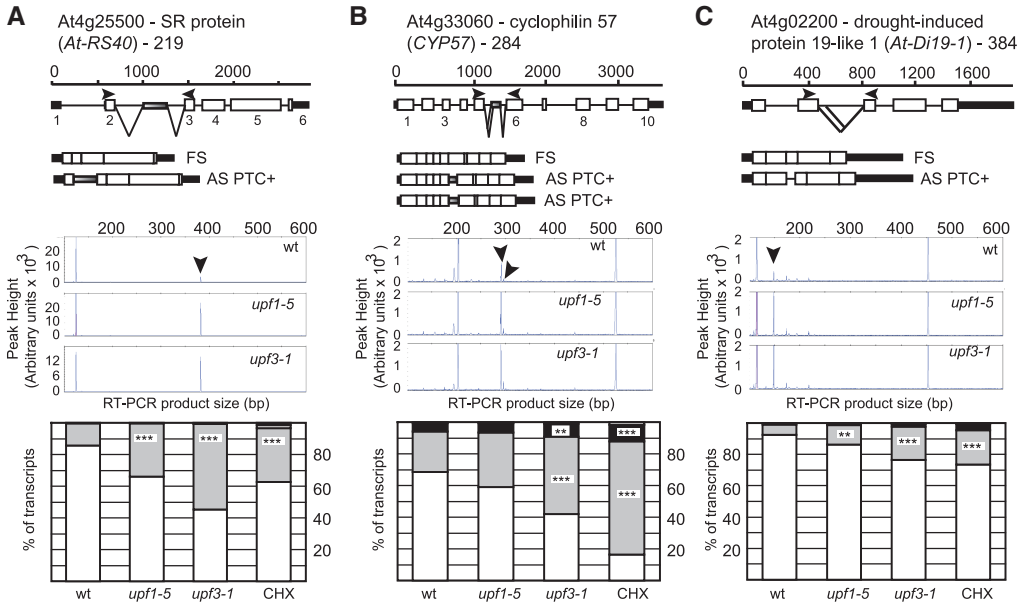


Figure 3. Genes with AS isoforms which increase in *upf* mutants. (**A**) At4g25500—SR protein gene, At-*RS40*. (**B**) At4g33060—cyclophilin 57, *CYP57*. (**C**) At4g02200—drought-induced protein gene, At-*Di19-1*. For all, the gene and transcript structures and relevant splicing events are shown. AS isoforms which increase in the *upf* mutants are labelled with arrows on ABI3730 scans and the ratios of transcripts are shown in histograms and significant increases are indicated. Significance: $***P < 0.01$; $**0.01 > P < 0.05$. For diagram key see legend to Figure 1.

products that showed a significant change, 165 increased in amount in at least one of the *upf* mutants or cycloheximide treatments (Figure 1B; Supplementary Table S2). Thirty-three transcripts increased in both mutants and cycloheximide-treated plants (Figure 1B). A total of 106 transcripts were increased in one or other mutant while 59 showed a significant increase only in the cycloheximide-treated plants. Cycloheximide is used widely as an NMD inhibitor but as cycloheximide is a general translational inhibitor, other RNA degradation pathways or cellular processes might also be affected by this treatment and impact on transcript levels. Therefore, the 106 transcripts which increased in the *upf* mutants and the 165 which increased in mutants and cycloheximide treatment represent a range of naturally occurring alternatively spliced transcripts which are putatively turned over by NMD and make up 11–17% of the total transcripts analysed and 16–25% of the alternatively spliced transcripts analysed.

We found that 87 and 121 genes of the 270 AS genes on the panel (Figure1C; Supplementary Table S3) had at least one AS isoform with increased abundance in the *upf* mutants or in the mutants plus CHX treatment, respectively, suggesting that ~32% and 45% of AS genes are regulated by NMD to some extent. At least 42% (9273 genes out of 22 302) of intron-containing genes in *Arabidopsis* are alternatively spliced (36). With the caveat that our gene set may contain some bias, we can extrapolate to suggest that around 13–18% of intron-containing genes may be regulated by AS and NMD in *Arabidopsis*.

GRP7 and *GRP8* are genes encoding components of a slave oscillator and are known to be auto- and cross-regulated by alternative splicing and NMD (44,45) and were included as controls. For both genes, the AS isoform which is turned over by NMD increased significantly as expected. The *GRP7* isoform increased significantly in both mutants and cycloheximide treatment (Figure 2A; Table 1, primer pair 206) and significant increases in the *GRP8* isoform were observed in *upf3-1* and cycloheximide treatment (Figure 2B; Table 1, primer pair 90). These results demonstrate that the AS RT–PCR panel is able to detect significant changes in AS isoforms due to NMD. Other examples of AS/NMD transcripts are shown in Figure 3. At-*RS40,* an SR protein gene, *CYP57*, a peptidyl-prolyl *cis-trans* isomerase gene, and the drought-induced protein 19-like 1 gene, At-*Di19-5,* have alternative splicing events which introduce PTCs and increase significantly in mutants and cycloheximide treated plants. In general, the increase in levels of the NMD-sensitive AS isoforms in the mutants and after cycloheximide treatment varied with different genes. Interestingly, the steady state levels of the AS transcripts turned over by NMD varied greatly in wild-type plants from being virtually undetectable to tens of percent of the transcripts from a gene (Figure 3B and C; Table 1 and Supplementary Table S2). Thus, AS/NMD transcripts from different genes are differentially abundant in wild-type plants which reflects the different efficiency of alternative splicing and rates of turnover by NMD. This analysis shows that coupled AS/NMD can influence gene expression significantly and that even low abundant alternative transcripts (at steady state level) might turn over a quite significant proportion of the RNA produced from a gene.

upf1-5, *upf3-1* and cycloheximide treatment impair nonsense-mediated decay to different degrees

More transcripts (102 transcripts) increased significantly in the *upf3-1* mutant than in *upf1-5* (47 transcripts) while 136 transcripts increased significantly in the

cycloheximide-treated plants (Figure 1B). This suggests that in terms of NMD impairment, the *upf3-1* allele is stronger than the *upf1-5* allele which conforms to the severity of the phenotype of the particular mutants which was described previously (26,27,31) and that *UPF3* transcripts are up-regulated in *upf1-5* (60). In addition to the described late flowering phenotype, we have observed that both *upf1-5* and *upf3-1* mutants have accelerated senescence and again *upf3-1* showed the stronger phenotype (Supplementary Figure S2). Some transcripts only increased with cycloheximide treatment and were not visible in the wild-type or mutant plants. Thus, cycloheximide has a much stronger effect on the number of transcripts which increased in abundance and on the degree of increase than *upf3-1* with the smallest increases being seen in *upf1-5* (Supplementary Tables S4 and S5). Some transcripts did not follow this pattern suggesting perhaps differential or additional functions of the two different NMD factors in different mechanisms of NMD (28). There was substantial overlap between transcripts which increased in abundance in cycloheximide treatment and in the mutants (Figure 1B). In addition, we analysed 11 transcripts which increased only in the cycloheximide treatment in detail and the majority showed NMD characteristics (PTCs/downstream splice junctions or uORFs) (Supplementary Table S2). This suggests that AS transcripts which increase in the cycloheximide treatment are turned over by NMD with the caveat that some transcripts may be turned over by other degradation pathways.

Features of alternatively spliced transcripts which are sensitive to nonsense-mediated decay

Previously, 1% of plant protein-coding genes was shown to have increased transcript levels in NMD-deficient plants using tiling arrays but the features of the AS transcripts which were NMD targets were not specifically investigated (50). The 165 endogenous AS isoforms which show significant increases in levels in the mutants and cycloheximide-treated plants (Table 1; Supplementary Table S2) are expected to be turned over by NMD and therefore should contain NMD signals. To identify the NMD features, each transcript which increased significantly in at least one of the mutants and 11 of the transcripts which increased only in cycloheximide-treated plants (117 transcripts in total) was characterized in terms of whether they contained PTCs, had splice junctions downstream of the authentic stop codon or PTCs, had long 3′-UTR sequences or contained an upstream ORF (Figure 1D). In addition, other AS events in the same gene (annotated in TAIR8) and novel RT–PCR products for which sequence was generated were analysed. Of the 117 transcripts, the sequences of 8 remain elusive and could not be characterized. Of the remaining 109 transcripts, 94 (86%) clearly contained NMD features. This includes a major group of 74 transcripts containing PTCs more than 50–55 nt upstream of splice junctions and long 3′UTRs, classical features of NMD substrates (61). The NMD-sensitivity of a further nine transcripts could be explained on the basis of long 3′-UTRs and/or where the distance between the authentic stop codon and splice junction of an intron in the 3′-UTR was changed due to alternative splicing. In addition, the alternative splicing events of 12 genes (14 transcripts) involved introns in the 5′-UTR which affected the presence or absence, length and position of uORFs. In 11 of these transcripts, the presence of one or more uORFs correlated with NMD. Finally, for the remaining 12 transcripts, the AS event monitored did not explain the turnover by NMD but of these, four genes had known AS events elsewhere in the transcripts which could generate NMD. The remaining transcripts could either be genes/transcripts which have unknown NMD-inducing AS events elsewhere in the gene or may represent genes where changes in AS isoform levels are due to secondary effects on mRNA accumulation. Therefore, the vast majority of the endogenous NMD-sensitive transcripts analysed had characteristic features of NMD substrates in plants.

Alternative splicing in 3′-UTRs modulates nonsense-mediated decay

It is not widely appreciated that alternative splicing in the untranslated regions of a gene (which does not create a PTC) may induce NMD sensitivity and be a mechanism of fine regulation of transcript abundance. Also, little is known about the consequences of AS of such UTR introns. We have identified two genes in our panel where AS occurred in 3′-UTR introns and thus did not create a PTC, but at least one AS isoform in each gene was sensitive to NMD. Alternative splicing in the 3′-UTR of At2g38880 (*NF-YB1*/*HAP3A* transcription factor) and At1g72560 (*PAUSED*/Exportin-t) generated two and three isoforms, respectively, with different sensitivity to NMD (Figure 4). Only the isoforms with a distance >50–55 nt from the authentic stop codon were subjected to NMD (Table 1). Thus, alternative splicing can affect the distance between the authentic stop codon and downstream splice junction in 3′-UTRs and determine whether a transcript is turned over by NMD.

The position of PTCs defines the length of 3′-UTRs which can trigger nonsense-mediated decay

Current models for NMD suggest that the length of the 3′-UTR (distance between the stop codon or PTC and 3′-end of the transcript—long 3′-UTR) can be one of the triggers of NMD (Figure 1c). For all of the analysed AS transcripts containing PTCs, we analysed the distance between the first PTC and the 3′-end of the transcript (Figure 5; Supplementary Table S2). These distances appeared to show a bimodal distribution with the majority (58 transcripts) being longer than 600 nt and twenty transcripts (from 16 genes) shorter than 550 nt. When these latter transcripts were examined, two-thirds were from genes with relatively short coding sequences (500–720 bp) and all but five had splice junctions downstream of the PTC such that they conformed to expected features of NMD substrates. The five transcripts without a downstream splice junction had PTC to 3′-end distances of 366, 370, 440 nt and two transcripts had 441 nt and

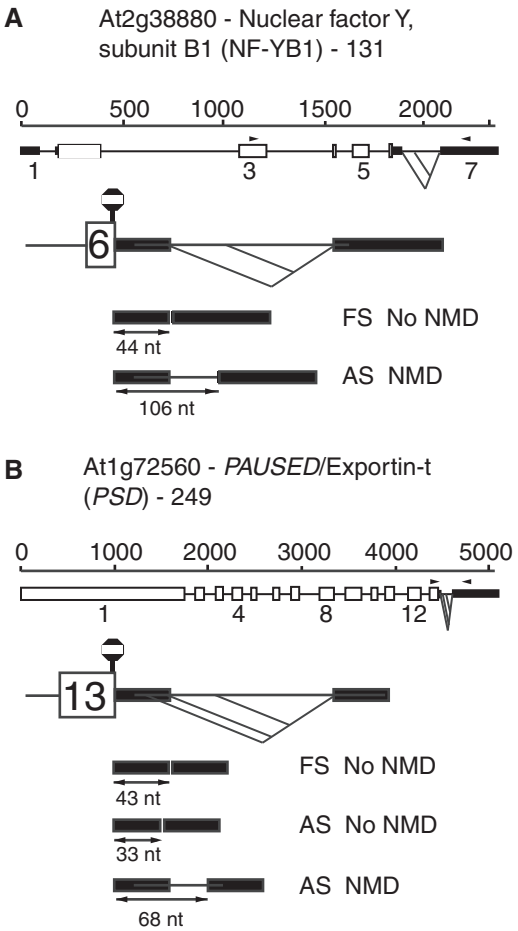


Figure 4. Alternative splicing of introns in the 3′-UTR influences turnover of AS isoforms by NMD. Exon–intron structures of genes and transcripts of (**A**) At2g38880—NF-YB1 transcription factor, *NF-YB1*/*HAP3a* and (**B**) At1g72560—*PSD*/*exportin-t*. Alternative splicing of the 3′-UTR introns in these genes generate transcripts with different distances between the authentic stop codon and downstream splice junction consistent with the 50–55 nt rule where distances >50–55 nt trigger NMD. Stop codon to splice junction distances are indicated along with whether the AS isoform is turned over by NMD or not. For diagram key see legend to Figure 1.

assuming that there are no other AS events in the genes causing NMD, these could represent 'long 3′–UTR' transcripts. We also determined the distance between the authentic stop codon and 3′-end of the gene in the normally spliced transcripts from the same genes. The mean distance was 242 nt and the majority of transcripts were in the range of 22–350 nt (Figure 5). Eight fully spliced transcripts (from 7 genes) had a 3′-UTR of >350 nt ranging from 354 to 718 nt. Taken together, these data suggest that long 3′-UTRs which trigger NMD in *Arabidopsis* mRNAs are in general >350 nt but there are exceptions where transcripts with longer 3′-UTRs do not show evidence of NMD.

uORFs overlapping the main start codon induce nonsense-mediated decay

Previous studies in *Arabidopsis* suggest that uORF can trigger NMD (35,60). Preliminary features of such

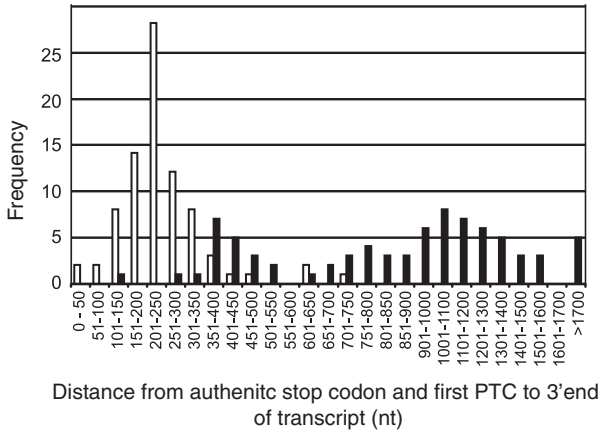


Figure 5. Distribution of stop codon to 3′-end distances. Distribution of frequency of distances in nucleotides (nt) between the first premature termination codon and the 3′-end of the transcript (shaded) compared to the distance between the authentic stop codon and 3′-end of the transcript of cognate genes (unshaded).

uORFs were defined using model constructs as being the first uORF in a transcript, at least 10 nt from the 5′-end and longer than 35–50 amino acids (35). However, the uORF in *AtMHX* was only 13 amino acids long and affected mRNA levels and translational efficiency (60). In addition, the link between AS in 5′UTRs and the presence, size, and positions of uORFs which may then trigger NMD has never been investigated in plants or in endogenous transcripts. In our study, we identified 12 genes with alternative splicing of 5′-UTR introns where transcripts increased significantly in *upf* mutants and/or cycloheximide treated plants. uORFs of between 3 and 123 amino acids were present in the fifteen different AS isoforms of these genes.

Seven genes had uORFs with the interesting unifying feature that the AS isoforms turned over by NMD contained an uORF which overlapped the translation start site of the main ORFs (Figure 6A–C; Supplementary Figure S3; Table 1). For example, in the zinc finger protein gene (At2g02960), alternative splicing produces six different AS transcripts through use of multiple alternative 3′ splice sites (Figure 6A). Four of these produce a 26 amino acid uORF which overlaps the AUG translation start site of the main ORF and all four are NMD sensitive. The fully spliced transcript and two shorter alternatively spliced isoforms contained short uORFs which do not overlap with the AUG. AS in the 5′UTR of At3g49430 (At-*SR34a*) generated three new uORFs of 13, 30 and 61 amino acids (Figure 6B). The stop codon of the 61 amino acid uORF overlapped the AUG of the main ORF and this transcript is NMD sensitive. Similarly, the fully spliced product of At3g20270 contains a uORF of 22 amino acids where the stop codon of the uORF lies downstream of the AUG of the main ORF (Figure 6C) and is NMD sensitive whereas the other AS transcripts without this feature are NMD resistant. Other examples of genes showing this phenomenon are shown in Supplementary Figure S3. Taken together, our analysis shows that many transcripts containing a uORF which

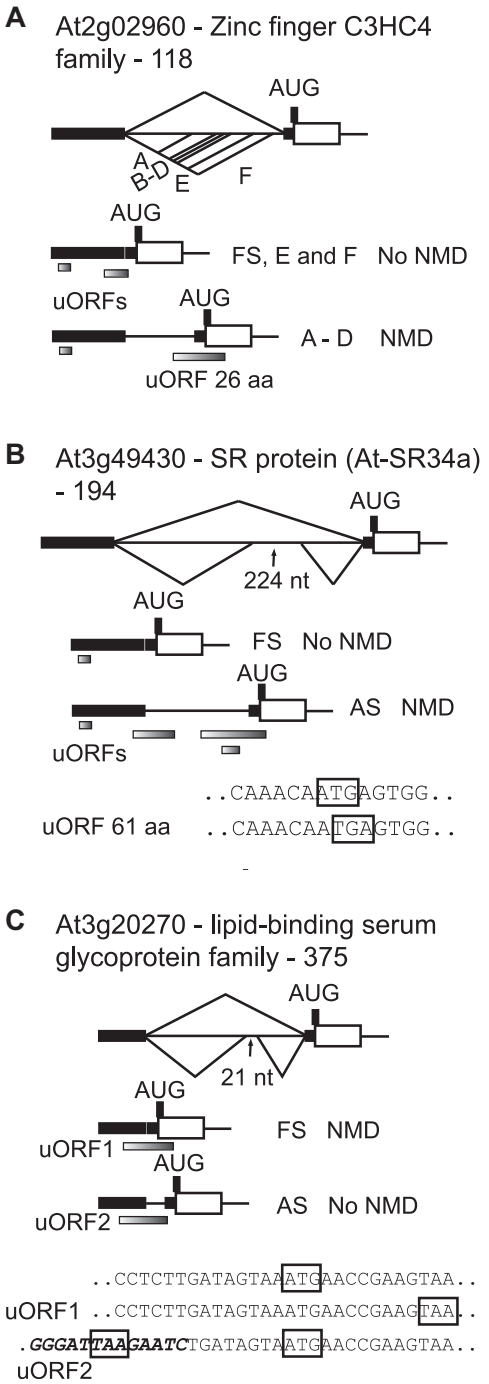


Figure 6. Alternative splicing of introns in the 5′-UTR affects the presence, size and position of uORFs and influences turnover of AS isoforms by NMD. Exon–intron structures of genes and transcripts of (**A**) At2g02960—zinc finger transcription factor, (**B**) At3g49430—SR protein gene, At-*SR34a*, and **(C)** At3g20270—lipid-binding serum glycoprotein gene. These examples illustrate alternative splicing events in 5′-UTR introns which generate uORFs which overlap the main ORF and correlate with NMD. (**A**) four of the AS isoforms contain a 26 amino acid uORF (A–D) which overlaps the translation start site of the main ORF and correlates to NMD; (**B**) an uORF of 61 amino acids overlaps the authentic translation start codon in the AS product; and (C) uORF1 overlaps the translation start of the main coding sequence in the fully spliced transcript while in the alternatively spliced isoform the stop codon of uORF2 lies upstream of the main translation start site. Shaded rectangles below transcripts—uORFs; FS—fully spliced; AS—alternatively spliced. Sequences below the figures show the relationship between the stop codon of uORFs and the translational start AUG of the main ORF. For diagram key see legend to Figure 1.

overlaps the authentic start codon are subject to NMD and identifies a new feature of uORFs which is capable of inducing NMD in plants. The analysis also indicates that not all features of uORFs which trigger NMD have been resolved as we find examples where there is no correlation of the presence of uORFs of between 43 and 92 amino acids and NMD (see 'Discussion' section).

Splice isoforms with retained introns are not sensitive to nonsense-mediated decay

Intron retention is the most abundant type of alternative splicing in plants—41% (62). In general, due to the UA-richness of plant introns, most intron retention events create PTC transcripts. These transcripts are considered as potential targets of NMD as it has been shown in other organisms that transcripts with retained introns are turned over by NMD (63). Of the 90 characterized AS transcripts which increased in one or both of the *upf* mutants, only four had retained introns (Supplementary Table S2) suggesting that IR transcripts were underrepresented. We therefore examined all of the readily detectable intron retention events on our panel where the IR transcripts made up at least 2% of the total. Of the 29 such IR transcripts, nineteen contained PTCs with downstream splice junctions and/or had long 3′-UTRs (all >400 nt) and therefore have features of typical NMD substrates (Table 2; Figures 1D and 7). Despite containing these NMD signals, these intron retention isoforms did not increase in abundance in the *upf* mutants and/or cycloheximide treatment suggesting that they are not turned over by the NMD pathway. Interestingly, in a number of these genes, other alternatively spliced transcripts were produced which contained PTCs and were subject to NMD. In two cases in particular, the other AS events involved the same intron as the retained intron (Figure 7A and B). In At5g37055 (*SEF—SERRATED LEAVES AND EARLY FLOWERING*), there are three different intron retention transcripts involving introns 1 and 2, all containing PTCs, none of which is a target of NMD (Figure 7C). However, use of an alternative 3′ splice site in exon 3 generates a PTC+ transcript which is turned over by NMD (Figure 7C and Table 2). A further example is At5g24270, coding for *SOS3—SALT OVERLY SENSITIVE 3*, a calcineurin-like protein, where retention of intron 5 was NMD resistant while use of an alternative 5′ splice site in the next intron, also creating a PTC, significantly up regulated the latter transcript in *upf3-1* and cycloheximide-treated plants (Figure 7D and Table 2). The first PTC generated in these two transcripts are only 30 nt apart (Figure 7D) arguing against a position effect of the PTC. Thus, transcripts from the same gene which generate PTCs in very similar positions either through alternative splicing or intron retention events can be differentially sensitive to NMD. It appears that if a transcript is generated where an intron has not been spliced (retained intron) then the transcript is not NMD sensitive.

Besides the above PTC+ NMD-insensitive events, six other IR transcripts were effectively PTC- and were not targets of NMD (Supplementary Table S6). In these

Table 2. PTC$^+$, NMD-insensitive intron retention transcripts and NMD-sensitive AS transcripts from the same genes

Primer pair	GeneID	Name	Exon no.	Primer sites	AS event	Transcript	wt	upf1-5	upf3-1	NMD
50	At5g43910	pfkB-type carbohydrate kinase family protein	12	Ex8-10	FS		0.58	0.46	0.39	
					IR9	PTC$^+$; ds SJ	0.05	0.05	0.04	No NMD
					E10 Alt 3′ss(−19)	PTC$^+$; ds SJ	0.35	0.48	0.56	NMD
346	At4g23260	CRK18	7	Ex1–3	FS		0.09	0.09	0.17	
					IR1	PTC$^+$; ds SJ	0.90	0.90	0.82	No NMD
					unknown (245 bp)		0.00	0.00	***0.01***	NMD
316	At2g28550	TOE1	8	Ex1–4	FS		0.93	0.94	**0.96**	
					IR2	PTC$^+$; ds SJ	0.07	0.06	0.04	No NMD
312	At5g13790	AGL15	8	Ex1–7	FS		0.45	0.47	0.47	
					IR2 and IR3	PTC$^+$; ds SJ	0.06	0.06	0.07	No NMD
363	At5g24270	SOS3/CBL4	8	Ex4–7	FS		0.83	0.82	0.77	
					E5 Alt5′SS(+29)	PTC$^+$; ds SJ	0.06	0.08	**0.13**	NMD
					IR4	PTC$^+$; ds SJ	0.03	0.03	0.02	No NMD
					unknown (299 bp)		0.02	0.02	***0.03***	NMD
					unknown (330 bp)		0.00	0.01	***0.01***	NMD
313	At5g37055	SEF	4	Ex1–4	FS		0.79	0.78	0.77	
					IR1 and IR2	PTC$^+$; ds SJ	0.03	0.03	0.03	No NMD
					IR1	PTC$^+$; ds SJ	0.13	0.13	0.12	No NMD
					IR2	PTC$^+$; ds SJ	0.01	0.01	0.01	No NMD
					E3 Alt 3′ss(−11)	PTC$^+$; ds SJ	0.02	***0.04***	**0.05**	NMD
372	At1g76460	RRM-containing protein	7	Ex1–3	FS		0.88	0.84	0.79	
					IR1	PTC$^+$; ds SJ	0.03	0.03	0.02	No NMD
					E2 Alt 3′ss(+9)	PTC$^+$; ds SJ	0.04	***0.05***	**0.07**	NMD
					E2 Alt 3′ss(+54)	PTC$^+$; ds SJ	0.05	**0.08**	**0.12**	NMD
345	At1g49730	protein kinase	9	Ex4–7	FS		0.89	0.81	0.77	
					IR6	PTC$^+$; ds SJ	0.02	0.03	0.03	No NMD
					E7 Alt 3′ss(+37)	PTC$^+$; ds SJ	0.08	**0.16**	**0.21**	NMD
321	At4g27410	RD26	3	Ex1–3	FS		0.93	0.92	0.93	
					IR1	PTC$^+$; ds SJ	0.02	0.02	0.02	No NMD
325	At2g47890	Zn finger (B-box type) protein	4	Ex1–4	FS		0.92	0.93	0.88	
					IR3	C-terminal change	0.06	0.06	0.10	No NMD
					IR2	PTC$^+$; ds SJ	0.01	0.01	0.01	No NMD
373	At3g13224	RRM-containing protein	6	Ex4–5/6	FS		0.90	0.89	0.90	
					IR5	PTC$^+$; long 3′ UTR (1395 nt)	0.10	0.11	0.10	No NMD
335	At5g66210	CPK28	13	Ex11–13	FS (I12 in 3′ UTR)		0.01	0.01	0.01	
					IR11 and IR12	PTC$^+$; long 3′UTR (686 nt)	0.90	0.91	0.90	No NMD
					ES12	C-terminal change	0.02	0.02	0.01	No NMD
326	At1g69250	NTF2	8	Ex6–8	FS		0.92	0.92	**0.94** **	
					IR6	PTC$^+$; long 3′UTR (415 nt)	0.08	0.08	0.06	No NMD
333	At3g16800	PP2C	6	Ex4–6	FS		0.97	0.97	0.97	
					IR5	PTC$^+$; long 3′UTR (469 nt)	0.03	0.03	0.03	No NMD
317	At2g28550	TOE1	8	Ex6–8	FS		0.97	0.98	0.98	
					IR7	PTC$^+$; long 3′UTR (493 nt)	0.02	0.02	0.01	No NMD
366	At5g25610	RD22	4	Ex1–3	FS		0.97	0.95	0.97	
					IR2	PTC$^+$; dsSJ	0.03	0.05	0.02	No NMD
374	At4g36960	RRM-containing protein	13	Ex1–3	FS		0.84	0.74	0.61	
					IR1 (5′ UTR)	PTC$^+$; uORF in 5′ UTR	0.05	**0.15**	**0.32**	NMD
					IR2	PTC$^+$; ds SJ	0.11	0.10	0.07	No NMD

Primers are positioned in the exons indicated and amplify across at least two introns. **En Alt5′ss** and **En Alt3′ss**—alternative 5′ splice site and alternative 3′ splice site in respective exon, where n is an exon number. The number in brackets indicates the number of nucleotides added (+) or removed (−) from the exonic sequence. **IRn**—retention of intron, where n is an intron number. The number in brackets indicates the size (nt) of the intron. Exons and introns are numbered with respect to the TAIR reference splice variant. **FS**—fully spliced; **PTC+**—transcript containing premature termination codon(s); **ds SJ**—presence of downstream splice junction(s); Numbers in bold—significant difference >3%; numbers in bold italic—significant difference <3%.

transcripts, the introns were either (i) in frame (no PTC), (ii) towards the end of the transcript such that the PTC was close to the authentic stop and would lead to a change in C-terminal sequence, (iii) in the 5′-UTR with uORFs which do not trigger NMD, or (iv) there was no evidence of an intron at the suggested position in the transcript.

Only four intron retention transcripts increased significantly in the mutants and/or cycloheximide treatment (Supplementary Tables S2 and S6). One of these (At4g36960) retained intron 1 in the 5′UTR which generated an uORF overlapping the authentic translation start site and therefore is expected to trigger NMD. In the other cases, potential NMD-causing AS events may occur elsewhere in the gene.

In conclusion, our analysis suggests that plant transcripts with retained introns are usually not targets for

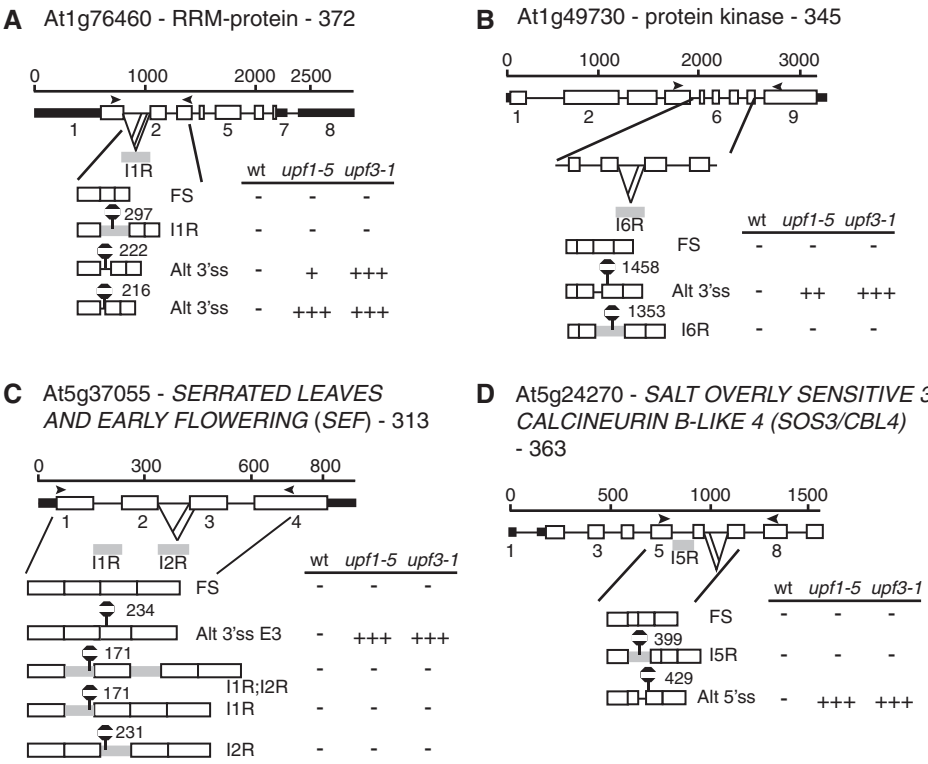


Figure 7. Intron retention transcripts are not turned over by NMD. Schematic figures of genes which produce detectable intron retention transcripts and other alternatively spliced transcripts with different NMD phenotypes: **(A)** At1g76460; **(B)** At1g49730; **(C)** At5g37055 and **(D)** At5g24270. Below each gene, structures of transcripts in the amplified region are shown and display different alternative splicing events. Data for each transcript is shown alongside (−: no change; +, ++ and +++: transcript level increases significantly in *upf* mutants with $P < 0.1$; $0.01 > P < 0.05$ and $P < 0.01$, respectively. For diagram key see legend to Figure 1. Grey lines below introns labelled IR—retained introns; FS—fully spliced; AS—alternatively spliced.

NMD provided there is no other alternative splicing event which produces features for NMD in the transcript.

DISCUSSION

Alternative splicing is a major determinant in the production of variant mRNA transcripts some of which contain PTCs and might be targeted by NMD. This pathway has a significant impact on the expression of genes involved in plant development and adaptation [reviewed (24,25)]. This raises the important questions of how frequently the expression of plant genes is regulated by coupled AS/NMD and what are the structural features of endogenous NMD substrates. Using a high-resolution RT–PCR system we have examined a large population of 950 endogenous transcripts from 270 genes and have characterized over 100 NMD-sensitive AS transcripts in detail. This represents the most extensive and accurate analysis of AS and AS/NMD in endogenous transcripts in plants. We demonstrate (i) a previously unknown high overall prevalence of AS and AS/NMD; (ii) that NMD-sensitive transcripts are readily detected in wild-type plants often representing substantial proportions of the total transcripts of a gene; (iii) that AS in 5′-UTRs and 3′-UTRs regulates transcript levels by rendering them NMD sensitive; (iv) that uORFs overlapping the start codon can trigger NMD; and (v) that transcripts with intron retention events in plants do not trigger NMD even though they possess classical features inducing NMD.

Coupling AS to NMD is a frequent event in plant gene expression

In the course of our analysis we have discovered an unexpectedly high number of novel AS transcripts. This follows from the sensitivity of the RT–PCR system which can detect transcripts of <1% of the total transcripts of a gene; on the other hand, many of the novel transcripts were abundant but not represented in databases. Thus, clearly much more AS is occurring in *Arabidopsis* than is currently estimated and annotated, especially considering that we have assessed AS/NMD only in plants at one developmental stage (3 week old) and have not taken into account many developmental stage-, tissue- or condition-specific AS events.

Genome-wide analyses in eukaryotes have shown that up to 20% of the transcriptome can be affected by NMD [reviewed in (12,53,64)] and that around 20–30% of alternatively spliced transcripts in humans contain PTCs and are potential targets of NMD (37). In plants, little is known about the contribution of NMD to regulation of gene expression. A recent genome-wide tiling array

analysis in *Arabidopsis* found that only around 1% of plant protein-coding genes were affected (50) while an RNA-Seq analysis which showed that 42% of *Arabidopsis* genes underwent AS predicted that around 78% of AS isoforms could be putative targets of NMD (36). By comparing our AS/NMD gene set to those detected using expression/tiling arrays (31,50) we found only two genes in common (data not shown) and, in addition, known AS/NMD substrates such as *GRP7* and *GRP8* or *SR* genes (45,46) were not detected using expression or tiling arrays (31,50). Similarly, very little overlap in NMD-affected gene sets was found in *Drosophila* when comparing expression microarray and splicing-sensitive array results (53). This discrepancy is most likely due to the sensitivity and resolution of the AS RT–PCR panel which is able to detect significant changes in individual transcript levels which would not be detected in microarray experiments where the usual cut-off is >1.5–2-fold.

In this study, 11–17% of the total number of transcripts and 16–25% of the alternatively spliced transcripts analysed were potential NMD substrates suggesting that about 32–44.8% of AS genes are regulated by NMD. Extrapolating from these values and the estimate of the frequency of AS (36), 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* (64,65).

Features of NMD-sensitive transcripts in plants

Rules for NMD in plants have been established based on the behaviour of a small number of genes or artificial constructs (28,32–35). While the general principles of intron-based and long 3′-UTR dependent NMD have been described (Figure 1C), little investigation of this behaviour in endogenous NMD-sensitive transcripts has been performed until now. Here, we determined the features of individual AS transcripts and found that the majority of AS/NMD transcripts (~85%) contained PTCs with downstream splice junctions and/or long 3′-UTRs and therefore comply with existing NMD rules. In plants the average length of the 3′-UTR is 241 nt (66) and our results show that a 'long 3′-UTR' capable of triggering NMD in *Arabidopsis* is usually >350 nt. A similar estimate was obtained using NMD-test constructs where instead of the 3′-UTR length, the distance between a PTC and the authentic stop codon was defined previously as around 300 nt (34). In addition, however, we identified a number of exceptions where transcripts with 3′-UTRs >350 nt were not turned over by NMD suggesting that additional yet unidentified features are involved in triggering NMD.

Our results demonstrate that alternative splicing of introns in either the 3′-UTR or 5′-UTR can determine whether transcripts of endogenous genes are targets of NMD or not and thereby regulate transcript levels. We identified two genes where AS in 3′-UTR introns rendered AS transcripts NMD-sensitive by increasing the distance between the authentic stop codon and the splice junction to more than 50 nt. This agrees with the rules for EJC complexes trigging NMD (Figure 1D). Interestingly, one of the genes showing regulation by AS/NMD in a 3′-UTR intron is the NF-YB1 transcription factor subunit involved in photoperiod-regulated flowering and in drought stress responses, and whose over-expression leads to increased drought resistance (67).

Alternative splicing of introns in 5′-UTRs can change the length and sequence of the 5′-UTR or remove the authentic AUG (again altering the 5′-UTR) and thereby affect the presence/absence, number, size and position of uORFs. In human, polymorphisms or mutations which create or remove uORFs can suppress mRNA and protein levels and cause disease (68). uORFs in 5′-UTR regions can affect gene expression by different mechanisms: encoding an active peptide, affecting translational efficiency or reducing transcript levels by triggering NMD (68–70). In eukaryotes, ribosomes generally load onto mRNAs at the 5′-end and scan to the first AUG translation start codon. If an uORF is translated, the uORF stop codon might be recognized as a PTC (with the additional features of creating a long 3′-UTR and the high likelihood of downstream splice junctions) and thereby targeting the transcript to the NMD pathway. Around 20% of plant genes contain uORFs (35,71) but their fate in terms of whether they are translated or scanned through, trigger NMD or allow re-initiation of translation is not known. We found a strong correlation between presence of an uORF which overlapped the AUG of the main ORF and NMD most likely triggered by generating a long 3′-UTR and downstream splice junctions. We also found other genes where the presence of 'fully upstream' uORFs correlated with activation of NMD which indicates inefficient reinitiation of translation of their main ORFs. However, other AS transcripts contained short uORFs and/or 'long' uORFs (e.g. 43, 55 and 92 amino acids) which did not trigger NMD. Thus, the factors which determine whether or not particular uORFs activate NMD are clearly complex and poorly understood. With around 20% of *Arabidopsis* genes containing uORFs and the frequent occurrence of AS in 5′-UTRs, AS/NMD involving uORFs is likely to be important in regulation of expression of many plant genes.

Retained introns do not trigger NMD

Besides identifying NMD-sensitive AS transcripts, we also identified AS transcripts which contained NMD signals but which were immune to NMD. A comprehensive analysis of AS/NMD in mammalian tissues indicated that not all characteristics of NMD-targeted RNAs have been identified and that not all RNAs containing known NMD features are in fact turned over by NMD (52). Surprisingly, we found that the majority of intron retention transcripts which we analysed were not turned over by NMD despite containing PTCs, downstream splice junctions and long 3′-UTRs. However, transcripts from the same gene with other types of alternative splicing events in the same or nearby intron which generated PTCs in very similar positions were sensitive to NMD. This unexpected finding is in contrast to current assumptions that plant transcripts with retained introns and PTCs

are subject to NMD as such transcripts have been found on ribosomes (72,73), a prerequisite for NMD. More importantly, in other organisms transcripts with retained introns and PTCs are subjected to NMD suggesting a different strategy in plants as intron retention transcripts avoid the NMD machinery and have a different fate [see below; (63,74)]. In addition, our clear demonstration that intron retention events which create a PTC are NMD insensitive may explain the high frequency of IR events identified in plants where it constitutes the major AS event.

We previously detected aberrant mRNAs in the nucleolus of *Arabidopsis* of which around 80% were intron retention events (75). We also found UPF2 and UPF3 to localize to the nucleolus and hypothesized that this may be the site of assembly of NMD factors onto aberrant mRNAs prior to NMD. From the data presented here, IR transcripts avoid NMD and their accumulation in the nucleolus may therefore have a different function. Although the current model of NMD in mammals is that PTCs are recognized in the pioneer round of translation as the mRNA exits the nuclear pore, it is not clear whether this model applies to plants. However, if translation is required for NMD, one possible explanation is that transcripts containing introns or intron fragments are recognized as aberrant prior to export by virtue of proteins binding to the UA-rich intron sequences (75) and therefore do not connect with the NMD machinery. Further research will be needed to determine the fate of different plant PTC-containing transcripts in terms of their intranuclear and intracellular localization, dynamics and transport and how and why intron retention transcripts escape NMD in plants.

Significance of AS/NMD for plant gene expression pattern

The endogenous NMD-sensitive transcripts showed great variation in their steady state levels (i.e. levels detectable in wild-type plants) and in the degree of increased abundance in the different mutants and cycloheximide treatment. This variation is likely to reflect gene-specific differences in transcription levels, frequency of AS producing the different isoforms, or tissue-specific AS occurring only in particular organs or cell types. In addition, the transcript-specific efficiency of NMD turnover could reflect features of the transcripts such as position of PTC and downstream splice junctions, length of 3′-UTR and RNA secondary structure. Importantly, for some genes, non-productive mRNAs (PTC-containing; unable to produce full-length protein) which includes transcripts targeted by NMD and transcripts which are NMD-insensitive such as intron retention transcripts, form a significant proportion of steady state levels of transcripts in wild-type plants. One consequence is that traditional expression microarrays are unable to distinguish between productive and non-productive mRNAs and therefore functional transcript abundance for these genes is over-estimated. Clearly, alternative splicing information must be integrated with transcriptional data to provide true measures of gene expression.

The coupling of AS and NMD is an important general mechanism in gene expression regulation. It modulates the relative levels of mRNA isoforms from a gene which are either productive (protein-coding) or unproductive AS variants and thereby regulates protein levels. Recent examples of plant genes regulated or putatively regulated by AS/NMD in plants are *GRP7/8* and *SOC1* (involved in the circadian clock and flowering control, respectively), *SR* and *PTB* protein splicing factors (involved in a range of developmental and stress response processes) and *HSF2A* (a heat shock factor) (43–48,76). Here, despite only around 270 genes being analysed, AS/NMD has been identified in 121 genes (Table1; Supplementary Table S3) playing central roles in cellular processes: transcription factors, splicing factors, RNA-binding proteins, RNA helicases, spliceosome and exon junction complex proteins, tRNA export, signal recognition particle and ribosomal proteins. Components of developmental pathways also show NMD-mediated turnover of AS transcripts, for example, different *MAF* genes and *VRN2* (flowering time) and *CCA1* and *PRR9* (core circadian clock). Finally, a number of genes involved in signalling and stress response pathways undergo AS/NMD: the calcium-dependent salt stress signalling pathway protein genes *SOS2* and *SOS3*, phosphatases and kinases (e.g. the SNF1-like protein kinase, At*KIN11*) and various temperature, drought and salt response factors (e.g. *SRF2*, *HSF2A*). The identification of many genes in a wide range of processes and pathways suggests that AS/NMD is a widespread regulatory mechanism in plants.

ACCESSION NUMBERS

The Arabidopsis Genome Initiative numbers for *UPF1* and *UPF3* are At5g47010 and At1g33980. AGI locus identifiers of genes analyzed in this article are listed in Supplementary Table S1.

SUPPLEMETARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–6, Supplementary Figures S1–3.

ACKNOWLEDGEMENTS

The authors thank Monika Maronova for help in preparing plant material, and TAIR and especially Philippe Lamesch for providing custom *Arabidopsis* datasets.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC) [BB/G000212/1 and BB/G024979/1 (ERA-NET Plant Genomics (PASAS)]; Scottish Government Rural and Environment Research and Analysis Directorate (RERAD) WP114; EU FP6 Programme Network of Excellence on Alternative Splicing (EURASNET) LSHG-CT-2005-518238; Austrian Science Fund (FWF) [SFB 1710, 1711; DK W1207; ERA-NET Plant

Genomics (PASAS) I254]; Austria Genomic Program (GENAU II, III; ncRNAs). H.Q.D. acknowledges support from the WWTF Science Chair to Arndt von Haeseler. Funding for open access charge: Austrian National Science Foundation (FWF).

Conflict of interest statement. None declared.

REFERENCES

1. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
2. Chen,M. and Manley,J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, **10**, 741–754.
3. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
4. Wahl,M.C., Will,C.L. and Luhrmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
5. 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.
6. 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.
7. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
8. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.
9. Xue,Y., Zhou,Y., Wu,T., Zhu,T., Ji,X., Kwon,Y.S., Zhang,C., Yeo,G., Black,D.L., Sun,H. *et al.* (2009) Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, **36**, 996–1006.
10. Zhong,X.Y., Wang,P., Han,J., Rosenfeld,M.G. and Fu,X.D. (2009) SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol. Cell*, **35**, 1–10.
11. 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.
12. Isken,O. and Maquat,L.E. (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.*, **21**, 1833–1856.
13. Behm-Ansmant,I., Kashima,I., Rehwinkel,J., Sauliere,J., Wittkopp,N. and Izaurralde,E. (2007) mRNA quality control: an ancient machinery recognizes and degrades mRNAs with nonsense codons. *FEBS Lett.*, **581**, 2845–2853.
14. Chang,Y.F., Imam,J.S. and Wilkinson,M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, **76**, 51–74.
15. Shyu,A.B., Wilkinson,M.F. and van Hoof,A. (2008) Messenger RNA regulation: to translate or to degrade. *EMBO J.*, **27**, 471–481.
16. Stalder,L. and Muhlemann,O. (2008) The meaning of nonsense. *Trends Cell Biol*, **18**, 315–321.
17. Rebbapragada,I. and Lykke-Andersen,J. (2009) Execution of nonsense-mediated mRNA decay: what defines a substrate? *Curr. Opin. Cell Biol.*, **21**, 394–402.
18. Nicholson,P., Yepiskoposyan,H., Metze,S., Zamudio Orozco,R., Kleinschmidt,N. and Muhlemann,O. (2010) Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell Mol. Life Sci.*, **67**, 677–700.
19. Le Hir,H., Izaurralde,E., Maquat,L.E. and Moore,M.J. (2000) The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J.*, **19**, 6860–6869.
20. Le Hir,H., Gatfield,D., Izaurralde,E. and Moore,M.J. (2001) The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.*, **20**, 4987–4997.
21. Kashima,I., Yamashita,A., Izumi,N., Kataoka,N., Morishita,R., Hoshino,S., Ohno,M., Dreyfuss,G. and Ohno,S. (2006) Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev.*, **20**, 355–367.
22. Ishigaki,Y., Li,X., Serin,G. and Maquat,L.E. (2001) Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell*, **106**, 607–617.
23. Sauliere,J., Haque,N., Harms,S., Barbosa,I., Blanchette,M. and Le Hir,H. (2010) The exon junction complex differentially marks spliced junctions. *Nat Struct Mol Biol*, **17**, 1269–1271.
24. Belostotsky,D.A. and Sieburth,L.E. (2009) Kill the messenger: mRNA decay and plant development. *Curr. Opin. Plant Biol.*, **12**, 96–102.
25. Chiba,Y. and Green,P. (2009) mRNA degradation machinery in plants. *J. Plant Biol.*, **52**, 114–124.
26. Arciga-Reyes,L., Wootton,L., Kieffer,M. and Davies,B. (2006) UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis. *Plant J.*, **47**, 480–489.
27. Hori,K. and Watanabe,Y. (2005) UPF3 suppresses aberrant spliced mRNA in Arabidopsis. *Plant J.*, **43**, 530–540.
28. Kerenyi,Z., Merai,Z., Hiripi,L., Benkovics,A., Gyula,P., Lacomme,C., Barta,E., Nagy,F. and Silhavy,D. (2008) Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J.*, **27**, 1585–1595.
29. Riehs,N., Akimcheva,S., Puizina,J., Bulankova,P., Idol,R.A., Siroky,J., Schleiffer,A., Schweizer,D., Shippen,D.E. and Riha,K. (2008) Arabidopsis SMG7 protein is required for exit from meiosis. *J. Cell Sci.*, **121**, 2208–2216.
30. Wu,J., Kang,J.H., Hettenhausen,C. and Baldwin,I.T. (2007) Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in Nicotiana attenuata. *Plant J.*, **51**, 693–706.
31. Yoine,M., Ohto,M.A., Onai,K., Mita,S. and Nakamura,K. (2006) The lba1 mutation of UPF1 RNA helicase involved in nonsense-mediated mRNA decay causes pleiotropic phenotypic changes and altered sugar signalling in Arabidopsis. *Plant J.*, **47**, 49–62.
32. Hori,K. and Watanabe,Y. (2007) Context analysis of termination codons in mRNA that are recognized by plant NMD. *Plant Cell Physiol.*, **48**, 1072–1078.
33. Kertesz,S., Kerenyi,Z., Merai,Z., Bartos,I., Palfy,T., Barta,E. and Silhavy,D. (2006) Both introns and long 3'-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res.*, **34**, 6147–6157.
34. Schwartz,A.M., Komarova,T.V., Skulachev,M.V., Zvereva,A.S., Dorokhov Iu,L. and Atabekov,J.G. (2006) Stability of plant mRNAs depends on the length of the 3′-untranslated region. *Biochemistry*, **71**, 1377–1384.
35. Nyiko,T., Sonkoly,B., Merai,Z., Benkovics,A.H. and Silhavy,D. (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol. Biol.*, **71**, 367–378.
36. 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.
37. 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. USA*, **100**, 189–192.
38. Hillman,R.T., Green,R.E. and Brenner,S.E. (2004) An unappreciated role for RNA surveillance. *Genome Biol.*, **5**, R8.
39. Lareau,L.F., Green,R.E., Bhatnagar,R.S. and Brenner,S.E. (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.*, **14**, 273–282.
40. Baek,D. and Green,P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl Acad. Sci. USA*, **102**, 12813–12818.

41. Lareau,L.F., Inada,M., Green,R.E., Wengrod,J.C. and Brenner,S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929.
42. Ni,J.Z., Grate,L., Donohue,J.P., Preston,C., Nobida,N., O'Brien,G., Shiue,L., Clark,T.A., Blume,J.E. and Ares,M. Jr (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708–718.
43. Kalyna,M., Lopato,S., Voronin,V. and Barta,A. (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res.*, **34**, 4395–4405.
44. 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.
45. 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.
46. 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.
47. Stauffer,E., Westermann,A., Wagner,G. and Wachter,A. (2010) Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant J.*, **64**, 243–255.
48. Song,H.R., Song,J.D., Cho,J.N., Amasino,R.M., Noh,B. and Noh,Y.S. (2009) The RNA binding protein ELF9 directly reduces SUPPRESSOR OF OVEREXPRESSION OF CO1 transcript levels in arabidopsis, possibly via nonsense-mediated mRNA decay. *Plant Cell*, **21**, 1195–1211.
49. 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.
50. Kurihara,Y., Matsui,A., Hanada,K., Kawashima,M., Ishida,J., Morosawa,T., Tanaka,M., Kaminuma,E., Mochizuki,Y., Matsushima,A. *et al.* (2009) Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **106**, 2453–2458.
51. Barberan-Soler,S. and Zahler,A.M. (2008) Alternative splicing regulation during C. elegans development: splicing factors as regulated targets. *PLoS Genet.*, **4**, e1000001.
52. Pan,Q., Saltzman,A.L., Kim,Y.K., Misquitta,C., Shai,O., Maquat,L.E., Frey,B.J. and Blencowe,B.J. (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.*, **20**, 153–158.
53. Hansen,K.D., Lareau,L.F., Blanchette,M., Green,R.E., Meng,Q., Rehwinkel,J., Gallusser,F.L., Izaurralde,E., Rio,D.C., Dudoit,S. *et al.* (2009) Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS Genet*, **5**, e1000525.
54. Simpson,C.G., Fuller,J., Maronova,M., Kalyna,M., Davidson,D., McNicol,J., Barta,A. and Brown,J.W. (2008) Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J.*, **53**, 1035–1048.
55. Valvekens,D., Van Montague,M. and Van Lijsebettens,M. (1988) Agrobacterium tumefaciens-mediated transformation of Arabidopsis thaliana root explants using kanamycin selection. *Proc. Natl Acad. Sci. USA*, **85**, 5536–5540.
56. Murashige,T. and Skoog,F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Plant Physiol.*, **15**, 473–497.
57. Simpson,C.G., Clark,G.P., Lyon,J.M., Watters,J., McQuade,C. and Brown,J.W.S. (1999) Interactions between introns via exon definition in plant pre-mRNA splicing. *Plant J.*, **18**, 293–302.
58. Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.
59. Carter,M.S., Doskow,J., Morris,P., Li,S., Nhim,R.P., Sandstedt,S. and Wilkinson,M.F. (1995) A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts in vivo is reversed by protein synthesis inhibitors in vitro. *J. Biol. Chem.*, **270**, 28995–29003.
60. Saul,H., Elharrar,E., Gaash,R., Eliaz,D., Valenci,M., Akua,T., Avramov,M., Frankel,N., Berezin,I., Gottlieb,D. *et al.* (2009) The upstream open reading frame of the Arabidopsis AtMHX gene has a strong impact on transcript accumulation through the nonsense-mediated mRNA decay pathway. *Plant J.*, **60**, 1031–1042.
61. Nagy,E. and Maquat,L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, **23**, 198–199.
62. Barbazuk,W.B., Fu,Y. and McGinnis,K.M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.*, **18**, 1381–1392.
63. de Lima Morais,D.A. and Harrison,P.M. (2010) Large-scale evidence for conservation of NMD candidature across mammals. *PLoS ONE*, **5**, e11695.
64. Ramani,A.K., Nelson,A.C., Kapranov,P., Bell,I., Gingeras,T.R. and Fraser,A.G. (2009) High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective Caenorhabditis elegans. *Genome Biol.*, **10**, R101.
65. Rehwinkel,J., Letunic,I., Raes,J., Bork,P. and Izaurralde,E. (2005) Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA*, **11**, 1530–1544.
66. Pesole,G., Liuni,S., Grillo,G. and Saccone,C. (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. *Gene*, **205**, 95–102.
67. Nelson,D.E., Repetti,P.P., Adams,T.R., Creelman,R.A., Wu,J., Warner,D.C., Anstrom,D.C., Bensen,R.J., Castiglioni,P.P., Donnarummo,M.G. *et al.* (2007) Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc. Natl Acad. Sci. USA*, **104**, 16450–16455.
68. Calvo,S.E., Pagliarini,D.J. and Mootha,V.K. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl Acad. Sci. USA*, **106**, 7507–7512.
69. Morris,D.R. and Geballe,A.P. (2000) Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.*, **20**, 8635–8642.
70. Meijer,H.A. and Thomas,A.A. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem J.*, **367**, 1–11.
71. Kochetov,A.V., Sirnik,O.A., Rogosin,I.B., Glazko,G.V., Komarova,M.L. and Shumny,V.K. (2002) Contextual features of higher plant mRNA 5' untranslated regions. *Mol. Biol.*, **36**, 510–516.
72. 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.
73. Jiao,Y. and Meyerowitz,E.M. (2010) Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol. Syst. Biol.*, **6**, 419.
74. Sayani,S., Janis,M., Lee,C.Y., Toesca,I. and Chanfreau,G.F. (2008) Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol. Cell*, **31**, 360–370.
75. Kim,S.H., Koroleva,O.A., Lewandowska,D., Pendle,A.F., Clark,G.P., Simpson,C.G., Shaw,P.J. and Brown,J.W. (2009) Aberrant mRNA transcripts and the nonsense-mediated decay proteins UPF2 and UPF3 are enriched in the Arabidopsis nucleolus. *Plant Cell*, **21**, 2045–2057.
76. Sugio,A., Dreos,R., Aparicio,F. and Maule,A.J. (2009) The cytosolic protein response as a subcomponent of the wider heat shock response in Arabidopsis. *Plant Cell*, **21**, 642–654.
77. 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. *Plant Cell*, **22**, 2926–2929.

Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis.

Marquez, Y., Brown, J. W., Simpson, C., Barta, A., & Kalyna, M. (2012).

Genome Res, 22(6), 1184-1195.

Author contribution: Experiment design, samples preparation and bioinformatic analyses.

Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*

Yamile Marquez, John W.S. Brown, Craig Simpson, et al.

Genome Res. 2012 22: 1184-1195 originally published online March 5, 2012
Access the most recent version at doi:10.1101/gr.134106.111

Resource

Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*

Yamile Marquez,[1] John W.S. Brown,[2,3] Craig Simpson,[2] Andrea Barta,[1,4] and Maria Kalyna[1,4]

[1]*Max F. Perutz Laboratories, Medical University of Vienna, A-1030 Vienna, Austria;* [2]*Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom;* [3]*Division of Plant Sciences, University of Dundee at JHI, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom*

Alternative splicing (AS) is a key regulatory mechanism that contributes to transcriptome and proteome diversity. As very few genome-wide studies analyzing AS in plants are available, we have performed high-throughput sequencing of a normalized cDNA library which resulted in a high coverage transcriptome map of *Arabidopsis*. We detect ~150,000 splice junctions derived mostly from typical plant introns, including an eightfold increase in the number of U12 introns (2069). Around 61% of multiexonic genes are alternatively spliced under normal growth conditions. Moreover, we provide experimental validation of 540 AS transcripts (from 256 genes coding for important regulatory factors) using high-resolution RT-PCR and Sanger sequencing. Intron retention (IR) is the most frequent AS event (~40%), but many IRs have relatively low read coverage and are less well-represented in assembled transcripts. Additionally, ~51% of *Arabidopsis* genes produce AS transcripts which do not involve IR. Therefore, the significance of IR in generating transcript diversity was generally overestimated in previous assessments. IR analysis allowed the identification of a large set of cryptic introns inside annotated coding exons. Importantly, a significant fraction of these cryptic introns are spliced out in frame, indicating a role in protein diversity. Furthermore, we show extensive AS coupled to nonsense-mediated decay in *AFC2*, encoding a highly conserved LAMMER kinase which phosphorylates splicing factors, thus establishing a complex loop in AS regulation. We provide the most comprehensive analysis of AS to date which will serve as a valuable resource for the plant community to study transcriptome complexity and gene regulation.

[Supplemental material is available for this article.]

Alternative splicing (AS) is a widespread mechanism which increases transcriptome and proteome complexity and controls developmental programs and responses to the environment in higher eukaryotes. The splicing process, removal of introns and ligation of exons, is performed by a large RNA-protein complex, the spliceosome, consisting of five small nuclear RNAs (snRNAs) and about 180 proteins with different functions (Wahl et al. 2009). Assembly of the spliceosome on introns in a precursor messenger RNA (pre-mRNA) is directed by *cis* elements and *trans*-acting factors (Black 2003; Stamm et al. 2005). The *cis* sequences include the splice sites, branchpoint, and polypyrimidine tract which have degenerate consensus sequences in higher eukaryotes. While many splice sites are selected in all transcripts (constitutive splicing), others are used to various levels, resulting in alternative transcripts. Selection of such alternative splice sites is affected by auxiliary *cis* elements located within exonic and intronic sequences, termed splicing enhancers and silencers. These elements are binding sites for *trans*-acting splicing factors, for example, hnRNP and SR proteins. These proteins, in addition to their functions in constitutive splicing, play a key role in AS by inhibition or promotion of selection of particular splice sites. The presence and abundance of different splicing factors in different cell types, tissues, developmental stages, and environmental conditions determines the AS profiles of expressed genes and ultimately shapes the transcriptome. In addition, alternative transcripts can code for protein isoforms with altered amino acid and domain composition affecting their activity, interaction capacity, localization, and stability, thus affecting the proteome (Stamm et al. 2005).

Alternative splicing was first described in 1977 as peculiar rearrangements in the adenovirus type 2 mRNA (Berget et al. 1977; Chow et al. 1977). Since the discovery of the first example of AS in an endogenous mammalian gene coding for calcitonin (Rosenfeld et al. 1981), the alignment of expressed sequence tag (EST) contigs to genomic DNA allowed the identification of a large number (~35%) of alternatively spliced genes in humans (Mironov et al. 1999). Estimates of AS in many different organisms have been made using EST/cDNA libraries (Okazaki et al. 2002; Zavolan et al. 2003; Iida et al. 2004; Cusack and Wolfe 2005; Wakamatsu et al. 2009). With the advent of tiling arrays and high-throughput sequencing, the number of genes which undergo AS has continued to increase (Jones-Rhoades et al. 2007; Weber et al. 2007; Kwan et al. 2008; Mortazavi et al. 2008; Pan et al. 2008). In particular, the application of high-throughput sequencing to transcriptomes (RNA-seq) has now demonstrated that AS occurs in ~95% of intron-containing genes in human (Pan et al. 2008).

In plants, estimates of the occurrence of AS have been hampered by a low number of ESTs (Brett et al. 2002). However, the levels of AS have continued to increase with greater EST/cDNA coverage: 1.2% (Zhu et al. 2003), 5% (Zhu et al. 2003), 11.6% (Iida et al. 2004), 21.8% (Wang and Brendel 2006), 29% (Xiao et al. 2005), and >30% (Campbell et al. 2006). Many transcriptome studies using high-throughput sequencing have been performed in plants, but few have been used to examine AS (Weber et al. 2007;

[4]Corresponding authors.
E-mail mariya.kalyna@univie.ac.at.
E-mail andrea.barta@meduniwien.ac.at.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.134106.111. Freely available online through the *Genome Research* Open Access option.

22:1184–1195 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

Filichkin et al. 2010; Lu et al. 2010; Zhang et al. 2010). The most recent estimate based on RNA-seq is that ~42% of *Arabidopsis* intron-containing genes undergo AS (Filichkin et al. 2010).

In terms of identifying AS in plants, the expression profile itself influences the representation of many transcripts in databases. For example, an *Arabidopsis* transcriptome study using 454 Life Sciences (Roche) sequencing (Weber et al. 2007) showed that the top 10 most highly expressed genes represent 25% of the total mapped reads, thus tremendously compromising the representation of less abundant transcripts. To improve gene representation and discovery of AS events in *Arabidopsis*, we have used RNA-seq of a normalized cDNA library made from *Arabidopsis* seedlings and flowers. We have shown that normalization significantly increases the coverage of reads across the genes, and we have identified a large number (~47 k) of new splice junctions. Taking advantage of a high-resolution RT-PCR panel (Simpson et al. 2008a,b), we were able to validate many novel AS events. Altogether, our results show that at least 61% of intron-containing genes are alternatively spliced under normal growth conditions, which indicates a high complexity of the *Arabidopsis* transcriptome.

Results

RNA-seq of a normalized cDNA library improves gene representation

To facilitate the discovery of alternative splicing events in *Arabidopsis*, we generated a normalized cDNA library (Supplemental Figs. S1, S2) and subjected it to paired-end Illumina sequencing. The library was constructed from total RNA isolated from *Arabidopsis* Col-0 wild-type flowers and 10-d-old seedlings, mixed in equal proportions (Supplemental Fig. S1). Sequencing resulted in nearly 116 million paired-end reads (75-nt read length). To our knowledge, this is the first time that a normalized cDNA library of *Arabidopsis* has been subjected to high-throughput sequencing.

Paired-end reads were aligned to the *Arabidopsis thaliana* reference genome version TAIR9 (www.arabidopsis.org) using Bowtie (Langmead et al. 2009), included in the TopHat software v1.0.14 (Trapnell et al. 2009). We mapped nearly 74% of the starting reads to the reference genome, of which almost 97% aligned uniquely (Fig. 1A). Additionally, the number of aligned reads per chromosome correlated with the chromosome size (Fig. 1B), implying extensive chromosome coverage. The read distribution along each chromosome in windows of 1 kb is shown in Figure 1C (see Methods).

To test whether the uniqueness of the reads was due to the length of the read and/or the cDNA normalization procedure, we compared our results with the 36-nt single-end read Illumina sequences from non-normalized cDNA libraries published previously (Filichkin et al. 2010). In addition, to allow a direct comparison of reads of the same size, we truncated our 75-nt-long reads to 36 nt (Supplemental Table S2). These comparisons suggested that the increased read length and cDNA normalization both significantly influenced read mapping, resulting in an increase in the total number of aligned reads and a higher proportion of uniquely mapped reads (Supplemental Table S2).

Most of the reads mapped to annotated genes (Supplemental Fig. S5A), illustrating that the majority of genes have been predicted in the TAIR9 annotation. Examination of the median coverage along the transcription unit (Supplemental Fig. S5B; see Methods) revealed that the cDNAs are highly covered in almost

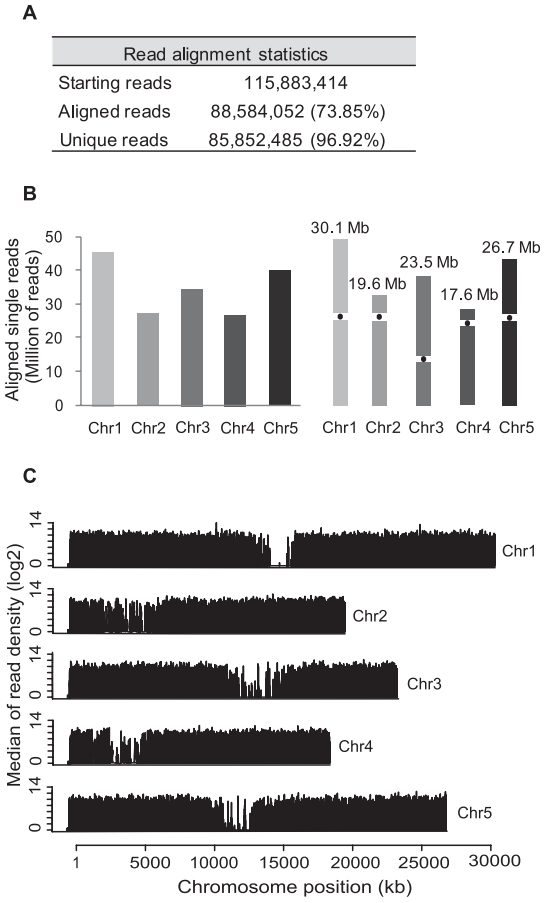


Figure 1. Read alignment using Bowtie and TopHat. (*A*) Table showing statistics of the aligned reads to the *Arabidopsis* genome (TAIR9). (*B*) Aligned single reads to *Arabidopsis* chromosomes (*left*) and schematic representation of *Arabidopsis* chromosomes (*right*). (*C*) Log_2 scale of median read density in windows of 1 kb by chromosome.

their entire length (median of 600 reads in 80% of the total length of the transcription unit) (Supplemental Fig. S5B).

To test the efficiency of normalization in reducing the abundance of highly expressed transcripts, we analyzed the read coverage of the top 10 most abundantly expressed genes in our normalized library and showed that they only contributed 1.05% of the total of reads (Supplemental Material; Supplemental Table S1). This level of coverage was substantially lower than found in non-normalized libraries sequenced with the Illumina (Filichkin et al. 2010) and 454 (Weber et al. 2007) platforms, where the proportion of reads in the top 10 most abundant genes was 26.76% and 7.76%, and 25%, respectively (Supplemental Table S1). The much lower read coverage of highly expressed genes in our libraries was independent of the read length and paired-end protocol (Supplemental Table S1; Supplemental Fig. S3A,B). A detailed description of the normalization efficiency is given in the Supplemental Material (Supplemental Data; Supplemental Table S1; Supplemental Figs. S3, S4). In general, the comparison between library preparation and sequencing methods suggested that the normalized library achieved much better gene coverage and significantly influenced read mapping, resulting in an increase in the total number of aligned reads and a higher proportion of uniquely mapped reads.

High level of detection of novel splice junctions in *Arabidopsis*

For splice junction detection, we used TopHat (Trapnell et al. 2009). To reduce the number of false positives, the minimum intron size was set to 60 nt, which is close to that established experimentally (Goodall and Filipowicz 1990) This is in contrast to other studies (Filichkin et al. 2010; Zhang et al. 2010) which used smaller intron sizes of 20 nt and 1 nt, respectively. We filtered the splice junctions originally detected (using alignment with two mismatches) (see Methods) by two criteria: quality of the alignment and coverage of the splice junction (Supplemental Fig. S1). Splice junctions that were obtained from the alignment with two mismatches (see Methods) were removed if the splice junction in question was within 10 nt of another splice junction with better support and if it was not supported by the alignment of perfectly aligned reads (alignment with no mismatches) (Supplemental Fig. S1; see Methods). For the remaining junctions, we have only included those supported by at least three reads (two mismatches). Our results showed that the majority of the splice junctions reside in annotated genes (131,523 of 149,925) (Table 1) and that, in protein-coding genes, they are located mainly in the coding sequence (CDS) (Fig. 2A), highlighting the potential of splice junctions to affect the final protein sequence. From the total of predicted splice junctions, 46,955 (31%) were novel and not annotated in the TAIR9 Database (Table 1). This is most likely due to the use of a normalized library together with longer reads which have considerably improved splice junction detection (Supplemental Table S2).

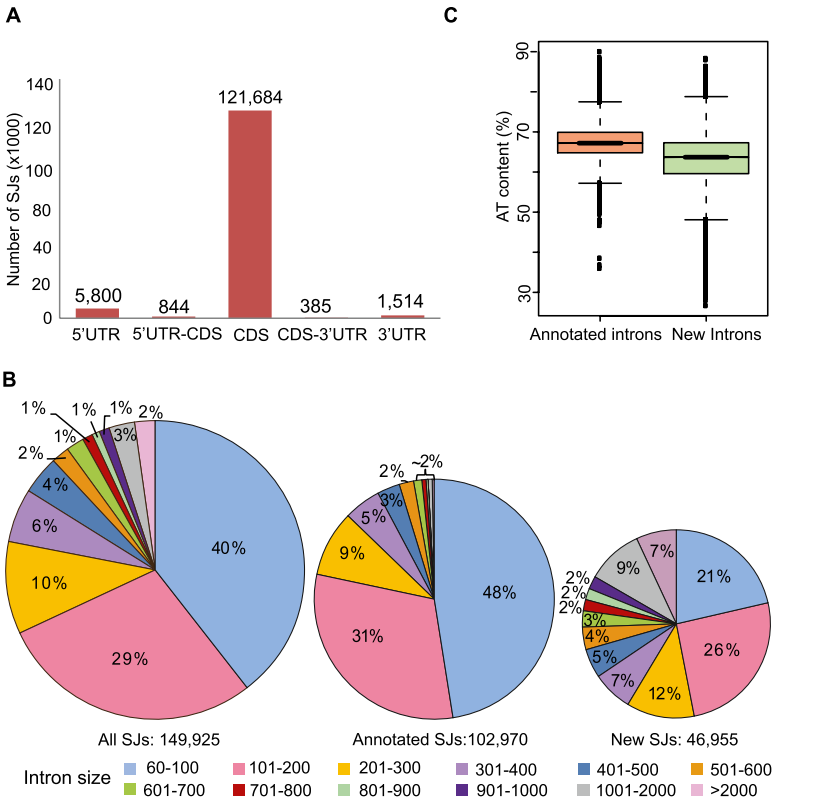


Figure 2. Features of splice junctions and predicted introns. (*A*) Distribution of splice junctions along gene features in protein-coding genes, according to TAIR9. (*B*) Distribution of intron sizes defined by splice junctions (SJ) predicted by TopHat (*left* circle). Splice junctions were classified as annotated if they were already in TAIR9 (*middle* circle); otherwise, they were classified as new (*right* circle). (*C*) Distribution of the AT content (%) in introns generated by annotated splice junctions and new splice junctions.

The majority of splice junctions define typical plant introns

In plants, several features that impact splicing efficiency have been described. These include a minimum intron size, sequences in the 5′ and 3′ splice sites, the branch point, and a minimum percentage of AU-richness (Goodall and Filipowicz 1990; Brown and Simpson 1998; Lorkovic et al. 2000). Although the presence of these signals will define an intron, it is the combination and strength of each signal which determines the splicing efficiency of intron removal (Brown and Simpson 1998).

Inspection of the intron sizes produced by all of the predicted splice junctions had a mean of 298 nt (median = 114 nt) and showed that the majority (~70%) of them were smaller than 200 nt (Fig. 2B). However, the sizes of introns generated by the new splice junctions were bigger than those from the junctions annotated in TAIR (mean = 588.7 nt versus mean = 166.5 nt, respectively, Mann-Whitney-Wilcoxon test, *P*-value < 0.00001). This was also highlighted by only ~47% of them being smaller than 200 nt (Fig. 2B). This is probably due to many of the novel splice junctions defining new AS events and that introns with AS tend to be larger than introns with no evidence of AS (see Fig. 5A, below).

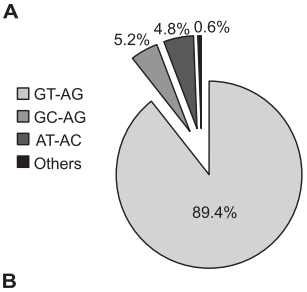
Examination of the dinucleotides at the intron borders exhibited an increase in the number of GC-AG and AT-AC splice sites in comparison to TAIR9 annotations (5.2% and 4.8% vs. 1% and 0.6%, respectively) (Fig. 3A). In order to demonstrate that the introns predicted by these splice junctions were bona fide introns, we looked for sequence signatures of plant introns using position weight matrices (PWMs) (Sheth et al. 2006). Using PWMs of the 5′ and 3′ splice sites of U2 and U12 annotated introns in *Arabidopsis*, together with their respective branch point signatures (Fig. 3B; see Methods), we were able to classify ~93% of the introns. As expected, most of them were classified as U2 introns, but we also identified 2069 potential U12 introns in genic regions (Fig. 3A;

Table 1. Predicted splice junctions (SJs)

	Score ≥ 1	Score ≥ 2	Score ≥ 3
Total of SJs	172,651	159,758	149,925
SJs in annotated genes	144,628	137,409	131,523
Annotated SJs	105,237	104,070	102,970
New SJs	67,414	55,688	46,955
Genes with at least one intron[a]	20,144	19,800	19,505
Genes with overlapping introns[a]	11,681 (58.0%)	10,879 (54.9%)	10,103 (51.8%)

[a]Introns are defined by the predicted splice junctions (SJs) (see text).

A



B

	U2 introns	U12 introns	NC	Total
Genic	124,383	2,069	5,071	131,523
Antisense	6,071	193	4,765	11,029
Intergenic	6,086	150	1,137	7,373
Total	136,540	2,412	10,973	149,925

Figure 3. Classification of introns defined by predicted splice junctions. (*A*) Splice sites of predicted introns. (*B*) Classification of introns according to Sheth et al. 2006 (see Methods). (NC) Not classified.

Supplemental Fig. S6; see the list of U12 introns in the Supplemental Material), and only 165 and 246 of such introns were identified computationally (Zhu and Brendel 2003; Alioto 2007). Functional classification (GO annotation) of the genes possessing U12 introns shows that they are 2.7-fold enriched in DNA and RNA metabolism class (*P*-value = 1.039×10^{-10}; data not shown).

We also examined the AU-richness of introns predicted by the splice junctions and showed that ~92% had >59% AU, which is the minimum AU% required for an intron to be efficiently spliced in dicots (Goodall and Filipowicz 1989; Fig. 2C). The AU-richness of introns predicted by the novel splice junctions was slightly lower (~64%) than that predicted by the known (annotated) splice junctions (~67%) (Fig. 2C). This might reflect the fact that the majority of the novel splice junctions are derived from new AS events (see below). Together, the results suggest that a high proportion of the introns generated from the predicted splice junctions in this study have typical characteristics of introns in *Arabidopsis*.

To investigate the potential of the introns (defined by the predicted splice junctions) to produce alternative transcripts, we looked for overlapping introns inside the same gene, as two overlapping introns cannot coexist in the same transcriptional unit. Using this approach, we estimate that 51.8% of the genes in our sample have more than one transcript and, therefore, have evidence of AS (Table 1). It is important to note that this analysis of overlapping introns does not provide data on intron retention events.

Assembly of putative transcripts predicts that 61.2% of intron-containing genes are alternatively spliced

To determine putative transcripts generated by RNA-seq, we have used the gene coordinates from TAIR9, the aligned reads, and the predicted splice junctions that have passed the quality criteria described above. Putative exons were defined as continuous stretches with at least four reads of coverage. The final transcripts were assembled by joining the putative exons with the predicted junctions. To test whether the predicted transcripts were supported by the read alignment, the putative transcripts were evaluated using Cufflinks (Trapnell et al. 2010). From this analysis, a total of 57,447 transcripts were successfully assembled that corresponded to 23,901 annotated genes (Table 2; see the set of assembled transcripts in the Supplemental Material). In agreement with what was observed in the read alignment (Fig. 1B), the number of genes and transcripts that were assembled by chromosome also correlates with the chromosome size (Table 2). Moreover, in the final set of predicted transcripts, 91% of the genic splice junctions reported in Table 1 (score ≥ 3) are contained in at least one putative transcribed unit.

Next, we evaluated whether the set of putative assembled transcripts corresponded to known isoforms in TAIR9. For this purpose, we calculated how many genes with assembled transcripts have at least one isoform that resembled an annotated transcript in TAIR9, using the definition that the assembled transcript covered contiguously at least 80% of the total length of the annotated isoform. We found that 75% of genes with assembled transcripts had at least one transcript that is annotated in TAIR9 (data not shown).

We estimated the number of *Arabidopsis* intron-containing genes with AS by calculating the ratio of genes with more than one putative transcript divided by the number of genes with at least one splice junction in our sample (intron-containing genes). Therefore, 61.2% of genes with introns in our sample have evidence of AS (Table 2). The difference between this value and the estimate of 51.8% based only on overlapping introns (Table 1) reflects the addition of intron retention transcripts and the filtering criteria which results in loss of some splice junctions (only 91% of splice junctions are included).

To determine the types of AS events in our set of assembled transcripts, we used ASTALAVISTA software (Foissac and Sammeth 2007). Intron retention is the most common event (~40%, Figure 4A; Supplemental Fig. S7) agreeing with previous studies in plants (Iida et al. 2004; Ner-Gaon and Fluhr 2006; Wang and Brendel 2006; Barbazuk et al. 2008; Filichkin et al. 2010) although it is not as high as the estimate of ~56% of the total events (Wang and Brendel 2006). We found that the use of alternative 3′ splice sites is more than twofold more frequent than the use of alternative 5′ splice sites (Fig. 4A; Supplemental Fig. S7), consistent with prior findings (Iida et al. 2004; Wang and Brendel 2006). Moreover, our transcripts show that while skipping of a single exon is a rare event (2.73% of the total events), it is relatively common that multiple exons are skipped together and that exon skipping can utilize alternative 5′ and/or 3′ splice sites (Fig. 4A; Supplemental Fig. S7). In general, ~42% of the events in the assembled transcripts are complex (Fig. 4A; Supplemental Fig. S7).

Table 2. Putative assembled transcripts and AS estimates

Putative assembled transcript statistics		
Chr	**Gene number**	**Transcript number**
Chr1	6,253	14,888
Chr2	3,761	9,391
Chr3	4,750	11,253
Chr4	3,610	8,935
Chr5	5,527	12,980
Total	23,901	57,447

Alternative splicing estimates			
	RNA-seq (this study)	TAIR9	TAIR10
Intron-containing genes	18,719	22,403	22,523
AS genes	11,465	4,626	5,885
% of AS	61.2%	20.7%	26.1%

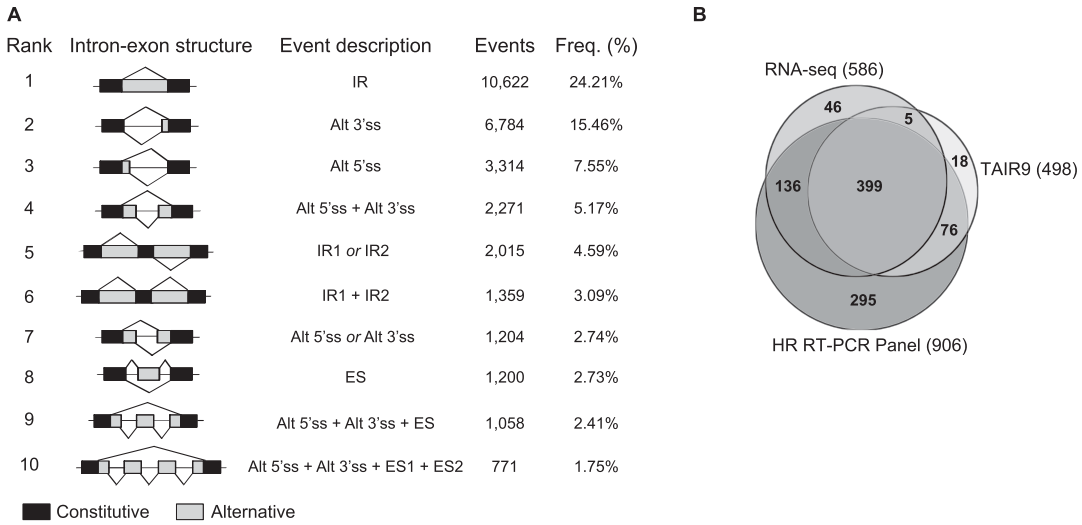


Figure 4. Alternative splicing events and validation of assembled transcripts. (*A*) Top 10 most frequent types of AS in the predicted transcripts according to ASTALAVISTA. The first column illustrates the intron-exon structure of the AS event, followed by its description, the raw number of events found in the sample, and their frequency. (IR) Intron retention, (Alt 5′ss) alternative 5′ splice site, (Alt 3′ss) alternative 3′ splice site, (ES) exon skipping. (*B*) Venn diagram of the number of fragments obtained in the HR RT-PCR panel, putative assembled transcripts of RNA-seq, and TAIR9-annotated transcripts according to primer pairs of the HR RT-PCR panel (see text and Methods).

Validation of assembled transcripts

To validate our set of putative assembled transcripts (PATs), we have used a high-resolution RT-PCR alternative splicing panel (HR RT-PCR panel), which allows monitoring of AS events in multiple genes under different conditions (Simpson et al. 2008a,b). This HR RT-PCR panel is capable of distinguishing AS events involving small size differences in transcripts (as few as 2–3 nt). We analyzed 256 different genes containing AS events which were either published, annotated in TAIR, or found in AS databases (Supplemental Table S3). Applying the HR RT-PCR panel primers to our assembled transcripts, we extracted the sizes of expected RT-PCR products for each of the genes on the panel and compared them to actual products detected on the panel and to predicted products from TAIR9-annotated AS transcripts. Of the 586 products predicted from the RNA-seq assembled transcripts, 92% were supported by at least one of these two resources (Fig. 4B; Supplemental Table S4). Furthermore, 136 (23.2%) of the predicted products are not described in the TAIR9 database but were confirmed by products on the HR RT-PCR panel (Fig. 4B; Supplemental Table S4), again illustrating that many AS events are not annotated. As expected, the HR RT-PCR panel detected the highest number of AS events. This is due, first, to its sensitivity and second, to the different samples and conditions employed. Nevertheless, 33 transcripts of low abundance (relative transcript abundance of <2%) in the HR RT-PCR panel were also indentified by RNA-seq. Moreover, RNA-seq identified a number of AS events which had not been scored on the HR RT-PCR panel previously because they represented very small peaks but were now confirmed by manual inspection. Additionally, some RT-PCR products were subjected to Sanger sequencing, and a total of 124 sequences supported the assemblies in our RNA-seq experiment. Altogether, the results highlight the validity of RNA-seq to detect novel AS events and demonstrate that AS is much more extensive in *Arabidopsis* than previously thought.

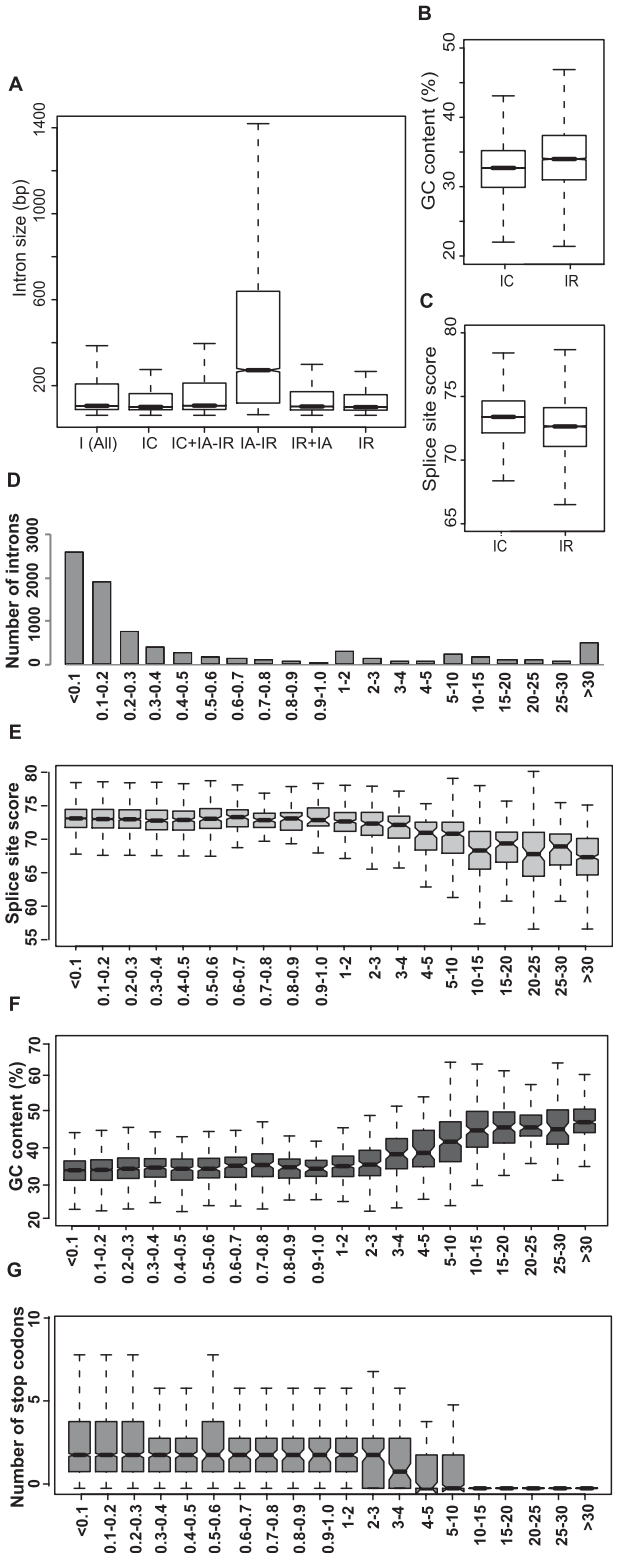
Low abundance of many intron retention events

Examination of the putative assembled transcripts showed that the predominant type of AS event is intron retention, with ~40% of the total events (Fig. 4A; Supplemental Fig. S7; and see above). We analyzed the characteristics of retained introns identified by RNA-seq and transcript assembly. We compared the size of retained introns (IR) to the sizes of introns with no evidence of retention or any other types of AS (constitutive intron - IC) and to alternatively spliced introns with no evidence of intron retention (IA-IR). We found that retained introns were significantly smaller than non-retained introns which were involved in other types of AS (IR vs. IA-IR, mean 161.39 bp vs. 548.86 bp, *P*-value < 0.00001) (Fig. 5A). On the other hand, there was no significant difference in the size distribution of retained introns compared to constitutive introns (IR vs. IC, *P*-value = 0.5) (Fig. 5A). Therefore, intron size cannot be used as a parameter to distinguish retained introns. Nevertheless, we found that retained introns had a higher GC content (*P*-value < 0.00001) (Fig. 5B) and weaker splice signals (*P*-value < 0.00001) (Fig. 5C) than constitutive introns, which might contribute to lower splicing efficiency and, therefore, a higher probability of intron retention.

We obtained an indication of the degree to which an intron is retained by calculating the median number of reads along the retained intron divided by the number of reads supporting the splice junction in question (see Methods). This value, the intron retention ratio (IRR), reflects to some extent the preference of an intron to be retained. Nearly 73% of the retained introns had an IRR of <1 and around 48% had an IRR of <0.2 (Fig. 5D), suggesting that these introns have a lower read coverage. Deeper analysis of the intron retention events showed that the preference for removing an intron (low IRR) correlated to some extent with stronger splice sites in comparison with introns that have higher IRR ($R^2 = -0.36$, *P*-value = 2.7×10^{-269}). For example, introns with an IRR < 4 have stronger splicing signals ($P < 0.05$) (Fig. 5E).

We, therefore, analyzed the abundance of intron retention events on the HR RT-PCR panel. Primers were designed to amplify across at least two introns, including the intron(s) which had been described in different plant databases as retained. Of 51 intron retention events from 45 genes, 33 were detected by RT-PCR and were supported by RNA-seq. Nine of the IR products were not de-

tectable on the HR RT-PCR panel, and eight of these were not detected by RNA-seq (Supplemental Table S5). For a further nine IR events, the IR products were barely detectable on the HR RT-PCR panel (<1% of the total transcripts from the gene) and had low coverage of RNA-seq reads (Supplemental Table S5; see also Supplemental Fig. S9). Thus, there is good agreement between the read coverage and detection by RNA-seq and the relative abundance of intron retention transcripts detected experimentally by the HR RT-PCR panel. Based on the lack of detection or very low abundance of some IR events which appear in plant databases, it is most likely that many are derived from ESTs representing partially spliced mRNAs or DNA artifacts.



Identification of an abundant class of cryptic introns

We examined the distribution of the retained introns along the transcripts to see whether they distribute differently in comparison to all the introns (Fig. 2A). We obtained a significant enrichment of retained introns in the 5′ and 3′ UTRs of genes (twofold and 3.4-fold, respectively, χ^2 test, $\chi^2 = 1176$, P-value $= 2.54 \times 10^{-253}$) (Supplemental Table S6), suggesting that intron retention is selected against in the coding regions. This observation appeared to contradict the finding that 1838 introns were preferentially retained in our sample (IRR > 1), and many had very high IRR scores (Fig. 5D). More detailed analysis of these introns with IRR > 1 showed that 66.86% (1227) were inside annotated exons. In addition, introns with an IRR > 3 had higher GC content than introns with lower IRRs (P-value < 0.05) (Fig. 5F) and, in general, retained introns are more GC rich than nonretained introns (P-value < 0.0001) (Fig. 5B; Supplemental Fig. S8). The position of these retained introns within annotated exons and the higher GC content suggests that they are exonlike. This is also supported by an analysis of the protein coding capacity of these retained intron sequences. First, they contribute significantly to known protein domains compared to nonretained introns or retained introns from UTRs (Supplemental Tables S7, S8), and second, the retained introns have a significantly lower proportion of stop codons than expected from their nucleotide frequency (χ^2 test, $\chi^2 = 1313.7$, P-value $= 1.19 \times 10^{-287}$). In particular, retained introns with an IRR > 4 have fewer stop codons (P-value < 0.05) (Fig. 5G). Additionally, the codon usage of retained introns with an IRR > 3 is more similar to the codon usage of exons than from nonretained introns (Supplemental Table S9; see Methods).

The exonic features in these retained introns suggest that they represent cryptic introns rather than retained introns. In the following analysis, we define a cryptic intron as an intron with both splice sites inside an annotated coding exon. To validate the removal of such exonic regions identified by RNA-seq, we selected 10 cryptic introns with different IRRs and tested them by RT-PCR. We used cDNA from our initial RNA preparations to exclude any possibility that our results are the consequence of the library preparation or computational analysis. Primer pairs were designed across more than one exon to distinguish any DNA contamination or unprocessed products. From the 10 genes tested, nine showed

Figure 5. Features of retained introns. (*A*) Size distribution of different intron classes. [I(All)] All introns in our sample, (IC) the constitutive introns, (IC+IA−IR) and (IA−IR) the categories without retained introns either including or excluding constitutive introns, respectively, (IR+IA) retained introns that are also involved in other AS events, (IR) introns that are only retained. (*B*) GC content distribution of constitutive introns (IC) and retained introns (IR). (*C*) Splice site score distribution of constitutive introns (IC) and retained introns (IR). (*D*) Histogram of number of introns in each intron retention ratio (IRR) category. (*E*) Boxplots of splice site score distribution by IRR category. (*F*) Boxplots of GC content distribution by IRR category. (*G*) Boxplots of stop codons distribution by IRR category. For the splice site score distributions in *C* and *E*, the means of 5′ and 3′ splice sites scores (calculated according to Sheth et al. 2006) were used.

an RT-PCR product that corresponded to cryptic intron splicing (Supplemental Fig. S10; Supplemental Table S10).

In total, we have identified 1289 cryptic introns in annotated coding exons (14.1% of all retained introns). Almost half of these cryptic introns (602) are spliced out in frame, thus having a potential of removing amino acid stretches within a full-length protein and generating a new protein isoform.

Alternative splicing of *AFC2* Clk/STY (LAMMER-type) kinase

To demonstrate the utility of our RNA-seq data, we have investigated the AS of the *AFC2* gene. AFC2 belongs to Clk/STY (LAMMER-type) protein kinases involved in regulation of constitutive and AS in plants and animals through phosphorylation and interaction with serine/arginine-rich (SR) proteins (Colwill et al. 1996a,b; Duncan et al. 1997; Golovkin and Reddy 1999; Savaldi-Goldstein et al. 2000, 2003). Extensive regulation of Clk/STY kinase genes by usage of alternative promoters and/or AS has been shown in animals (Duncan et al. 1995, 1997; Kpebe and Rabinow 2008), but little is known about regulation of these genes in plants except that two splice variants of *AFC2* have been reported in TAIR.

Our RNA-seq data detected the two splice variants annotated in TAIR and a total of 10 new putative assembled transcripts with an RPKM (reads per kilobase of exon model per million mapped reads) value ≥ 1 for *AFC2* (Fig. 6A). By comparison to the TAIR reference model (AT4G24740.1), the PATs showed retention of introns 1, 2, 4, 5, 4 + 5, 6, 9, and 10, skipping of exons 2, 3, and 5 + 6, and the use of an alternative 5′ splice site in exon 2 and an alternative 3′ splice site in exon 4. We have validated experimentally the majority of the novel AS events and assembly of the PATs by the HR RT-PCR panel and Sanger sequencing of the RT-PCR products using two primer pairs designed to exons 1 and 4 (primer pair 227) and to exons 4 and 7 (primer pair 226) (Supplemental Table S4). Products corresponding to skipping of exon 3 (PAT3, 278 bp), retention of intron 1 (PAT1, 960 bp, product size is outside the range of HR RT-PCR), and retention of intron 5 (PAT9) were not observed. Retention of intron 4 (PAT10, 460 bp) or intron 6 (PAT8, 418 bp) products are barely detectable by the HR RT-PCR panel (Fig. 6A,B, upper panel), in agreement with RPKM values of the intron retention events (PAT7-10) being much lower than for the two skipped exons 5 and 6 (PAT4-6) or fully spliced PATs (data not shown).

To examine the consequences of the AS events in *AFC2,* we have predicted the putative proteins, taking into consideration alternative transcription start sites (TSSs) and two possible translation initiation sites (TISs) reported in TAIR9 (Fig. 6A). The full-length protein of 427 amino acids containing the N-terminal noncatalytic domain followed by the catalytic kinase domain is encoded by AT424740.1. Translation of other PATs results in either N-terminally or C-terminally truncated proteins which lack either the putative nuclear localization signal encoded by the first exon or the highly conserved LAMMER motif, respectively (Fig. 6A). Interestingly, protein alignment shows that the alternatively spliced exon 2 of *AFC2* corresponds to exon B which undergoes skipping in mammalian orthologs (Duncan et al. 1995; data not shown). It remains to be seen whether *AFC2* putative protein isoforms are, indeed, produced and what are the functional consequences of these variations.

The majority of the AS events in *AFC2* generate premature termination codons (PTCs) which are followed by downstream splice junctions (Fig. 6A) and may, therefore, be turned over by nonsense-mediated decay (NMD). We have analyzed the abundance of these PTC+ transcripts in *upf1-5* and *upf3-1* mutants that are impaired in the NMD pathway using the HR RT-PCR panel. The transcripts with skipped exons 5 and 6 showed a significant increase in the NMD mutants compared to the wild-type control, in contrast to PTC+ PATs with intron retention events (Fig. 6B). These results are in agreement with our recent data showing that intron retention containing transcripts are not targeted to the NMD pathway (Kalyna et al. 2011). Interestingly, intron retention transcripts have also been identified for the orthologous murine *Sty* gene and were shown to be retained in the nucleus (Duncan et al. 1995) and therefore possibly avoid NMD. This example shows that identification of previously unknown AS transcripts of *AFC2* by our RNA-seq experiment can provide insight into the regulation of this important gene and establish the basis for future experimental validation.

Discussion

Our RNA-seq analysis provides the most comprehensive information on alternatively spliced transcripts in *Arabidopsis* to date and will be of great value in addressing regulation of expression and gene/protein function. We have performed a genome-wide investigation of alternative splicing in *Arabidopsis* by RNA-seq and have demonstrated that over 61% of intron-containing genes are alternatively spliced. The significantly increased occurrence of AS found in this study is due to the use of a normalized cDNA library combined with 75-nt paired-end sequencing reads. The normalization process led to enhanced gene representation due to a significant reduction of read coverage of highly expressed genes, and the longer reads improved splice junction detection (Supplemental Tables S1, S2). The analysis detected almost 150,000 splice junctions, of which 31% were novel. The majority of these splice junctions define introns with features of typical plant introns, providing confidence in the splice junctions identified. The high proportion of splice junctions predicting high confidence introns reflects the filtering parameters based on experimental evidence that we have used in this study. In addition, we identified about eight times more (2069) U12 introns than previously found (165 and 246, Zhu and Brendel 2003, and Alioto 2007, respectively). This comprehensive set of U12 introns will stimulate research of the significance and evolution of minor spliceosomal introns.

Estimates of the number of genes showing AS were derived in two ways. First, the analysis based on overlapping introns which did not take intron retention events into account showed that around 52% of intron-containing genes were alternatively spliced. Second, putative transcripts were assembled from exons (defined by reads) and the splice junctions that were filtered on the basis of read coverage. This approach now including intron retention events indicates that 61.2% of intron-containing genes undergo AS. Although this number is significantly higher than previously reported, we expect that it still represents an underestimate because we used only flower and seedling material grown under normal conditions. Therefore, we have probably not captured AS events specific for other organs, developmental stages, and environmental conditions (abiotic and biotic stresses). Indeed, recent RNA-seq analysis showed that many AS events are stress-associated (Filichkin et al. 2010).

Assembly of transcripts from short sequencing reads remains a computational challenge. It is, therefore, essential to obtain an experimental support for an assembly procedure. We have taken advantage of a high-resolution alternative splicing RT-PCR panel to validate a large set of AS events and transcripts. This method

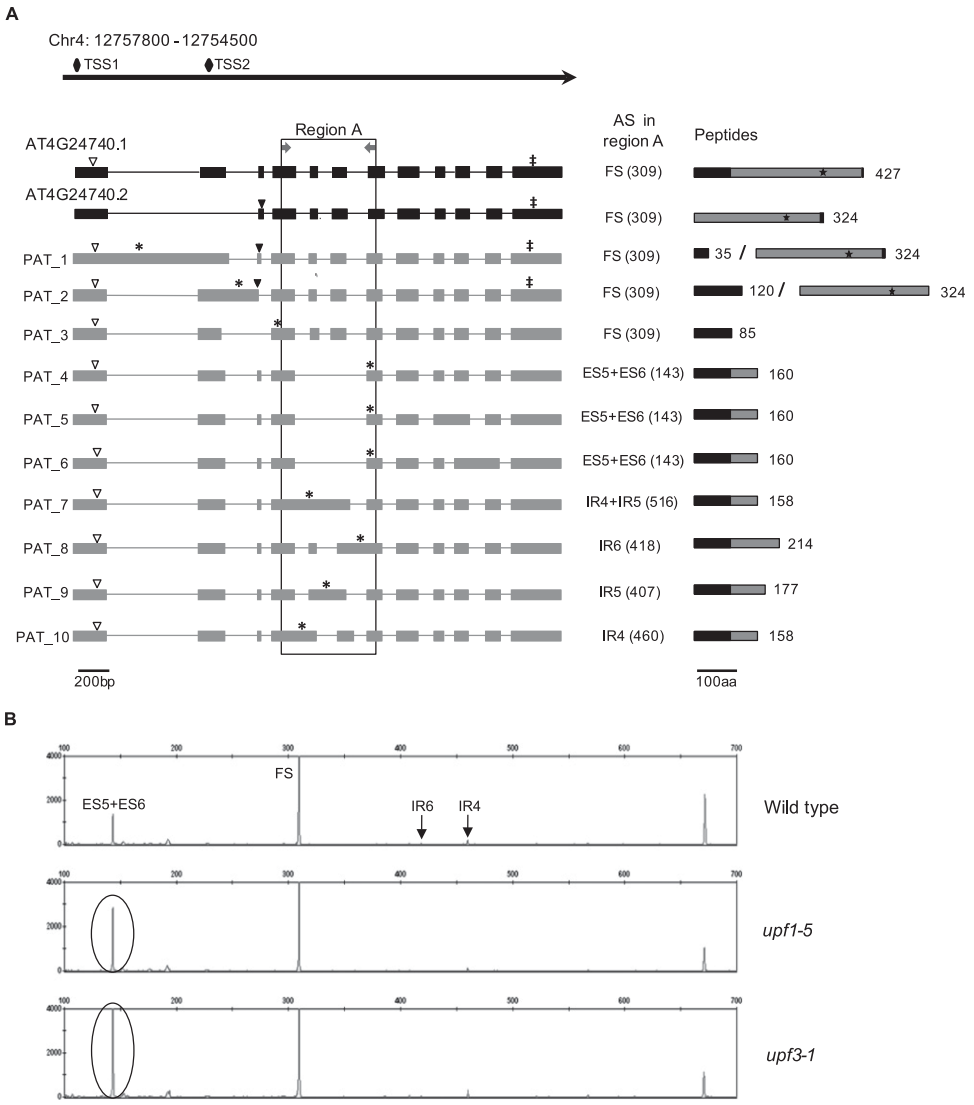


Figure 6. Alternative splicing in the *AFC2* gene. (*A*) The black arrow represents the chromosome location of *AFC2* and its orientation indicates the direction of transcription. (Diamonds) Two transcription start sites (TSS) reported by TAIR. (Black) TAIR models for the *AFC2* gene; (gray) new putative assembled transcripts (PATs). The triangles located over the TAIR transcripts and PATs indicate the start codons, while the asterisks or double daggers represent the stop codons. For PAT1 and PAT2, two putative start codons (full and empty triangles) and two putative stop codons (asterisk and double dagger) are depicted. Protein isoforms predicted using the shown start and stop codons are illustrated at the end of each splice variant followed by their sizes in amino acids. (Black) Noncatalytic domain; (gray) kinase catalytic domain. (Star) Location of the LAMMER motif. (Arrows) The primer pair 226 designed for the HR RT-PCR panel. The amplified region is highlighted by a box (region A). The AS events in the region A are shown after each transcript, and the number in parentheses denotes the size of the amplified product in nucleotides. Abbreviations for AS events: (FS) fully spliced, (IRn) intron retention where "n" is the number of the intron, (ESn) exon skipping of exon number "n." (*B*) Electropherograms of RT-PCR products for the primer pair indicated in *A* for wild type and *upf* mutants generated by GeneMapper. Numbers on the *x*-axis represent the size markers in bp; numbers on the *y*-axis represent relative fluorescence, reflecting transcript abundance. The AS event associated with each product is shown in the respective peak. (Ovals) The peaks which increase in abundance in the *upf1-5* and *upf3-1* mutants.

(and Sanger sequencing, when necessary) achieved an excellent level of validation, as we obtained experimental support for 535 of 586 transcripts predicted by RNA-seq for the regions covered on the HR RT-PCR panel. On the other hand, a considerable number of the HR RT-PCR panel transcripts were not identified in this RNA-seq sample (Fig. 4B). This is most probably due to the AS information on the HR RT-PCR panel being derived from a pool of different plant material and growth conditions. In addition, some of the AS transcripts unique to the HR RT-PCR panel were, in fact, explained by RNA-seq splice junctions which were lost due to stringent filtering parameters used in the assembly process. The additional transcripts on the HR RT-PCR panel also suggest that our estimate of the occurrence of AS is quite conservative.

Our transcript assemblies allowed us to identify the frequency of different types of AS events. In particular, we observed a complex alternative splicing landscape where multiple splicing events occurred within the same transcript, raising the possibility that these events are either independent or coordinated. In our analysis, intron retention was the most prominent event, and we conducted a genome-wide analysis of the features of retained introns from our RNA-seq data. In contrast to what was observed in animals (Stamm et al. 2000; Galante et al. 2004; Sugnet et al. 2004;

Zheng et al. 2005; Sakabe and de Souza 2007) and suggested for plant introns (Wang and Brendel 2006), we did not detect a correlation between the retention of an intron and its size. Therefore, in *Arabidopsis*, intron size cannot be used to predict whether an intron will be retained or not. However, introns alternatively spliced by other mechanisms than intron retention were considerably larger than the average intron size.

To further analyze intron retention, we developed an approach to indicate the potential for introns to be retained (the intron retention ratio). The use of IRR allowed us to detect many introns which have a high level of retention and a high GC content. Further examination led to the identification of a set of introns (1289) located within annotated coding exons which we define as cryptic introns. Interestingly, 602 of these cryptic introns are excised in frame and, therefore, would generate protein isoforms by removing a stretch of amino acids within a protein. We also identified a set of introns with low IRRs, many of which had low read coverage. By analyzing all of the intron retention events contained on the HR RT-PCR panel, we showed that ~20% were not detectable by RT-PCR or RNA-seq and a further 20% were barely detectable on the panel but were supported by RNA-seq at low read coverage. These low abundance IR events which are annotated in databases potentially could represent incompletely spliced mRNAs or DNA artifacts in cDNA/EST preparation. The impact of intron retention on alternative splicing in *Arabidopsis* in general is illustrated by the fact that, although IRs represent 40% of AS events, only 23.6% of the assembled transcripts contain one or more retained introns. Indeed, if these transcripts are discounted, 51.3% of intron-containing genes have an alternative transcript which does not include intron retention. This agrees well with our estimate based on overlapping introns from our set of predicted splice junctions (51.8%). Therefore, while intron retention is still the most abundant AS event in plants (~40%), the significance of intron retention for alternative splicing in plants is obviously much less than has been thought previously.

Alternative splicing has recently caught the attention of many plant researchers, as many developmental processes and response to environmental cues have been shown to be controlled by AS (for review, see Ali and Reddy 2008; Terzi and Simpson 2008; Petrillo et al. 2011). An essential level of regulation of pre-mRNA splicing is achieved through phosphorylation of SR proteins (Cao et al. 1997). To illustrate the potential application of our RNA-seq data, we have examined the AS of *AFC2*, coding for a highly conserved Clk/STY (LAMMER-type) kinase, involved in modulation of pre-mRNA splicing through the phosphorylation and interaction with the SR proteins and other splicing factors (Golovkin and Reddy 1999; Savaldi-Goldstein et al. 2000, 2003). In animal systems, alternative transcripts of Clk kinases generated through exon skipping contain PTCs (Duncan et al. 1995, 1997) and are probably degraded by NMD (Hillman et al. 2004). We have demonstrated that AS coupled to NMD plays an important role in the regulation of *AFC2* which, in turn, modulates splicing factors like SR proteins, suggesting a regulatory loop that results in a fine-tuned regulation of AS. Importantly, NMD affects levels of *AFC2* PTC+ transcripts generated through exon skipping; however, no effect was observed on PTC+ splice variants generated by intron retention. This observation is in agreement with our previous findings showing that, although many intron retention transcripts possess NMD features (such as PTCs followed by downstream splice junctions), they are not sensitive to NMD (Kalyna et al. 2011). Therefore, a PTC created by AS is not necessarily a signature to trigger NMD. Furthermore, AS events causing NMD might occur elsewhere in a transcript, making it necessary to study AS linked to NMD at the level of full-length transcripts in order to predict their fates. This will be greatly aided by the advent of high-throughput sequencing technologies providing longer reads.

In summary, we provide a high confidence set of AS events in *Arabidopsis* on a genome-wide level and, therefore, offer a valuable resource in understanding gene regulation. The question remains to what extent the extensive AS described here contributes to regulation of expression and proteome diversity in plants. Our data will be extremely valuable in addressing these aspects of the biological functions of AS. In addition, combining analyses of ecotype genetic variation and AS profiles would add tremendously in dissecting phenotypic features and mechanisms underlying the plasticity of plant development and stress response.

Methods

cDNA library preparation and high-throughput sequencing

Total RNA was extracted using RNeasy Mini Kit (Qiagen) from wild-type *Arabidopsis thaliana* (Col-0) flowers of different stages and 10-d-old seedlings grown at 22°C, 60% humidity, light intensity of 150 μmol m^{-2} s^{-1}, and 16-h light/8-h dark cycle. Total RNA samples from seedlings and flowers were mixed in a 1:1 ratio for the synthesis of the cDNA library. A full-length oligo-dT cDNA library was constructed using a Mint-Universal Kit (Evrogen), followed by cDNA library normalization using a Trimmer-direct Kit (Evrogen). The resulting library was fractionated using a Covaris sonicator (cycle 15%, intensity 6.0, cycles burst: 250), and fragments in a size range of 200–800 nt were purified using a QIAquick Gel Extraction Kit (Qiagen). Selected products were amplified by 18 PCR cycles. For high-throughput sequencing, the library was prepared according to the manufacturer's instructions and submitted to five lanes of 75-nt paired-end sequencing using the Illumina GA system.

Read alignment to the reference *A. thaliana* genome

A total of 115,883,414 paired-end reads were obtained from five lanes of Illumina sequencing. Read mapping to the *Arabidopsis* reference genome TAIR9 (The *Arabidopsis* Genome Initiative 2000), and splice junction detection was done using TopHat (Trapnell et al. 2009) with a maximum of two mismatches. Minimum and maximum intron lengths were fixed at 60 and 6000 nt, respectively. The rest of the parameters were left as default. A Pearson correlation coefficient between the sequencing lanes was calculated using the coverage values obtained by TopHat (Trapnell et al. 2009). As a good correlation between the replicates was obtained ($R^2 > 0.98$), the pool of sequences from the five lanes was used in further analyses.

To examine the influence of read-length and normalization on the read mapping, we used publicly available *Arabidopsis* Illumina data of 36-nt read length (Filichkin et al. 2010). Two libraries were analyzed: an oligo-dT cDNA library from 3-wk-old wild-type plants and one random primed (RP) cDNA library from 12-d-old seedlings (SRA files: SRX006704, SRX006692, SRX006192, SRX006191). To generate 36-nt reads for our data, we extracted the first 36 nt from the 75-nt original reads using homemade Perl scripts. For the single-end reads sets, we used only the left-side reads coming from the original pair-end set. All the sets were aligned with TopHat using the same intron size parameters as described above.

Read coverage along transcription units was obtained using annotated gene models in TAIR 9 (The *Arabidopsis* Genome Initiative 2000). If more than two gene models were annotated for a particular gene, then we chose the longest model as a representative of the gene. The start and the end of each transcription unit were normalized from 1% to 100%, respectively.

Splice junction detection

We used TopHat software (Trapnell et al. 2009) to predict splice junctions. To diminish the number of potential false positives, we performed a prediction of splice junctions without mismatches in addition to the original alignment of two mismatches (Supplemental Fig. S1). The above approach allowed us to discard splice junctions that are predicted by erroneous alignment, especially in regions where several dinucleotides resembling splice sites are found in tandem. We have filtered out those splice junctions that were not predicted in the alignment of no mismatches and that are in proximity with another better supported junction by <10 nt. The remaining splice junctions were used for further analysis if they were supported by at least three alignments in the original list (alignment with two mismatches).

We have defined as genic splice junctions those junctions that are inside the coordinates of annotated genes in TAIR9 and have the same strand as the gene in question. Those splice junctions that are inside gene coordinates but are in the opposite strand were named antisense. Finally, those junctions that are not inside any gene coordinates were called intergenic.

For the genic splice junctions, we took those coming from protein-coding genes and have examined whether the predicted splice junctions are inside the coding sequence, 5′ UTR or 3′ UTR. If one extreme of the junction was in the coding sequence and the other in the UTR, they were classified as 5′UTR-CDS or CDS-3′UTR.

To determine the signatures of U2 or U12 introns, we used 10 intronic and three exonic bases for the 5′ splice site and 14 intronic bases and three exonic bases for the 3′ splice site. We have evaluated the splice sites using the procedure and PWMs described by Sheth et al. (2006). A strong signal of U2 or U12 intron was considered if the splice site in question (5′ or 3′) score was ≥65 in the respective PWM. For the U12 branch point, a PWM was built using the branch point sequences deposited in the U12 database (Alioto 2007). The procedure of building the PWM was the same as in Sheth et al. (2006). To identify the branch point, we scanned the first 30 bases upstream of the 3′ splice site and evaluated this stretch of sequence with the U12 PWMs. The higher score was retained, and, if this score was ≥65, then it was considered to be a strong signal for a U12 branch point. To define U2 branch points, we have looked for the motif YURAY in the stretch of 30 nt. As U12 introns have more highly conserved 5′ splice site and branch point sequences than U2 introns, they were classified as U12 when they possessed both 5′ splice site and branch point with scores ≥ 65 in the respective U12 PWMs. An intron was considered as U2 if it had a good U2 signal in the 5′ splice site (score ≥ 65 in the respective U2 PWMs) and a good signature in the 3′ splice site (score ≥ 65 in the respective U2 PWMs) or the YURAY U2 branch point.

Gene ontology classification of the genes containing U12 introns was done using Classification SuperViewer Tool (Provart and Zhu 2003) available at: http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi.

Assembly of putative transcripts

To build the possible set of transcripts that can be explained by our reads, we first started by defining exons inside the longest transcription unit for each annotated gene in TAIR9. A putative exon was defined as a continuous stretch of at least four reads of coverage inside the transcription unit. If the gene in question had predicted junctions in the set of filtered junctions (score ≥ 3) (Table 1), then we used these junctions to connect the putative exons. All the possible combinations of exons that can be connected through the predicted junctions were built for each gene. The final set of transcripts was evaluated using Cufflinks v0.9.3. Those transcripts that passed the expression filter used by Cufflinks default settings (transcripts with <15% of expression level according to the highest expressed transcript in each gene are discarded) were used for the further analyses. To estimate how the above method is able to retrieve known transcripts, we have compared our set of transcripts with those annotated transcripts in TAIR9. If one putative transcript has the same combination of introns as in an annotated transcript and it covers at least 80% of its length contiguously, then it was considered that a known transcript was obtained for that particular gene.

To quantify the types of AS events in our final set of predicted transcripts, we have used ASTALAVISTA software (Foissac and Sammeth 2007).

Validation of putative assembled transcripts using the HR RT-PCR panel

For corroboration of our putative assembled transcripts, we have used the HR RT-PCR Alternative Splicing Panel (Simpson et al. 2008a,b). The 273 primer pairs were designed to cover AS events previously reported in 256 genes. The primer sequences used in the analyses can be found in Supplemental Table S3. We extracted the sequences defined by each pair of primers using our PATs and transcripts annotated in TAIR9 database. For each sequence, we have calculated its size. Fragments between 100 and 650 nt were used for comparison due to the size marker used in HR-RT PCR panel. A PAT fragment was considered to match between the two resources if their fragment sizes are exactly the same or have a similar size with a deviation of no more than 4 nt. Additionally, Sanger sequences obtained from fragments of the HR RT-PCR panel have been used to corroborate some of the putative transcripts obtained in this study.

Intron retention analysis

The intron retention events were obtained from our set of PATs. Thus, the retained introns in our analyses are completely covered by RNA-seq reads and have passed the expression filtering by Cufflinks. The intron retention ratio was calculated using the median of reads that aligned inside the intron divided by the number of reads supporting the splice junction. The median was used as the parameter of intron coverage because reads along the introns are not uniformly distributed and this measure is less sensitive to extreme values. Additionally, we have calculated the IRR by counting the reads containing the splice site and adjacent sequence of the retained intron divided by the reads supporting the splice junction. Both methods of IRR calculation gave a Pearson correlation of ~0.9 (data not shown). The further analyses were performed using the first method (median of reads along the intron). The differences between groups of intron features (intron size, GC content, splice site score, and number of stop codons) were evaluated using Mann-Whitney-Wilcoxon tests in groups of introns with the same sample size (Fig. 5).

We analyzed the coding potential of retained introns, which are located inside annotated coding sequences. Translation was performed using the frame of the 5′ neighboring exon annotated in TAIR. For those introns that are inside annotated exons, we have used the frame of the exon in which they are contained. As a negative control, we have used the same number of randomly selected introns inside coding sequences that have no evidence of retention in our sample, and the frame for their translation was obtained as for the retained introns. Furthermore, introns inside UTRs translated in the three frames were also used as a negative control. The resulting translated sequences (exon+intron+exon) were used to search Pfam protein domains employing hidden Markov models (HMMs) with HMMER (Pfam HMM database

release 25.0). Hits were only regarded as significant if they had an E-value ≤ 0.01. An intron was considered to be part of a protein domain if it contributes at least to 20% of the total domain.

The expected frequency of stop codons in introns was calculated by f(T) × f(A) × f(A) + f(T) × f(G) × f(A) + f(T) × f(A) × f(G), where f(A) represents the frequency of the nucleotide A. We obtained the codon usage of the exons and introns as described by Galante et al. (2004). We performed 100 pairwise comparisons between random distributions of 435,283 codons present in the retained introns along the 61 possible codons to obtain the average χ^2 and the standard deviations (Supplemental Table S6).

For testing of the cryptic introns by RT-PCR, we designed primer pairs for 10 randomly selected introns with an IRR > 1 (Supplemental Table S10). The cDNA employed for the experiment was synthesized from a 1:1 mix of total RNA from wild-type flowers and 10-d-old seedlings using the Reverse Transcription System kit (Promega). PCR was performed under the following conditions: denaturation 98°C for 3 min, followed by 30 cycles of denaturation 98°C for 10 sec, annealing 58°C for 10 sec, and extension 72°C for 30 sec, and final extension at 72°C for 5 min. The amplified products were visualized in a 2% agarose gel stained with ethidium bromide (Supplemental Fig. S8).

Data access

Illumina short read sequences generated in this study have been submitted to the NCBI Sequence Read Archive (SRA) (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession number SRA047499. Splice junctions, U12 introns, and putative assembled transcripts are available in the Supplemental Material.

Acknowledgments

We thank Andreas Sommer from the Vienna Biocenter CSF-sequencing facility for sequencing, the Center of Integrative Bioinformatics Vienna (CIBIV) for use of cluster facilities, and Gautier Koscielny from the EBI Hinxton for advice to Y.M. in high-throughput sequencing data management. This work was supported by the Austrian Science Fund (FWF) (DK W1207; ERA-NET Plant Genomics [PASAS] I254; SFB RNAreg F43-P10) and the Austria Genomic Program (GENAU III) (ncRNAs), the EU FP6 Programme Network of Excellence on Alternative Splicing (EURASNET) (LSHG-CT-2005-518238), the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/G024979/1, ERA-NET Plant Genomics [PASAS]), and the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD) (WP114).

References

Ali GS, Reddy AS. 2008. Regulation of alternative splicing of pre-mRNAs by stresses. *Curr Top Microbiol Immunol* **326:** 257–275.
Alioto TS. 2007. U12DB: A database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* **35:** D110–D115.
The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.
Barbazuk WB, Fu Y, McGinnis KM. 2008. Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Res* **18:** 1381–1392.
Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci* **74:** 3171–3175.
Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72:** 291–336.
Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. *Nat Genet* **30:** 29–30.
Brown JW, Simpson CG. 1998. Splice site selection in plant pre-mRNA splicing. *Annu Rev Plant Physiol Plant Mol Biol* **49:** 77–95.
Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* **7:** 327. doi: 10.1186/1471-2164-7-327.
Cao W, Jamison SF, Garcia-Blanco MA. 1997. Both phosphorylation and dephosphorylation of ASF/SF2 are required for pre-mRNA splicing in vitro. *RNA* **3:** 1456–1467.
Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12:** 1–8.
Colwill K, Feng LL, Yeakley JM, Gish GD, Caceres JF, Pawson T, Fu XD. 1996a. SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *J Biol Chem* **271:** 24569–24575.
Colwill K, Pawson T, Andrews B, Prasad J, Manley JL, Bell JC, Duncan PI. 1996b. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intracellular distribution. *EMBO J* **15:** 265–275.
Cusack BP, Wolfe KH. 2005. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol Biol Evol* **22:** 2198–2208.
Duncan PI, Howell BW, Marius RM, Drmanic S, Douville EM, Bell JC. 1995. Alternative splicing of STY, a nuclear dual specificity kinase. *J Biol Chem* **270:** 21524–21531.
Duncan PI, Stojdl DF, Marius RM, Bell JC. 1997. In vivo regulation of alternative pre-mRNA splicing by the Clk1 protein kinase. *Mol Cell Biol* **17:** 5996–6001.
Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. 2010. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* **20:** 45–58.
Foissac S, Sammeth M. 2007. ASTALAVISTA: Dynamic and flexible analysis of alternative splicing events in custom datasets. *Nucleic Acids Res* **35:** W297–W299.
Galante P, Sakabe N, Kirschbaum-Slager N, de Souza S. 2004. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10:** 757–765.
Golovkin M, Reddy AS. 1999. An SC35-like protein and a novel serine/arginine-rich protein interact with *Arabidopsis* U1-70K protein. *J Biol Chem* **274:** 36428–36438.
Goodall GJ, Filipowicz W. 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* **58:** 473–483.
Goodall GJ, Filipowicz W. 1990. The minimum functional length of pre-mRNA introns in monocots and dicots. *Plant Mol Biol* **14:** 727–733.
Hillman RT, Green RE, Brenner SE. 2004. An unappreciated role for RNA surveillance. *Genome Biol* **5:** R8. doi: 10.1186/gb-2004-5-2-r8.
Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, 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.
Jones-Rhoades MW, Borevitz JO, Preuss D. 2007. Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. *PLoS Genet* **3:** 1848–1861.
Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, Marshall J, Fuller J, Cardle L, McNicol J, et al. 2011. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res*. doi: 10.1093/nar/gkr932.
Kpebe A, Rabinow L. 2008. Alternative promoter usage generates multiple evolutionarily conserved isoforms of *Drosophila* DOA kinase. *Genesis* **46:** 132–143.
Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40:** 225–231.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.
Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W. 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci* **5:** 160–167.
Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, et al. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* **20:** 1238–1249.
Mironov AA, Fickett JW, Gelfand MS. 1999. Frequent alternative splicing of human genes. *Genome Res* **9:** 1288–1293.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Ner-Gaon H, Fluhr R. 2006. Whole-genome microarray in *Arabidopsis* facilitates global analysis of retained introns. *DNA Res* **13:** 111–121.
Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415.

Petrillo E, Sanchez SE, Kornblihtt AR, Yanovsky MJ. 2011. Alternative splicing adds a new loop to the circadian clock. *Commun Integr Biol* **4:** 284–286.
Provart N, Zhu T. 2003. A browser-based Functional Classification SuperViewer for *Arabidopsis* genomics. *Currents in Computational Molecular Biology* **2003:** 271–272.
Rosenfeld MG, Amara SG, Roos BA, Ong ES, Evans RM. 1981. Altered expression of the calcitonin gene associated with RNA polymorphism. *Nature* **290:** 63–65.
Sakabe N, de Souza S. 2007. Sequence features responsible for intron retention in human. *BMC Genomics* **8:** 59. doi: 10.1186/1471-2164-8-59.
Savaldi-Goldstein S, Sessa G, Fluhr R. 2000. The ethylene-inducible PK12 kinase mediates the phosphorylation of SR splicing factors. *Plant J* **21:** 91–96.
Savaldi-Goldstein S, Aviv D, Davydov O, Fluhr R. 2003. Alternative splicing modulation by a LAMMER kinase impinges on developmental and transcriptome expression. *Plant Cell* **15:** 926–938.
Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34:** 3955–3967.
Simpson CG, Fuller J, Maronova M, Kalyna M, Davidson D, McNicol J, Barta A, Brown JW. 2008a. Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J* **53:** 1035–1048.
Simpson CG, Lewandowska D, Fuller J, Maronova M, Kalyna M, Davidson D, McNicol J, Raczynska D, Jarmolowski A, Barta A, et al. 2008b. Alternative splicing in plants. *Biochem Soc Trans* **36:** 508–510.
Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang M. 2000. An alternative-exon database and its statistical analysis. *DNA Cell Biol* **19:** 739–756.
Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344:** 1–20.
Sugnet CW, Kent WJ, Ares M Jr, Haussler D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput* **9:** 66–77.
Terzi LC, Simpson GG. 2008. Regulation of flowering time by RNA processing. *Curr Top Microbiol Immunol* **326:** 201–218.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome: Design principles of a dynamic RNP machine. *Cell* **136:** 701–718.
Wakamatsu A, Kimura K, Yamamoto J, Nishikawa T, Nomura N, Sugano S, Isogai T. 2009. Identification and functional analyses of 11,769 full-length human cDNAs focused on alternative splicing. *DNA Res* **16:** 371–383.
Wang BB, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci* **103:** 7175–7180.
Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144:** 32–42.
Xiao YL, Smith SR, Ishmael N, Redman JC, Kumar N, Monaghan EL, Ayele M, Haas BJ, Wu HC, Town CD. 2005. Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2 reveals numerous transcript variants. *Plant Physiol* **139:** 1323–1337.
Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* **13:** 1290–1300.
Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, et al. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res* **20:** 646–654.
Zheng CL, Fu XD, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11:** 1777–1787.
Zhu W, Brendel V. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res* **31:** 4561–4572.
Zhu W, Schlueter SD, Brendel V. 2003. Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiol* **132:** 469–484.

Received October 28, 2011; accepted in revised form February 23, 2012.

Exitron splicing, a new type of alternative splicing event shaping the eukaryotic proteome

Marquez, Y., Höpfler, Markus., Ayatollahi, Zahra, Barta, A., & Kalyna, M. (2014).

Submitted.

Author contribution: Experiment design, samples preparation and bioinformatic analyses.

Exitron splicing, a New Type of Alternative Splicing Event Shaping the Eukaryotic Proteome

Yamile Marquez, Markus Höpfler[1], Zahra, Ayatollahi, Andrea Barta, Maria Kalyna*

Max F. Perutz Laboratories, Medical University of Vienna, Dr. Bohr-Gasse 9/3, A-1030 Vienna, Austria

[1] Current address: Department of Molecular Cell Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

*Correspondence to: mariya.kalyna@univie.ac.at.

Abstract

Alternative splicing is a key mechanism for increasing transcriptome and proteome diversity. Here, we identify exitron splicing as a novel type of alternative splicing event in plants and humans. Exitrons (**ex**onic **i**n**trons**) are internal regions of coding exons that can be removed by the spliceosome. Intriguingly, intronless genes can be also alternatively spliced *via* exitron usage, providing the first evidence of splicing in these genes. We show that exitron splicing is regulated in a tissue-specific manner, in response to stress and by splicing factors. Exitron-encoded sequences contain protein domains interacting with different molecules and are enriched in disordered regions and post-translational modifications; consequently their splicing impacts proteome dynamics and remodelling. We propose a "splicing memory" mechanism for the origin of exitrons whereby exitrons are derived from ancestral coding exons through a history of intron loss and maintenance of vestigial splicing regulatory elements that drives exitron evolution.

In the majority of eukaryotic genes, the protein coding information of exons is interrupted by introns [1]. Intron removal by the spliceosome relies mainly on the core splicing signals present in every intron: 5' and 3' splice sites (SS) and branch point [2]. However, in *Arabidopsis* and human these signals represent only part of the information required to define introns [3]. Splicing regulatory elements contribute significantly to the recognition of the core splicing signals. Splicing factors (SFs) bound to such elements can affect SS selection resulting in alternatively spliced (AS) transcripts. About 95% and 60% of human and *Arabidopsis* genes are alternatively spliced, respectively [4-6]. The main types of AS events are alternative 5' and 3' SS usage, exon skipping (ES) and intron retention (IR).

Features of exitrons, new players in AS

In a genome-wide analysis of AS in *Arabidopsis thaliana*, we previously identified a class of AS events, which we referred to as cryptic introns [6]. These introns constitute internal parts of the protein coding exons, do not contain stop codons, and possess all the canonical core splicing signals. Here, we name them exitrons (exonic introns, EI) (Fig. 1a). In AS events described to date, AS regions are either flanked by intron(s) (ES, usage of alternative 5' or 3' SS) or are introns themselves (IR). Exitrons are exonic sequences and are directly flanked by exonic sequences. Therefore exitron splicing (EIS) is the only AS event which does not involve canonical introns and this feature sets EIS apart from other types of AS events. Because of its novelty and potential to affect protein diversity we decided to investigate this novel class of AS events. For this purpose, we have defined a high confidence set of 1002 exitrons (Fig. 1b, Supplementary Table 1) by applying more stringent criteria to our RNA-seq data[6].

As exitrons constitute internal regions of exons, they are expected to not only have features of both exons and introns but also to be distinguishable from both of them.

We observed that exitrons are overall longer and have considerably higher GC content than both constitutive and retained introns (Fig. 1c and 1d). The sizes of exitrons are more comparable to the sizes of constitutive exons and exons of the exitron-containing genes than to the sizes of introns (Fig. 1e compare to Fig. 1c). Interestingly, exons that contain exitrons (EI exons) tend to be significantly longer than other groups of exons (Fig. 1e). Exitrons have GC content significantly lower than constitutive exons and, notably, than other exons in exitron-containing genes and EI exons (Fig. 1f) indicating that this is a specific feature of exitrons and not a general property of exitron-containing genes or EI exons. These features clearly distinguish exitrons from both constitutive exons and introns, and, more importantly, from IRs.

In *Arabidopsis*, 1002 exitrons are embedded within coding exons of 892 genes (Fig. 1b, Supplementary Table 1). Intriguingly, 18.9% of exitrons are located in 165 intronless genes (Supplementary Table 1) suggesting EIS as a so far overlooked generator of alternative transcript and protein isoforms from intronless genes. We have validated exitron-spliced isoforms including those in intronless genes by various methods (Extended Data text, Supplementary Table 2, Extended Data Figs. 1-3). Almost half of the exitrons have lengths of multiples of three nucleotides (EI_{x3}) and thus their splicing results in the removal of an internal protein sequence (Fig. 1a and 1b). Splicing of other exitrons changes the reading frame, which either generates an altered C-terminal sequence of the resultant protein or introduces a premature termination codon (PTC) downstream of the splice junction potentially triggering nonsense-mediated RNA decay (NMD) (Fig. 1a). Indeed, PTC+ exitron-spliced transcripts were elevated in the NMD mutants when compared to wild-type plants contrasting EI_{x3}-spliced variants (Extended Data text, Extended Data Fig. 4).

Therefore EIS results in proteome diversity and may regulate transcript and protein abundance.

A gene ontology (GO) classification of the EI_{x3}-containing genes (the set that will certainly result in protein isoforms) revealed an enrichment of gene categories with functions in plant adaptation and development, suggesting an important regulatory role for EIS (Fig. 1g).

AS is regulated in a tissue-specific manner, during development and in response to changing environmental conditions [4,5,7]. Analysis of randomly chosen EI_{x3}-containing genes (Fig. 2a) showed that the ratios of the full-length and exitron-spliced isoforms were significantly different in seedlings and flowers and that some exitron-spliced isoforms were abundant under normal growth conditions, up to 50% of total transcripts (Fig. 2b, Extended Data text, Extended Data Figs. 4a and 4b). This implies tissue-specific regulation of EIS and a role for EIS as a regulatory mechanism for proteome diversity in different tissues. Changes in EIS are also observed under a variety of stresses with methyl jasmonate (involved in plant defence and development) and mannitol (osmotic stress) treatments having the broadest effect (Fig. 2b, Extended Data text, Extended Data Fig. 4c-e). Differential changes under various stresses suggest that regulation of EIS is not the result of a general stress response but rather that it might be specific to certain conditions.

Splicing factors affect exitron splicing

Serine/arginine-rich (SR) proteins are splicing regulators that influence a variety of AS events [8]. Our analysis showed that *Arabidopsis* SR proteins can directly or indirectly affect EIS and that their effect is differential (Fig. 2b, Extended Data text, Extended Data Figs. 4f and 4g).

We further tested the regulation of EIS in the *SUPPRESSOR OF ABI* (*SUA*) mutant background. SUA is a homolog of the mammalian splicing factor RBM5 (LUCA15/G15/H37) [9-11] and is responsible for the inhibition of splicing of a cryptic intron in the *ABSCISIC ACID INSENSITIVE3* (*ABI3*) gene [9]. This cryptic intron is an internal part of a coding exon, thus it qualifies as an exitron. We used two mutant lines of *SUA* [9], *sua-1* and *sua-2*, in two *A. thaliana* ecotypes, Ler-0 and Col-0, respectively, to examine the effect of SUA on EIS. Consistent with the role of SUA in suppressing EIS in *ABI3*, we observed an increase of the exitron-spliced isoform ratio in the mutants (Fig. 2b, Extended Data Figs. 4h and 4i). Interestingly, we observed differential EIS between *Arabidopsis* ecotypes (Fig. 2b, Extended Data Figs. 4h and 4i) due to the presence of SNPs affecting either the 5' SS, or the 3' SS or the branch point (Fig. 2c). Therefore we checked SNP distribution in 82 *A. thaliana* ecotypes (Extended Data text, Supplementary Tables 5-8) and found that genetic variability affects core splicing signals of exitrons suggesting that differences in EIS can potentially contribute to ecotype diversification.

Considering the major effect of SUA on EIS, we checked whether exitrons contain binding motifs of RBM5 [12,13] and found that they are indeed enriched in exitrons (Fig. 2d). Conversely, a *de novo* motif search resulted in two motifs very similar to the RBM5 motifs tested before (Fig. 2d). Enrichment of RBM5 motifs in exitrons, together with the fact that SUA and RBM5 interact with common proteins in plants and animals [9-11], suggests that EIS in animals (see below) can be also regulated by RBM5. Interestingly, RBM5 regulates ES [10,11] but not IR [14] in human, further corroborating the difference between ES and IR and implying similarities between ES and EIS events.

Exitron-spliced isoforms are translated

To validate translation of exitron-spliced transcripts, we searched for peptide support in proteomic datasets [15,16]. We found peptides for EI_{x3} and non-EI_{x3} protein isoforms in eleven genes (Extended Data text, Supplementary Table 3). In total, EI-protein isoforms for 1.8% of all exitron-containing genes are supported by peptides that compares well to the peptides supporting 3.4% AS genes in mouse [17].

We further checked whether the exitron-containing and the exitron-spliced transcripts associate with ribosomes. Since exitrons can be misinterpreted as IRs, we examined published information [18] on the IR transcripts detected on ribosomes. Four transcripts with retained introns within coding regions (AT2G18690, AT2G33340, AT3G13300, and AT4G01690) were found to be associated with ribosomes and interpreted to be IR transcripts exported from the nucleus [18]. However, these transcripts are indeed transcripts with unspliced exitrons (Supplementary Table 1) and not *bona fide* IR transcripts. Importantly, the exitron-spliced isoforms of the AT3G13300 and AT4G07410 genes were also detected on ribosomes. Peptide support of exitron-spliced isoforms in proteomic datasets and their association with ribosomes show that both the exitron-containing and the exitron-spliced isoforms are exported to the cytoplasm and are translated.

Exitron splicing impacts protein function

Splicing of EI_{x3} results in the removal of an internal sequence in a coding exon and thus in a shorter protein isoform (Fig. 1a). Analyses of the protein sequences encoded by the EI_{x3}s showed that EIS influences protein domains, disordered regions, post-translational modification (PTM) sites, transmembrane domains and signal peptides. (Fig. 3a, Supplementary Tables 9, 11-13) indicating that EIS may impact the functional properties of the proteins. Surprisingly, we observed that 36% of exitron boundaries coincide with protein domain borders (Supplementary Table 10). Protein

domains are under-represented in the EI_{x3} set (Fig. 3b), suggesting that many exitron-encoded sequences do not contain structurally defined regions and are disordered. Indeed, we found that disordered regions and short linear motifs (SLiMs) that can undergo disorder-to-order transitions and are important platforms for protein interactions [19,20] are over-represented in the EI_{x3} set (Fig. 3b). Human tissue-specific cassette exons have similar properties, and it has been suggested that these AS events might play a role in defining protein interaction networks [21,22]. PTMs are crucial regulators of protein function affecting their localization, activity or affinity to other proteins. Analyses of published, experimentally validated, PTM sets revealed that EI_{x3}–encoded sequences carry sites for various PTMs including phosphorylation, sumoylation, ubiquitylation, S-nitrosylation, and lysine acetylation (Supplementary Table 11). Phosphopeptides are enriched in the sequences encoded by exitrons when compared to constitutive and alternative exons (11.2% vs 4.1% and 4.9%, respectively; p-value <0.001). Importantly, this is a feature of exitrons themselves rather than of the proteins encoded by exitron-containing genes (11.2% vs 5.8% of phospho-PTMs are in EI_{x3}s in comparison to other exons of exitron-containing genes, respectively; p-value <0.001). Interestingly, our analysis revealed the first examples of AS resulting in protein isoforms that can have differential ubiquitylation and sumoylation states. Presence of protein domains involved in interaction with target molecules (such as kinase, nucleotide-binding, leucine-rich-repeat and pentatricopeptide repeat (PPR) domains, Supplementary Table 9) and enrichment of disordered regions and PTMs in exitron-encoded sequences indicate that EIS may impact the dynamics and remodelling of the *Arabidopsis* proteome.

The impact of EIS on a protein domain is illustrated by the EI_{x3} in the gene AT1G54270, which codes for the eukaryotic translation initiation factor 4A (eIF4A),

a DEAD-box RNA helicase (Fig. 3c). EIS removes the highly conserved ATP binding motif (Walker A or motif I) together with two conserved phosphorylation sites (Fig. 3c and Extended Data Fig. 4). Unwinding of the substrate by this RNA helicase is ATP dependent [23], implying that EIS affects translation initiation process. This splicing event is supported by ESTs in *A. thaliana*, *Brassica napus* and *Raphanus raphanistrum* (Extended Data Fig. 4, Supplementary Table 2). Moreover, an EST for the human *eIF4A-1* shows EIS at the same position as in *Arabidopsis* (Extended Data Fig. 4). This high conservation suggests that EIS is important for the function of eIF4A and was probably present before the separation of plants and animals.

Exitron splicing is a general AS event

The high conservation of EIS in the *eIF4A* gene prompted us to estimate conservation of EIS events between *A. thaliana* and other plant species. Therefore, we produced a confident set of orthologous gene pairs in eight different plant species (for numbers of gene pairs see Fig. 5a) and tested it against EST collections. EIS in 63 genes is conserved in at least two species (Supplementary Table 14). Finding conserved EIS events depends on the level of EST coverage and will be improved with the availability of deeper transcriptome data. Conversely, EIS can be species-specific and, similarly to cassette exons [24], can provide another potential source for adaptation and speciation.

EIS in human *eIF4A* demonstrates that these events are not restricted to plants. Literature search revealed that IRs in the mammalian-specific genes *CCK2R*, *CD55*, *FMNL1* and *TGIF2* qualify as exitrons: they do not contain stop codons, code for important parts of the proteins, and their splicing results in changes of protein properties (Supplementary Table 15). Moreover, cases of splicing of intraexonic

sequences were described in *Caenorhabditis elegans*, and a hypothesis on their origin has been proposed [25] (see below).

To obtain further evidence of EIS in non-plant species, we analyzed the set of annotated IR transcripts in human Ensembl. Out of this set 1022 retained introns (in 827 genes) qualified as exitrons (Fig. 4a, Supplementary Table 16). EIS in *FMNL1* (Supplementary Table 15) is detected in this analysis, however, not in *CCK2R* and *CD55,* suggesting that this set is not exhaustive and more exitrons are present in human. Almost half of human exitrons are EI_{x3} (Fig. 4a, Supplementary Table 16). GO classification of EI_{x3}-containing genes showed enrichment in genes with important functions including DNA replication, immune response and coding for components of mediator or calcium channel complexes (Supplementary Table 17). Human and *Arabidopsis* exitrons share similar features, such as higher GC content in comparison to retained introns (Fig 4b), similar size distribution (Extended Data Fig. 6), whereby their sizes are closer to exons than to retained introns (Fig. 4c and 4d). In both species, EI exons are considerably longer than other exons (Fig. 4d and Fig 1e). Analysis of the proteins sequences encoded by the human EI_{x3} showed the same (but more striking) tendency of under-representation of protein domains and enrichment in disordered regions and SLiMs (Fig. 4e-g, Supplementary Table 18), suggesting similar functional implications for this type of AS event in both organisms. Similar to *Arabidopsis*, splicing boundaries of about one third of human exitrons coincide with borders of protein domains (Supplementary Table 19). Future analyses of human exitrons using RNA-Seq data sets will further refine their features and comparison to *Arabidopsis* exitrons. These findings illustrate that EIS is conserved, general type of AS event with similar features in plants and human.

Evolution of exitron splicing

Our finding of EIS raises the question of exitron origin. Two scenarios are the most likely generators of exitrons. The first is that an exitron is an ancient intron, which acquired its coding potential through substitutions while maintaining the original splicing signals. The opposite scenario is that an exitron is the product of an ancestral coding sequence that acquired the capacity to be spliced out. In the latter case, although substitutions would occur to generate splicing signals, exitron would be under selection pressure to prevent substitutions that disrupt its coding capacity. If an exitron is a retained intron which acquired coding capacity, then it would be a canonical, non-coding intron in orthologous genes. Alternatively, it would correspond to a coding region. We show that almost all *Arabidopsis* exitrons align to exonic coding sequences in other species (Fig. 5a, Supplementary Table 20). Up to 3% of exitrons align to introns that are not protein-coding. They may have lost their coding capacity in the modern species or they may represent evidence for the first scenario. The origin of exitrons from exonic coding sequences is further corroborated by a prevalence of substitutions at the third positions of codons as revealed by analyses of the exitron regions in 82 *Arabidopsis* ecotypes (Extended Data text) and of exitrons that reside in genes with paralogous copies (Extended Data Fig. 7). This contrasts the pattern usually found in intronic sequences. Importantly, substitutions in exitrons, when compared to paralogues, increase splice site scores and lower the GC content (Supplementary Table 21). Altogether, these results show that exitrons originate from ancestral exonic coding sequences that have acquired the capacity of being spliced out.

We next asked how EIS has evolved. Several observations led us to think that intron loss could play a role during evolution of EIS. Firstly, *Arabidopsis* and human exitron

SS boundaries often coincide with protein domain borders (Supplementary Tables 10 and 19) reminiscent of the strong correlation observed between the borders of exons and protein domains[26]. Secondly, EI exons are longer than other exons (Fig. 1e and 4e). Thirdly, 18.9% *Arabidopsis* exitrons reside in annotated intronless genes, including 29 *PPR* genes (Supplementary Table 1). *PPR* genes rarely contain introns in flowering plants, but are intron-rich in moss (primitive plant) providing evidence that retroposition and intron loss are responsible for their evolution [27]. Therefore we examined the structures of paralogues and orthologues of exitron-containing genes in different plant species. Indeed, introns are present in the regions homologous to exitrons (Supplementary Tables 22 and 23), some located closely or exactly at one or both of the exitron borders (Supplementary Table 22), whereby an exitron can correspond to a single or multiple ancestral exons (Extended Data Fig. 8). Intron loss during evolution of EIS in plants is illustrated by a gene coding for the highly conserved T-protein, a component of the glycine cleavage system [28,29]. Structures of orthologues demonstrate multiple intron losses and that the exitron was an exon of the same size precisely bordered by canonical introns in green algae (Fig. 5c and Extended Data Fig. 10a). Synonymous substitutions favoured the appearance of both SS in the *Arabidopsis* genus (Extended Data Fig. 10b). To obtain evidence for a role of intron loss in human, we analysed the intronless gene *HSPA1A* (*HSP70A-1*) [30]. EIS removes the first of two subdomains of the ATPase domain of this protein (Fig. 5b). Comparison of orthologous genes shows that the exitron corresponds to three exons in *Ciona* (Fig. 5b, Extended Data Fig. 9a). Interestingly, GT and AG dinucleotides are already present at the corresponding exon borders in *Ciona* (Extended Data Fig. 9b). These findings indicate that the evolution of EIS involved intron loss both in *Arabidopsis* and human.

Analysis of the exitron in T-protein gene resulted in two further interesting findings. Firstly, highly conserved short sequences with avoidance of synonymous substitutions are present close to the exitron of T-protein gene (Extended Data Fig. 10b) suggesting functions beyond protein coding. As exonic splicing regulatory elements are under purifying selection [31], we searched for SF binding sites. We found that these sequences correspond to potential binding sites of RBM5/SUA and ETR3/CELF2/BRUNOL3 (Extended Data Fig. 10b). BRUNO-like RNA-binding proteins are involved in multiple aspects of RNA processing including alternative splicing, they are present in animals and plants and their binding specificity is conserved from human to *Arabidopsis* [32,33]. Secondly, the region affected by EIS in *Arabidopsis* overlaps with the conserved in animals ES event in the orthologous *AMT* gene encoding the T-protein (Fig. 5c, Extended Data Fig. 10a and 10c). Both AS events remove the region involved in the protein enzymatic activity, thus resulting in a very similar functional outcome. Taken together, ES events generating two protein isoforms possibly existed in the ancestral gene. Upon intron loss in plants, changes occurred at synonymous sites with some regions being under purifying selection due to the presence of binding sites for RNA splicing/processing factors with functions beyond AS. Selection for the protein isoforms contributed to the maintenance of their binding sites and the emergence of core splicing signals restoring production of the two splice variants *via* EIS.

It has been proposed that mutations in protein coding sequences creating a PTC would promote intronization of the affected region to rescue at least a shortened ORF [34]. Exitron evolution must have proceeded differently as exitrons do not contain PTCs. On the contrary, splicing of non-EI_{x3} exitrons may actually result in PTCs, albeit downstream of exitron splice junctions (Fig. 1a). In *C. elegans*, another hypothesis for

intronization of exonic sequences has been proposed whereby GT/C and AG splicing boundaries are created by single substitutions [25]. Though these dinucleotides are required at SS, they are not sufficient to act as splicing signals. Moreover, we showed that they can be already present in the ancestral sequences (see *Ciona HSPA1A* orthologues, Extended Data Fig. 9b). Furthermore, this scenario is unlikely in *Arabidopsis* and human where core splicing signals contribute only about half of the information required to define introns [3].

Based on our findings we propose the hypothesis where genes, upon intron loss, "remember" the way they were originally spliced. This "splicing memory" is encoded by the vestigial exonic splicing regulatory elements which previously defined exons and determined the position-dependent information on ancestral AS patterns (Fig. 5d). These ancestral AS events could be ES where exitron borders coincide exactly with ancestral exons. Alternatively, one of the exitron borders can represent an alternative SS which remained upon intron loss thus facilitating exitron evolution. A role of exonic splicing regulatory elements is suggested by the presence of highly conserved sequences predicted to be binding sites of SFs, as in the T-protein gene. Furthermore, SF motifs are over-represented in human intronless genes [35]. Upon intron loss, binding of the proteins with multiple functions in RNA processing to the sites still present in the exonic sequences may connect a region which no longer contains introns to the extensive RNA processing and splicing network. Tethering the spliceosomal components to these regions could favour mutations beneficial for the emergence of core splicing signals (or maintenance of existing ones) thus promoting exitron origin and restoring production of alternative transcripts *via* EIS. The evidence presented here supports a mechanism for the origin of exitrons, where their

evolution was a consequence of intron loss leaving behind footprints of former exon borders, defined by splicing regulatory sequences.

Taken together, we have identified and characterized a previously overlooked type of AS event - splicing of exitrons. Though exitrons can be misinterpreted as retained introns, we show that they are clearly distinguishable from them. We demonstrate that exitrons originate from exons and their evolution involves intron loss. EIS is the only AS event which does not involve canonical introns setting it apart from other types of AS events. Therefore we consider EIS as a novel type of AS, that complements the repertoire of AS events by allowing for intraexonic coding sequences to be differentially spliced. Surprisingly, our analysis revealed that intronless genes can also undergo AS *via* exitron usage, thus further eroding the dogma of one gene, one transcript, one protein. The evolutionary conservation of exitrons and their presence in species from plants to human demonstrate that EIS is a general type of AS with important functional implications.

METHODS SUMMARY

The sets of exitrons, retained introns, constitutive and alternative exons were obtained using our RNA-Seq data, published earlier [6]. The complete pipeline for the computational analyses and statistical procedures performed in the present work are described in Extended Data Methods. To verify the existence of exitron-spliced transcripts we used two EST datasets for *A. thaliana:* http://www.plantgdb.org/ and ftp://ftp.arabidopsis.org/home/tair/ and the High Resolution RT-PCR Alternative Splicing Panel [36-38]. *In vitro* transcription and subsequent RT-PCRs of five exitron-containing genes were performed with the TranscriptAid™ T7 High Yield Transcription Kit (Fermentas) using *A. thaliana* Col-0 genomic DNA. To validate EIS in the intronless genes, flower total RNA was extracted with the RNeasy Mini Kit

(Qiagen) followed by oligo-dT cDNA preparation using the Reverse Transcription System Kit (Promega). Two proteogenomic data sets for *A. thaliana* [15,16] were used to verify translation of exitron-containing and exitron-spliced isoforms (Supplementary Table 3). To test NMD sensitivity of isoforms, 3-week old plants of *upf1-5 and upf3-1* [39] alongside wild-type *A. thaliana* (Col-0) control plants were used. Regulation of EIS in different tissues was checked using wild type (Col-0) 10-day old seedlings and flowers. The effect of stresses on EIS was determined using 10-day old seedlings. To evaluate the effect of SR proteins on exitron splicing, 10-day old seedlings of the overexpression/mutant lines of At-SR30, At-RS31 and At-RS2Z33 were used alongside with wild-type *A. thaliana* (Col-0) control plants. Three-week old plants of SUA mutant lines, *sua-1* and *sua-2* [9] were used together with 3-week old wild type plants as controls (Ler-0 and Col-0 ecotypes, respectively). Details of plant materials, treatments, RT-PCRs and quantification are described in Extended Data Methods. PCR primers are listed in Supplementary table 4.

Full Methods and any associated references are available in the Extended Data.

References

1 Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).

2 Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14**, 802-813, doi:10.1261/rna.876308 (2008).

3 Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**, 11193-11198, doi:10.1073/pnas.201407298 (2001).

4 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415, doi:ng.259 [pii]
10.1038/ng.259 (2008).

5 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:nature07509 [pii]
10.1038/nature07509 (2008).

6 Marquez, Y., Brown, J. W., Simpson, C., Barta, A. & Kalyna, M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res* **22**, 1184-1195, doi:10.1101/gr.134106.111 (2012).

7 Staiger, D. & Brown, J. W. Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. *Plant Cell*, doi:10.1105/tpc.113.113803 (2013).

8 Zhong, X. Y., Wang, P., Han, J., Rosenfeld, M. G. & Fu, X. D. SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell* **35**, 1-10, doi:S1097-2765(09)00429-8 [pii]
10.1016/j.molcel.2009.06.016 (2009).

9 Sugliani, M., Brambilla, V., Clerkx, E. J., Koornneef, M. & Soppe, W. J. The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *Plant Cell* **22**, 1936-1946, doi:10.1105/tpc.110.074674 (2010).

10 O'Bryan, M. K. *et al.* RBM5 Is a Male Germ Cell Splicing Factor and Is Required for Spermatid Differentiation and Male Fertility. *PLoS Genet* **9**, e1003628, doi:10.1371/journal.pgen.1003628 (2013).

11 Bonnal, S. *et al.* RBM5/Luca-15/H37 regulates Fas alternative splice site pairing after exon definition. *Mol Cell* **32**, 81-95, doi:10.1016/j.molcel.2008.08.008 (2008).

12 Fushimi, K. *et al.* Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM5. *Proc Natl Acad Sci U S A* **105**, 15708-15713, doi:10.1073/pnas.0805569105 (2008).

13 Song, Z. *et al.* Solution structure of the second RRM domain of RBM5 and its unusual binding characters for different RNA targets. *Biochemistry* **51**, 6667-6678, doi:10.1021/bi300539t (2012).

14 Wong, J. J. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583-595, doi:10.1016/j.cell.2013.06.052 (2013).

15 Castellana, N. E. *et al.* Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci U S A* **105**, 21034-21038, doi:10.1073/pnas.0811066106 (2008).

16 Baerenfaller, K. *et al.* Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**, 938-941, doi:10.1126/science.1157956 (2008).

17 Brosch, M. *et al.* Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* **21**, 756-767, doi:10.1101/gr.114272.110 (2011).

18 Ner-Gaon, H. *et al.* Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J* **39**, 877-885 (2004).

19 Davey, N. E., Shields, D. C. & Edwards, R. J. SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res* **34**, 3546-3554, doi:10.1093/nar/gkl486 (2006).

20 Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9 Suppl 2**, S1, doi:10.1186/1471-2164-9-S2-S1 (2008).

21 Buljan, M. *et al.* Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* **46**, 871-883, doi:10.1016/j.molcel.2012.05.039 (2012).

22 Ellis, J. D. *et al.* Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* **46**, 884-892, doi:10.1016/j.molcel.2012.05.037 (2012).

23 Cordin, O., Banroques, J., Tanner, N. K. & Linder, P. The DEAD-box protein family of RNA helicases. *Gene* **367**, 17-37, doi:10.1016/j.gene.2005.10.019 (2006).
24 Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593, doi:10.1126/science.1230612 (2012).
25 Irimia, M. *et al.* Origin of introns by 'intronization' of exonic sequences. *Trends Genet* **24**, 378-381, doi:10.1016/j.tig.2008.05.007 (2008).
26 Liu, M. & Grigoriev, A. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet* **20**, 399-403 (2004).
27 O'Toole, N. *et al.* On the expansion of the pentatricopeptide repeat gene family in plants. *Mol Biol Evol* **25**, 1120-1128, doi:10.1093/molbev/msn057 (2008).
28 Nanao, K. *et al.* Identification of the mutations in the T-protein gene causing typical and atypical nonketotic hyperglycinemia. *Human genetics* **93**, 655-658 (1994).
29 Bauwe, H. & Kolukisaoglu, U. Genetic manipulation of glycine decarboxylation. *J Exp Bot* **54**, 1523-1535, doi:10.1093/jxb/erg171 (2003).
30 Mayer, M. P. & Bukau, B. Hsp70 chaperones: cellular functions and molecular mechanism. *Cell Mol Life Sci* **62**, 670-684, doi:10.1007/s00018-004-4464-6 (2005).
31 Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* **2**, E268, doi:10.1371/journal.pbio.0020268 (2004).
32 Good, P. J., Chen, Q., Warner, S. J. & Herring, D. C. A family of human RNA-binding proteins related to the Drosophila Bruno translational regulator. *J Biol Chem* **275**, 28583-28592, doi:10.1074/jbc.M003083200 (2000).
33 Kim, H. S., Abbasi, N. & Choi, S. B. Bruno-like proteins modulate flowering time via 3' UTR-dependent decay of SOC1 mRNA. *New Phytol* **198**, 747-756, doi:10.1111/nph.12181 (2013).
34 Catania, F. & Lynch, M. Where do introns come from? *PLoS Biol* **6**, e283, doi:10.1371/journal.pbio.0060283 (2008).
35 Pozzoli, U. *et al.* Over-representation of exonic splicing enhancers in human intronless genes suggests multiple functions in mRNA processing. *Biochem Biophys Res Commun* **322**, 470-476, doi:10.1016/j.bbrc.2004.07.144 (2004).
36 Simpson, C. G. *et al.* Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J* **53**, 1035-1048 (2008).
37 Simpson, C. G. *et al.* Alternative splicing in plants. *Biochem Soc Trans* **36**, 508-510 (2008).
38 Kalyna, M. *et al.* Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res* **40**, 2454-2469, doi:10.1093/nar/gkr932 (2012).
39 Arciga-Reyes, L., Wootton, L., Kieffer, M. & Davies, B. UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis. *Plant J* **47**, 480-489 (2006).

Extended Data and Supplementary Information accompany this paper and contain Extended Data Text, Extended Data Methods, Extended Data Figures 1-10, additional references and Supplementary Tables 1-23.

Acknowledgments We thank to John WS Brown for discussions during preparation of this manuscript, to Wim JJ Soppe for providing seed of SUA mutants and all the members of the lab for comments. This work was supported by the Austrian Science Fund (FWF) [DK W1207; ERA-NET Plant Genomics (PASAS) I254; SFB RNAreg F43-P10].

Author contributions Y.M. and M.K. designed all studies and experiments; Y.M. performed all large-scale computational analyses; M.K. performed several small-scale bioinformatic analyses; Y.M. performed *in vitro* transcription and subsequent RT-PCRs; Y.M. and M.H. performed experiments on the regulation of exitron splicing in different tissues, stress conditions, and mutant backgrounds; Z.A. verified exitron splicing in the intronless genes; A.B. provided advices and commentary on data analysis, helped to organize and supported the project.; and M.K. supervised all studies. The manuscript was prepared by Y.M. and M.K, with the participation of all authors.

Competing financial interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to: Maria Kalyna

Figure Legends

Fig. 1. Exitron (EI) definition, statistics and features. **a**, Schematic representation of EIS and its consequences. Exitron - dark blue. Constitutive and alternative splicing - green and red carets, respectively. PTC - stop sign. **b,** General statistics of *Arabidopsis* exitrons. **c,** Comparison of exitron and intron sizes. **d**, Comparison of GC content of exitrons and introns. **e**, Comparison of exitron and exon sizes. **f**, Comparison of GC content of exitrons and exons. **g**, Gene ontology classification of EI_{x3} –containing genes.

Fig. 2. Exitron splicing is a regulated process. **a**, Structures of exitron-containing genes tested by RT-PCRs in **b**. **b,** Heat map for differential EIS in different tissues, stress conditions and SF mutant backgrounds. Changes with a p-value ≤ 0.1 were considered as significant. **c,** SNPs affect EIS in Ler-0 ecotype. SNPs in the splicing signals are indicated in red. RT-PCR products of the full-length (FL) and exitron-spliced (EI) isoforms (filled and open triangles, respectively) are shown. Ubiquitin was used as a loading control (*).**d**, RBM5/SUA motif density. Densities for RBM5 motifs and MEME predicted motifs are significantly higher in exitrons (*, p-value < 0.0001).

Fig. 3. Functional implications of exitron splicing. **a**, Statistics of functional features of the protein sequences encoded by exitrons. **b,** Comparison of PFAM domains (left), disordered regions (middle) and short linear motifs (SLiMs) (right) encoded by exitrons and different types of exons. **c,** EIS removes ATP-binding domain of the highly conserved eukaryotic initiation factor 4A.

Fig. 4. Characterization of human exitrons. **a**, General statistics of the human exitrons. **b,** GC content distribution of exitrons and retained introns. **c,** Size

distribution of exitrons and retained introns. **d**, Comparison of exitron and exon sizes. PFAM domains **e**,, disordered regions **f**, and SLiMs **g**, in the sequences encoded by exitrons and other types of exons.

Fig. 5. Origin and evolution of exitrons. **a**, Statistics of exitrons that align to plant orthologous coding sequences. **b,** Exitron evolution in the intronless human *HSPA1A* gene. **c,** Exitron evolution in the gene encoding glycine cleavage system T-protein. Phylogenetic reconstruction of intron loss events in different plant species (1 and 2 indicate paralogues). Gene structures are not to scale. Introns at the conserved positions are coloured. **d**, Mechanism for evolution of EIS: the "splicing memory" hypothesis. Dashed lines show potential ancestral AS events.

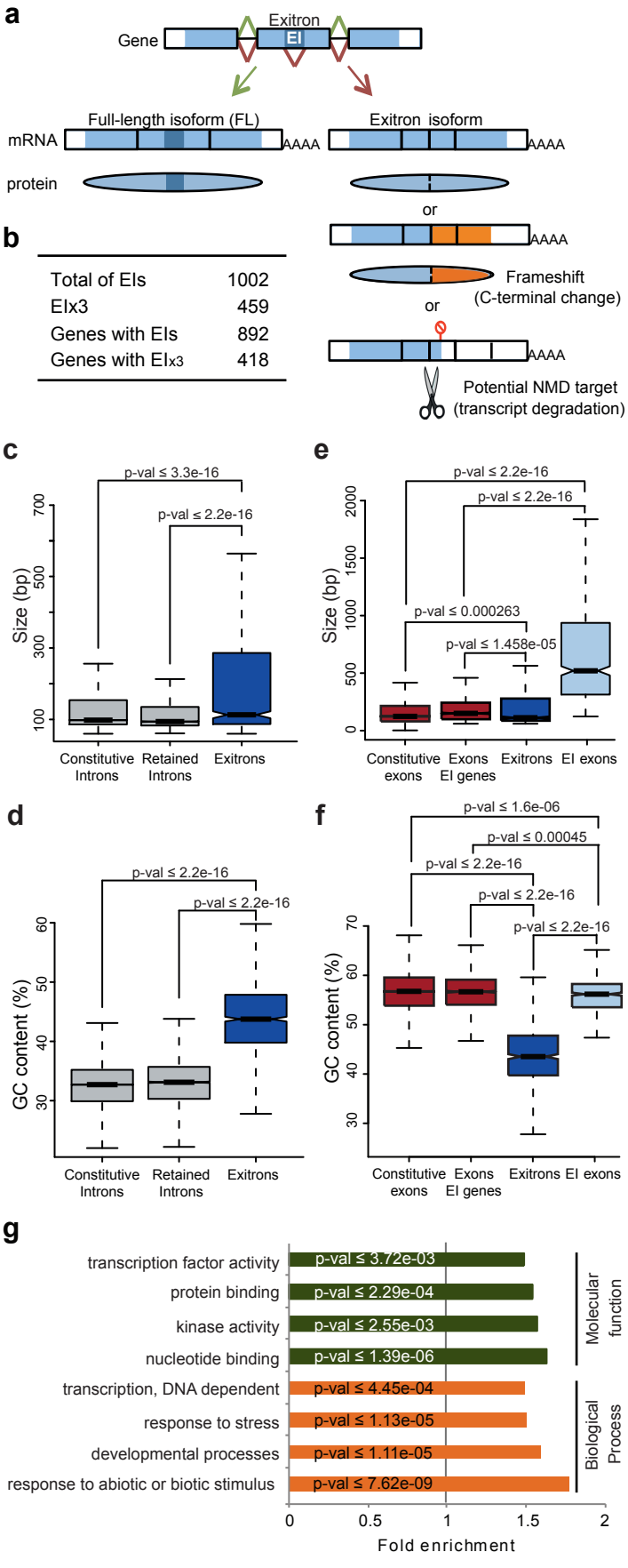
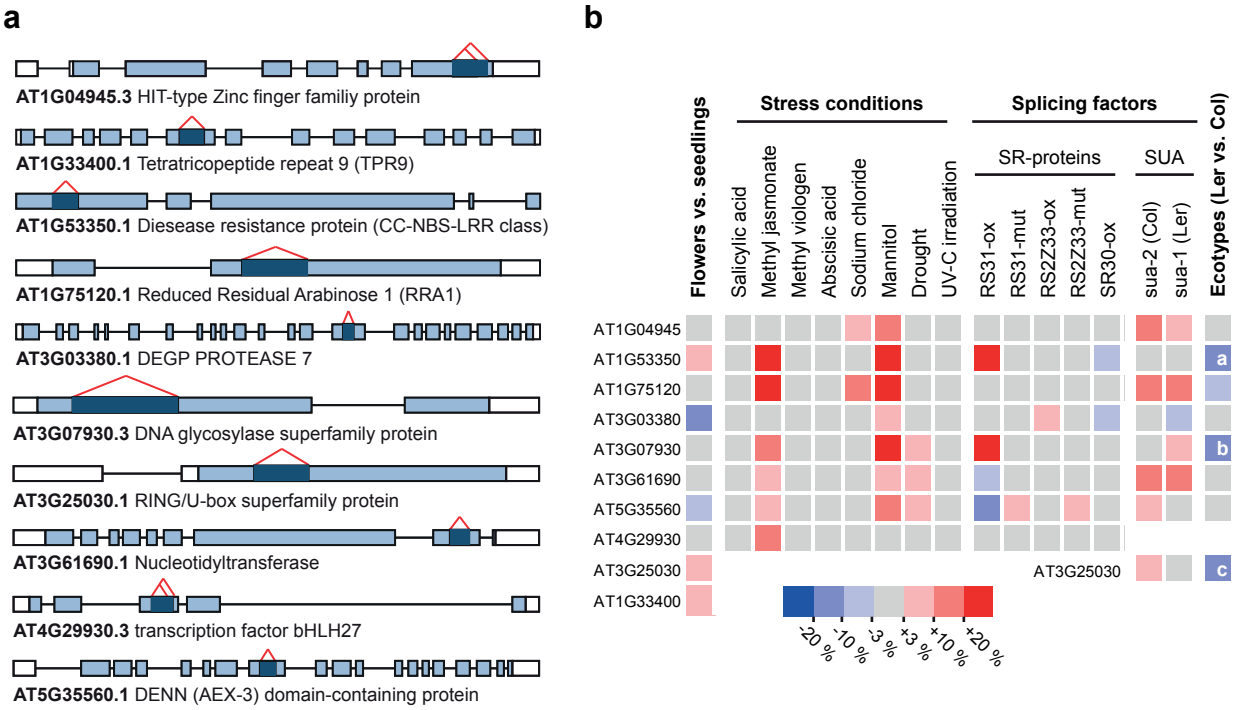

a
Gene
Exitron
EI
Full-length isoform (FL)
Exitron isoform
mRNA
AAAA
protein
or
Frameshift
(C-terminal change)
Potential NMD target
(transcript degradation)
b
Total of EIs 1002
EIx3 459
Genes with EIs 892
Genes with EIx3 418
c
p-val ≤ 3.3e-16
p-val ≤ 2.2e-16
Size (bp)
Constitutive Introns
Retained Introns
Exitrons
e
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
p-val ≤ 0.000263
p-val ≤ 1.458e-05
Size (bp)
Constitutive exons
Exons EI genes
Exitrons
EI exons
d
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
GC content (%)
Constitutive Introns
Retained Introns
Exitrons
f
p-val ≤ 1.6e-06
p-val ≤ 0.00045
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
GC content (%)
Constitutive exons
Exons EI genes
Exitrons
EI exons
g
transcription factor activity
protein binding
kinase activity
nucleotide binding
transcription, DNA dependent
response to stress
developmental processes
response to abiotic or biotic stimulus
p-val ≤ 3.72e-03
p-val ≤ 2.29e-04
p-val ≤ 2.55e-03
p-val ≤ 1.39e-06
p-val ≤ 4.45e-04
p-val ≤ 1.13e-05
p-val ≤ 1.11e-05
p-val ≤ 7.62e-09
Molecular function
Biological Process
Fold enrichment

Figure 1.



RBM5/SUA motif density per Kb						
	Exitrons	**Exitron exons**	**Constitutive introns**	**Retained introns**	**Constitutive exons**	**Genome**
RBM5 motifs						
CTCTTC	5.56	0.76	2.59	2.79	0.65	0.51
GAGAAG	10.51	1.24	0.81	0.76	1.06	0.57
Total	**16.07***	2.00	3.41	3.55	1.71	1.08
MEME motifs						
TCTTCT[T|C]C	3.26	0.43	1.52	1.72	0.32	0.32
GA[A|G]GAAG[A|C]	7.72	1.15	0.32	0.30	0.86	0.42
Total	**10.98***	1.58	1.84	2.01	1.18	0.74

Figure 2.

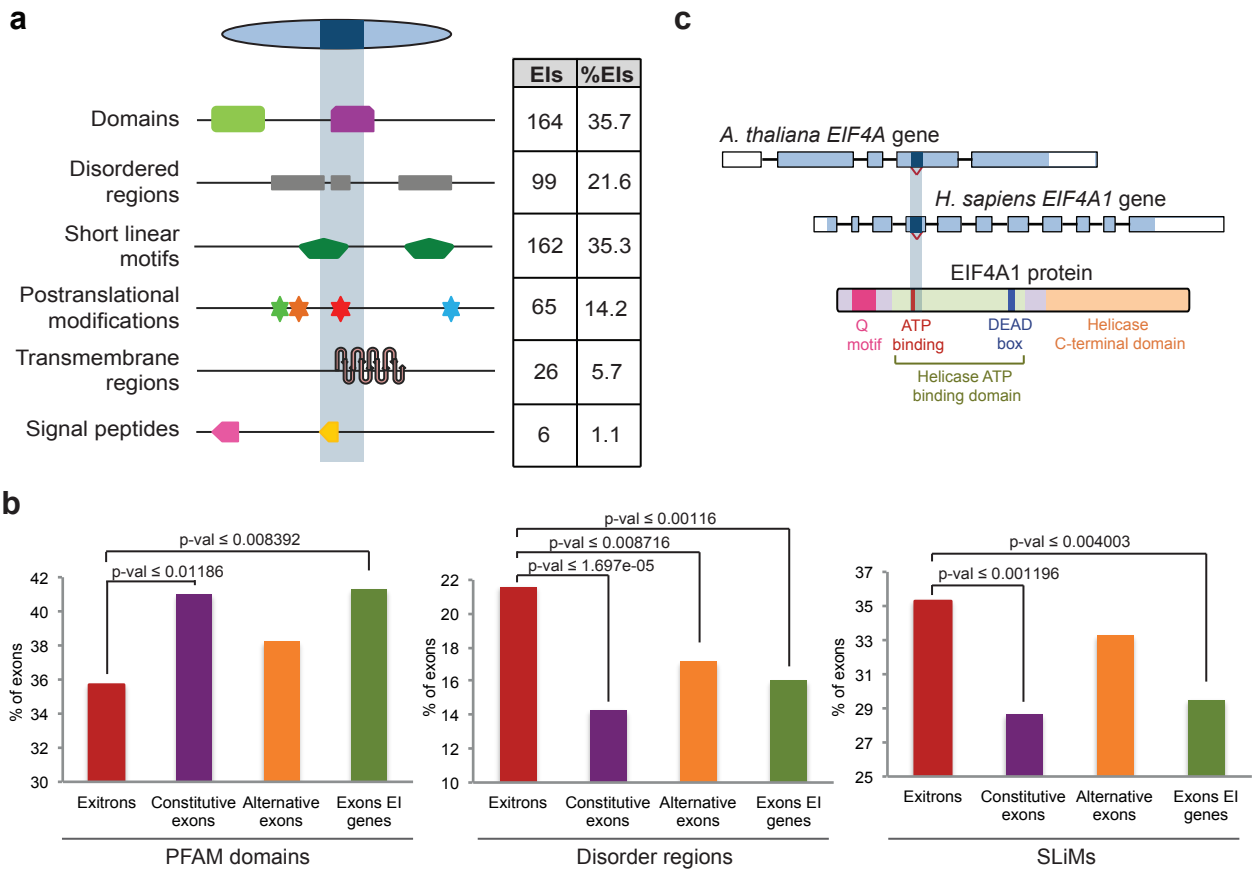

a
Domains
Disordered regions
Short linear motifs
Postranslational modifications
Transmembrane regions
Signal peptides
EIs
%EIs
164
35.7
99
21.6
162
35.3
65
14.2
26
5.7
6
1.1
c
A. thaliana EIF4A gene
H. sapiens EIF4A1 gene
EIF4A1 protein
Q motif
ATP binding
DEAD box
Helicase C-terminal domain
Helicase ATP binding domain
b
p-val ≤ 0.008392
p-val ≤ 0.01186
% of exons
Exitrons
Constitutive exons
Alternative exons
Exons EI genes
PFAM domains
p-val ≤ 0.00116
p-val ≤ 0.008716
p-val ≤ 1.697e-05
Disorder regions
p-val ≤ 0.004003
p-val ≤ 0.001196
SLiMs

Figure 3.

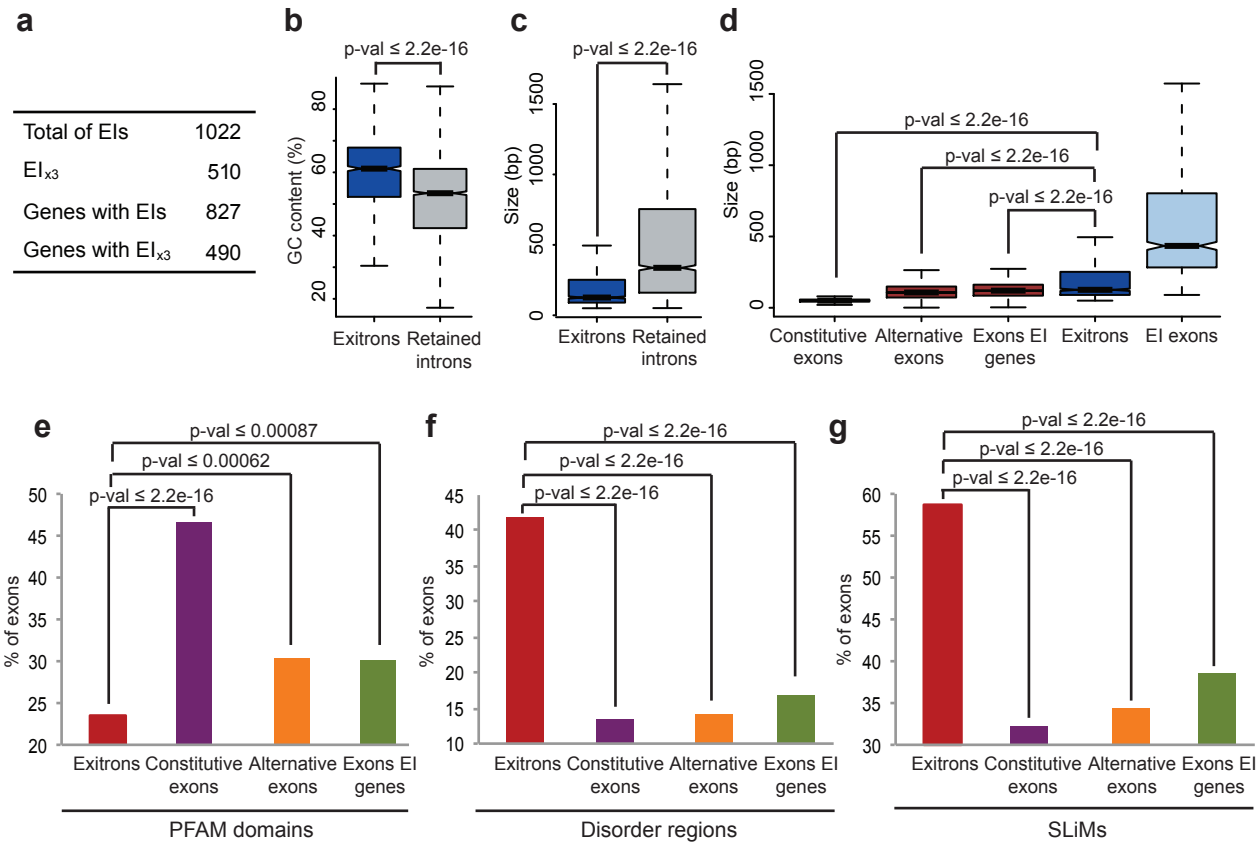
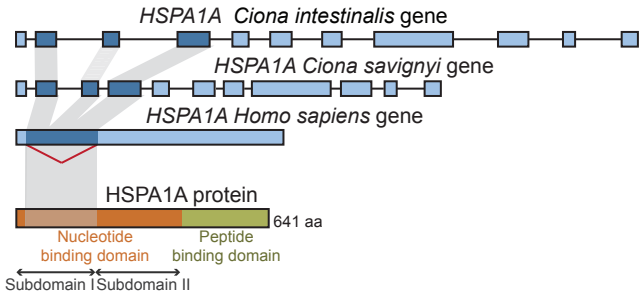

a
Total of EIs 1022
EIx3 510
Genes with EIs 827
Genes with EIx3 490
b
p-val ≤ 2.2e-16
GC content (%)
Exitrons
Retained introns
c
p-val ≤ 2.2e-16
Size (bp)
Exitrons
Retained introns
d
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
Size (bp)
Constitutive exons
Alternative exons
Exons EI genes
Exitrons
EI exons
e
p-val ≤ 0.00087
p-val ≤ 0.00062
p-val ≤ 2.2e-16
% of exons
Exitrons
Constitutive exons
Alternative exons
Exons EI genes
PFAM domains
f
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
% of exons
Exitrons
Constitutive exons
Alternative exons
Exons EI genes
Disorder regions
g
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
p-val ≤ 2.2e-16
% of exons
Exitrons
Constitutive exons
Alternative exons
Exons EI genes
SLiMs

Figure 4

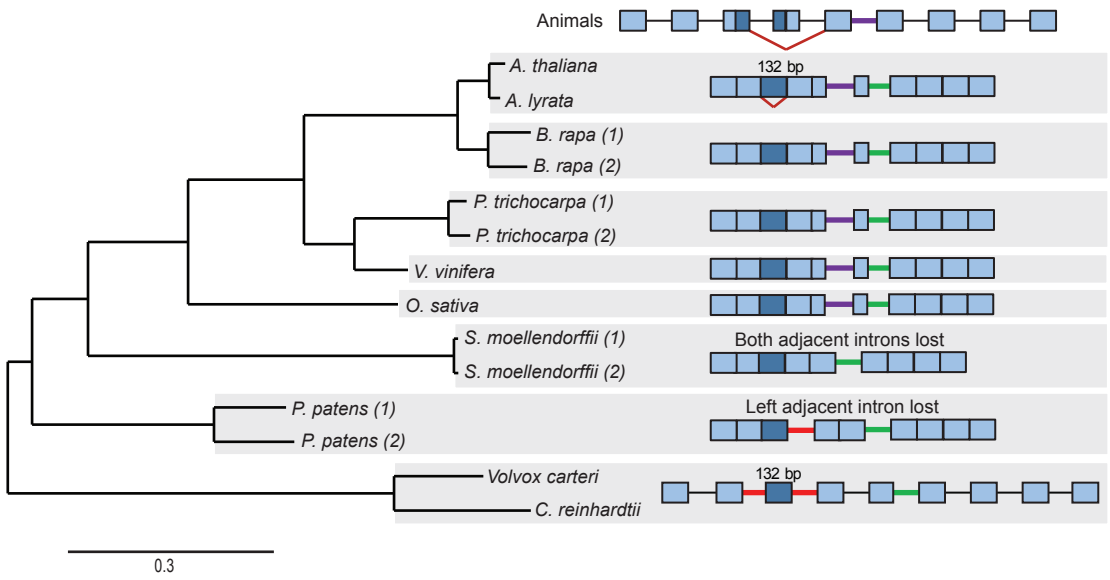
a

Majority of EIs align to coding sequences in orthologues				
Organism	**# of analyzed ortologues**	**Coding exon**	**Intron (coding)**	**Total coding (%)**
Arabidopsis lyrata	648	631	17 (4)	635 (98.0)
Brassica rapa	442	430	12 (1)	431 (97.5)
Populus trichocarpa	256	252	4 (0)	252 (98.4)
Vitis vinifera	252	246	6 (0)	246 (97.6)
Oryza sativa	163	158	5 (0)	158 (96.9)
Selaginella moellendorffii	62	62	0	62 (100)
Physcomitrella patens	73	72	1 (0)	72 (98.6)
Chlamydomonas reinhardtii	29	29	0	29 (100)

b



c



d

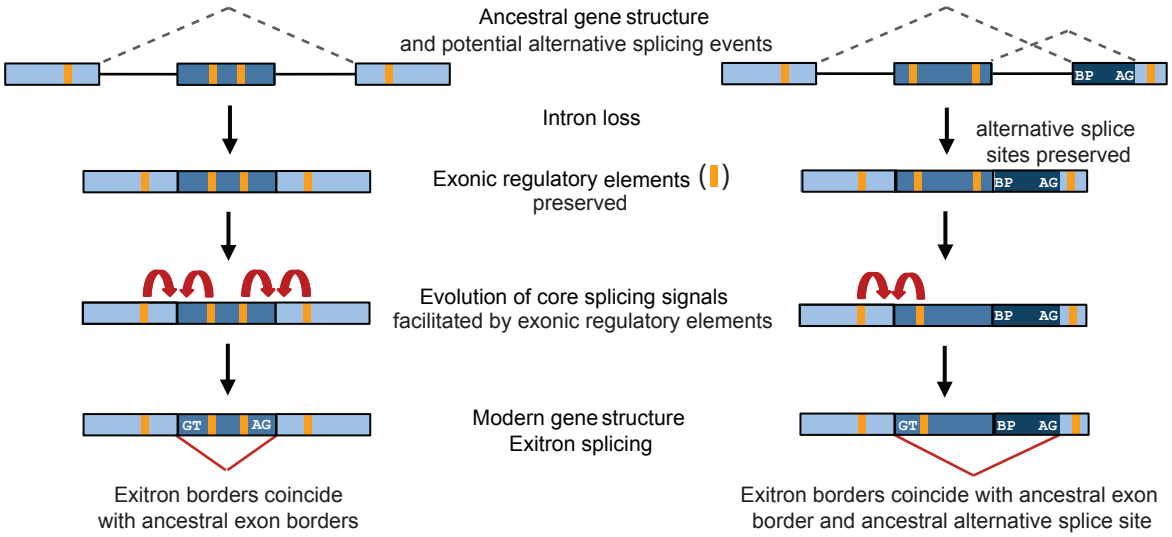


Figure 5

Concluding Discussion

Concluding Discussion

The coupling of alternative splicing and NMD regulate important genes in Arabidopsis, it follows the same rules as in other eukaryotes but with interesting exceptions.

The impact of NMD on the regulation of gene expression in plants has been largely unexplored. Few experiments based on specific transgenes revealed that in plants the same features of the transcript, as reported in yeast and mammals, operate to trigger NMD. These NMD eliciting features are: long 3' UTRs (≥350 nt), introns in the 3' UTR, a PTC located more than 50-55 nt upstream of a splice junction and the existence of uORFs (see introduction Figure 5). Analyses of plants impaired in NMD using tilling arrays suggested that NMD affected only 1% of the protein coding genes (Kurihara et al., 2009; Yoine, Ohto, Onai, Mita, & Nakamura, 2006). By contrast, ~78% of the splice variants detected in an RNA-Seq experiment have PTCs located 55 nt upstream of exon junctions, suggesting NMD as a major mechanism for controlling mRNA levels in Arabidopsis (Filichkin et al., 2010). These two apparent contradictory results implied that a more comprehensive analysis of NMD sensitivity of alternative splicing variants needed to be done.

In Kalyna et al. 2012 (see chapter), we have used the high resolution RT-PCR alternative splicing panel (HR RT-PCR AS panel, (Simpson et al., 2008)) for exploring the fate of the splicing variants produced in 270 genes (~950 transcripts) in the context of two NMD mutants (*upf1-5* and *upf3-1*) and treatment by cycloheximide (CHX), an inhibitor of translation and as consequence also of NMD. The advantages of using the HR RT-PCR AS panel compared to other technologies are its specificity, due to the gene specific primers that are designed to a particular region, its high resolution with a maximum of 3 nucleotides of size deviations, and its sensitivity enabling to detect minor changes in isoform ratios. Additionally, Sanger sequencing was performed to confirm ~300 novel spliced variants. Importantly, for the first time a substantial amount of endogenous alternatively splice variants were characterized in terms of NMD.

Our analysis of NMD impaired plants using the HR RT-PCR AS panel, led us to important findings in alternative splicing and its cross talk with NMD in plant systems. One of the major finding was that alternative splicing was strongly underestimated in plants, as the number of the identified AS variants increased by almost one third for the genes analyzed. Notably, around 67% of the AS transcripts were not affected in any of the NMD mutants or in the CHX treatment, suggesting that these variants are stable or not subjected to NMD. Furthermore, levels of NMD sensitive variants were not always low in wild type plants and could contribute tens of percent of the transcripts from a gene, suggesting that there is variability in regulating the amount of transcripts by NMD.

In agreement with its strong phenotype, UPF3 mutant line *upf3-1* had a broadest effect in increasing levels of NMD-sensitive transcripts in comparison to UPF1 mutant plants where a weak allele (*upf1-5*) has been used. Furthermore, CHX treated plants revealed the strongest effect in terms of variety of transcripts elevated when compared to wild type plants, many of them not overlapping with the *upf* mutants. Two main conclusions were done based on the results of the mutant lines and the CHX treatment. The first one is that presence of transcripts that only increased in CHX suggests that other degradation pathways linked to translation must exist besides NMD and/or the incidence of secondary effects due to translation inhibition by this drug. Secondly, the lack of correlation in the increase of transcripts levels between *upf3-1* and *upf1-5* (expected more increase under upf3-1) for some of the cases implies additional roles of these two NMD factors.

Our results from the 270 genes, show that 11-17% of the total transcripts analyzed are degraded by NMD and that 16% to 25% of AS variants analyzed are sensitive to NMD. Therefore we could extrapolate that around 13–18% of intron-containing genes may be regulated by AS and NMD in Arabidopsis, whereby 13% takes into account results only from *upf* mutants and 18% includes the CHX treated samples. This last result implies an important coupling of alternative splicing and the NMD pathway and agrees well with estimates for other eukaryotes and with a later RNA-Seq study of NMD deficient Arabidopsis plants on the genome-wide level showing that 13,3% of intron-containing genes are regulated by AS/NMD (Drechsel et al., 2013). We found that most of the NMD sensitive transcripts (~85%) followed the same rules described previously (see introduction Figure 5). Importantly, we observed that alternative splicing in the UTR regions has a strong influence on the fate of the isoform. In the case of the 3’ UTR, alternative splicing can affect the distance of the stop codon and

the 3' end or the positioning of an EJC after the stop due to the occurrence of 3"UTR introns. For the 5' UTR, AS events affecting the presence, the location and the length of uORFs have a strong impact on NMD sensitivity of a transcript, especially when the uORFs overlap with the original start codon. Finally, we observed that the presence of NMD eliciting features does not always result in NMD, indicating that additional rules might exist. Importantly, the intron retention transcripts, though having NMD features, are insensitive to NMD. We observed several exceptions when intron retention transcripts were indeed NMD sensitive, however that could be explained by additional AS events present in these transcripts and probably responsible for targeting them to the NMD pathway. Therefore in order to determine the final fate of a given transcript, the inspection of the splicing pattern of the whole isoform is mandatory. As intron retention is the major type of AS event in plants, the finding that intron retention transcripts are not subjected to NMD may account for high frequency of intron retention events in plant systems and additional research need to be done in order to understand the functions of intron retention in plants.

Altogether, our research revealed that NMD coupled to alternative splicing is a major mechanism for regulating gene expression in plants. Our findings confirmed the rules for NMD sensitivity described previously. Nevertheless, it highlighted that further research needs to be done for those cases were the rules did not explain the outcome as for the NMD insensitivity found in transcripts with retained introns. Very recently, colleagues in our lab explored the subcellular localization of some intron retention transcripts. They found that these mRNA variants are only located in the nucleus explaining why they escape the NMD machinery despite possessing NMD eliciting features (Gohring, Jacak, & Barta, 2014). Genome wide studies of nuclear and cytoplasmic fractions need to be done in order to determine whether the nuclear localization of intron retention transcripts is a general feature.

Alternative splicing is a major mechanism in the generation of transcript diversity in Arabidopsis

Several independent studies in plants together with our analysis of splice variants using the HR RT-PCR AS panel implied that there was a considerable underestimation of the extent of alternative splicing in Arabidopsis. Due to the fact that many of genes with AS variants were involved in important functions such as development and response to stress, the need to have a good annotation of the

repertoire of isoforms produced by AS became more evident. The availability of such information will allow evaluating gene expression at the level of transcripts and determining the biological impact of the AS variants in regulating the gene function.

In 2010, a study using RNA-Seq of plants grown under different stress conditions concluded that alternative splicing regulates 42% of the multiexonic genes in Arabidopsis. Although this work increased considerably the discovery of AS variants in Arabidopsis and supported the use of RNA-seq for discovery of novel mRNA isoforms, many technical challenges were still present, especially in terms of read length (only 36 nt) and sequencing depth. These technical disadvantages suggested that many isoforms failed to be detected and that the extent of alternative splicing was still underestimated.

In order to improve the detection of isoforms and determine the impact of alternative splicing on Arabidopsis transcriptome, we have used RNA-Seq of a normalized cDNA library. In respect to sample preparation, we generated a ds cDNA library of total RNA from two samples (equal concentrations of RNAs isolated from 10-day seedlings and flowers from plants grown under normal conditions), in order to improve the discovery of tissue-specific isoforms. Moreover, considering that in plants some mRNAs are very abundant (10 genes produce up to 25% of the total mRNA population), we subjected the library to a normalization procedure in order to reduce amount of these transcripts and to enable the detection of novel isoforms with lower expression levels. In terms of the sequencing technology, we used the Illumina platform with paired end 75 nt long sequencing reads to improve the finding of spliced reads and the quality of the alignment. Furthermore, we obtained more than 110 million of paired end reads for the ds cDNA library to increase the probability to capture low abundant isoforms.

Our approach resulted in a considerable reduction in the representation of highly abundant mRNAs, as only ~1% of the reads were aligned to the 10 highest abundant genes, in comparison to almost 7% to 25% of the reads in other HTS data. Notably, more than 30% of the predicted splice junctions are coming from not annotated introns. We observed that many of the new introns detected were slightly bigger and present higher GC content when contrasted to canonical introns in Arabidopsis. These features suggested that many of them are involved in alternative splicing events, where introns usually are longer than constitutive ones, and are generally less efficiently spliced due to their higher GC content. In terms of splice site signals,

we found that most of the newly identified introns possess U2 splice signal sequences, but we obtained a substantial amount of introns resembling U12 splicing signals (more than 2000 introns). In general, the splicing signals of the newly detected introns presented lower scores than the scores of the annotated ones, suggesting that other regulatory signals can be involved in the intron to improve their recognition in specific conditions.

As transcript assembly is a challenging task when using short reads, we first obtained an approximate of the number of alternatively spliced genes in our sample. The basis of this strategy relies on the presence of overlapping splice junctions that were mapped to each gene, since overlapping splice junctions must derive from different isoforms. Based on this, we found that at least 51% of the genes are alternatively spliced. It is important to note that, according to our strategy, this number does not consider cases of intron retention, indicating that AS events other than intron retention contribute significantly to increase the number of mRNA variants.

Finally, we assembled putative transcripts using all our RNA-Seq data. We have used an annotation-based assembly for this purpose, where we used the information of the start, end and strand of the known genes to predict possible splice variants. Additionally we have filtered the variants that were lowly supported by the read alignment. Using these criteria, we obtained more than 50,000 transcripts and we manage to recapitulate at least one known isoform for 75% of the protein coding genes expressed in our sample. Furthermore, we have used the information of the isoforms on the HR RT-PCR AS panel to confirm the transcript assemblies. For the set of 270 genes in the AS panel, 92% of the putative transcripts assembled using RNA-Seq were also detected by the HR RT-PCR AS panel. Importantly, the HR RT-PCR AS panel reported more isoforms for these genes than detected by RNA-Seq implying that our set of AS transcripts is not exhaustive.

Considering the set of assembled transcripts, we estimate that alternative splicing regulates at least 61% of the intron containing genes in Arabidopsis. We explored which types of AS events affect the transcripts and in agreement with previous studies, intron retention was the most frequent type of AS event. Due to the high frequency of retained introns, we decided to analyze their features and compared them to constitutive introns in our sample. We observed that the majority of the retained introns were lowly covered by reads, implying that intron retention

transcripts are low abundant. Moreover, many of the retained introns have weak splicing signals and high GC content, correlating positively with the number of reads covering the intron, explaining their low splicing efficiency. Furthermore, we identified a fraction of highly retained introns with particular features, such as very high GC content and the lack of stop codons. We found that these introns were located inside coding exons and we named them cryptic introns (see chapter). Altogether, our analysis of retained introns led us to conclude that despite the high frequency of this AS event, most of the intron retention containing transcripts are low abundant, inferring that their contribution to the total mRNA populations is minor. Moreover, if we subtracted the transcripts with retained introns, we obtained that 51% of the genes produce isoforms that do not involve intron retention. This estimate agrees well with the one obtained using only overlapping splice junctions (see above). This result indicates that Arabidopsis employs mostly other alternative splicing events to increase its repertoire of mRNA variants.

The combination of a normalized library with RNA-Seq in Arabidopsis revealed that alternative splicing is a major mechanism in the generation of transcript variability in Arabidopsis, with at least 61% of its genes being affected by this process. Moreover, the applied approach allowed the identification of novel AS variants to complement and contribute the annotation of isoforms in this model plant. Finally, we explored in detail at the genome-wide level the features of retained introns and their impact on Arabidopsis transcriptome. Altogether, our work provides a very valuable resource for plant scientists involved in studies of the regulation of gene expression, setting the basis for future research to explore the biological impact of the alternatively spliced variants. Even though our strategy increased considerably the discovery of new isoforms, these transcripts are probably representing only a snapshot of all the existing mRNA variants in Arabidopsis. As we have analyzed only two biological samples obtained from seedlings and flowers grown under normal conditions, we anticipate that the availability of sequencing data from more tissues and environmental conditions will increase the current estimate of genes undergoing alternative splicing in Arabidopsis.

Exitron splicing is a new type of alternative splicing event

By studying the alternative splicing landscape of Arabidopsis using RNA-Seq and more specifically, while evaluating the features of intron retention, we identified a set

of retained introns with unusual characteristics that we named cryptic introns. Majority of these introns have weaker splice sites, higher GC content and are more prone to be retained when compared to other retained introns. Analysis of their location in the gene structures revealed that they are embedded inside coding exons therefore we renamed them as exitrons (**ex**onic i**n**t**rons**). The unique feature of this event, where internal parts of coding exons can be differentially spliced, led us to investigate this so far unexplored alternative splicing event.

Using our previous RNA-Seq experiment from wild type 10-day seedlings and wild type flowers, we obtained a total of 1,002 exitrons in 892 genes. Interestingly around 46% of exitrons have a length that is divisible by three and, therefore, their splicing does not introduce any stop codon and will result in a protein variant lacking an internal stretch of aminoacids when compared to the full-length protein isoform. Splicing of the remaining exitrons with sizes no divisible by three might generate protein isoforms with different C-termini or alter transcript variants stability, for example by targeting the transcript for degradation by NMD.

As exitrons were identified in the set of retained introns, we inspected exitron features to determine if they are distinguishable from retained introns. Analyses of exitron features showed that they indeed are different from retained introns and constitutive introns, as they have significant bigger sizes and higher GC content. The high GC content observed in the exitrons is in agreement with their intrinsic nature of being parts of the protein coding sequences. Together with this, comparison of exitrons and constitutive exons showed that they have similar size distribution but the GC content of exitrons is lower. The lower GC content of exitrons in contrast to constitutive exons may play a role in exitron recognition by the splicing machinery. Interestingly, the exons that are hosting the exitrons are significantly longer than constitutive exons.

An indication for the biological relevance of alternatively splice variants is their regulation under specific tissues and/or environmental conditions. In order to test whether exitron splicing is responsive to tissue or stress conditions, we performed RT-PCRs of ten distinct genes containing exitrons with lengths divisible by three. Remarkably, we found that exitron splicing is differentially regulated in distinct tissues (seedlings *vs* flowers), indicating that exitron splicing can be influenced by developmental stages. Additionally, we found that exitron splicing is changing under stress conditions. Notably, for some of the conditions tested we could observe that

the exitron-spliced isoform can contribute to up to 50% of the total population of transcripts in specific genes, implying that they might have important roles in particular conditions. Furthermore, we found that exitron splicing can be affected by splicing factors. The splicing factor having the widest effect on exitron splicing (affecting 80% of the exitrons tested) was SUA. Consistent with its inhibitory effect on splicing of a cryptic intron in *ABI3* gene reported before (Sugliani, Brambilla, Clerkx, Koornneef, & Soppe, 2010), the levels of the exitron-spliced isoforms increased in SUA deficient plants. We found that binding motifs of RBM5, the mammalian orthologue for SUA having a role in exon skipping regulation, were enriched in the exitrons sequences. In addition, *de novo* motif finding confirmed the enrichment of RBM5-like binding motifs. These results suggest that SUA plays a critical role in exitron splicing.

In the course of the RT-PCR experiments using SUA deficient plants, we tested exitron splicing in two ecotypes of Arabidopsis, Col-0 and Ler-0, where we observed differences in the abundance of some of the exitron-spliced transcripts between these two accessions. Inspection of these genes revealed that single nucleotide polymorphisms (SNPs) located in splicing signals could explain the observed differences on exitron splicing in these two ecotypes. Moreover, our further analyses showed that these cases are not isolated as SNPs in 82 Arabidopsis ecotypes can affect the splicing signals of exitrons when compared to the reference ecotype Col-0. These natural variations of exitron sequences, especially in the splicing signals, in different Arabidopsis accessions in conjunction with that many exitron-containing genes have functions related to stress response, imply that exitron splicing might be playing a role in ecotype adaptation.

As exitrons are internal parts of coding exons, we evaluated the features of the encoded protein isoforms to determine the possible functional impact of exitron splicing. For that purpose we have analyzed the exitrons with sizes divisible by three, as these exitron-spliced isoforms will lack an internal part of amino acid sequences when compared to the full-length protein isoforms, in which the exitron is not spliced out. We found that protein domains are underrepresented in sequences enconded by exitrons, but for the cases when exitron splicing affects protein domains, these protein domains are mainly involved in the interactions with other molecules. In line with this result, we have found that amino acid sequences encoded by exitrons are enriched in intrinsically disordered regions and in short linear motifs (SLiMs). Interestingly, SLiMs can form binding platforms for other molecules, be sites of

posttranslational modifications (PTMs), targeting signals for subcellular compartmentalization or cleavage sites for endopeptidases forming “degrons”, degradation signals for a proteolytic apparatus. Moreover, we have found the some protein sequences encoded by exitrons also carry post-translational modification sites that are known to regulate protein function. Altogether, the features found in the sequences encoded by the exitrons suggest that exitron splicing can have a role in regulating protein interaction networks.

By using ESTs from other plant species, we got evidence that exitron splicing is not restricted to Arabidopsis and can be conserved. Additionally, we also explored the presence of this new type of alternative splicing in human annotated transcripts. We found 1022 exitrons in human having features similar to Arabidopsis exitrons, such a higher GC content when compared to retained introns and similar size distribution as Arabidopsis exitrons. Notably, human exitrons are also coding mainly for disordered regions. This last feature is also characteristic for tissue-specific alternative exons in humans, implying that alternative splicing can be critical in regulating the dynamics of protein networks. Evidence of exitrons in non-plant species indicates that this new type of alternative splicing event is not restricted to plants and might have a role in rewiring protein interaction networks in both organisms.

Due to the intraexonic nature of exitrons and their particular features, we decided to investigate the origin of exitrons and determine how these sequences acquired the ability to be spliced. By analyzing structures of ortologues of exitron-containing genes in different plant species, we have found out that exitrons originate from ancestral coding exons. Moreover we found evidence that introns loss occurred at the exitron borders or close to them, including the cases of exitrons in intronless genes. Detailed analysis of the exitron in the gene coding for the glycine cleavage system T-protein revealed that this exitron corresponds to a coding exon bordered by canonical introns in the algae orthologues. In addition, we found stretches of highly conserved nucleotide sequences inside and bordering this exitron that resemble binding sites for splicing and mRNA processing factors. Evidence of intron loss was also found in the human intronless gene coding for the heat shock protein HSPA1 (Hsp70), indicating that intron loss could be a general mechanism behind exitron evolution. Considering all the evidence, we proposed the “splicing memory” hypothesis for exitron evolution. Our hypothesis states that intron loss occurred exactly or close to the borders of the exitrons, where these ancestral exons “remember” that they were

previously subjected to splicing by the influence of binding sites for mRNA processing factors that result in their current splicing.
In this work we characterized a new type of alternative splicing event, exitron splicing that complements the already known set of AS events. Moreover, for the first time the origin of an alternative splicing event has been elucidated. Altogether, our study sets the bases for studying exitrons in other species and their impact on organism biology.

Future perspectives of alternative splicing in plants

Latest research has revealed that the finely tuned gene expression and protein production in a given condition lies in regulatory layers that result from the cross talk between splicing and others pre-mRNA processing pathways affecting the final fate of the mRNA product (Braunschweig et al., 2013). In recent years, alternative splicing has been established as an important mechanism for regulating gene expression and for the production of protein variants. Although many studies have highlighted that alternative splicing is an important regulator of gene expression by targeting mRNA variants for degradation, little experimental evidence exists about protein isoforms and further research is needed in this respect. Nevertheless, the complexity and dynamics of the alternative splicing profiles in the course of vertebrates evolution have suggested that alternative splicing had a fundamental contribution during speciation (Barbosa-Morais et al., 2012; Merkin et al., 2012), supporting its biological significance.

In plants, few studies are related to alternative splicing, even though its importance for development and response to stress is well documented. In the course of my PhD studies, I have had the opportunity to contribute in determining the impact of alternative splicing on the *Arabidopsis thaliana* transcriptome. We found that alternative splicing is also a crucial regulator of gene expression in plants by targeting mRNA variants to the NMD pathway. We observed that the majority of the cases follow the same rules described previously, though novel features have been identified as well. Nevertheless, additional investigation is necessary to explore the link between alternative splicing and NMD in specific cells, tissues and/or conditions as a starting point to evaluate contribution of these processes to the observed phenotypes. Moreover, using RNA-Seq of seedlings and flowers, we showed that alternative splicing is a major contributor to transcriptome complexity in Arabidopsis,

with more than 60% of the multi-exonic genes being regulated by this process. The availability of a more complete annotation of alternative splicing events will allow studying genes at the level of isoforms and evaluating their contribution to the gene function. We expect that upcoming sequencing data from different tissues and conditions will increase our latest estimate and will continue to improve the annotation of alternative splicing events and transcript variants in Arabidopsis for future studies involving gene regulation and protein isoforms. Finally, we identified and characterized a new type of alternative splicing event, exitron splicing, where intraexonic protein coding regions can be alternatively spliced. This study showed that exitrons have unique properties that distinguish them from other types of introns and that they originate from ancestral coding exons through intron loss events in the course of their evolution. The intrinsic characteristic of exitrons as internal parts of coding exons will encourage the study of the mechanistic principles behind their recognition by the spliceosome.

We expect that ours and other studies of alternative splicing in plants will ultimately change the vision of plant scientists when studying gene regulation.

References

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol, 8*(2), 135-141. doi: 10.1016/j.pbi.2005.01.001

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., . . . Ast, G. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep, 1*(5), 543-556. doi: 10.1016/j.celrep.2012.03.013

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., & Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res, 19*(10), 1732-1741. doi: 10.1101/gr.092353.109

Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res, 38*(14), 4570-4578. doi: 10.1093/nar/gkq211

Baek, D., & Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci U S A, 102*(36), 12813-12818.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., . . . Frey, B. J. (2010). Deciphering the splicing code. *Nature, 465*(7294), 53-59. doi: nature09000 [pii]

10.1038/nature09000

Barbazuk, W. B., Fu, Y., & McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res, 18*(9), 1381-1392. doi: gr.053678.106 [pii]

10.1101/gr.053678.106

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., . . . Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science, 338*(6114), 1587-1593. doi: 10.1126/science.1230612

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

Barta, A., Kalyna, M., & Reddy, A. S. (2010). Implementing a rational and consistent nomenclature for serine/arginine-rich protein splicing factors (SR proteins) in plants. *Plant Cell, 22*(9), 2926-2929. doi: tpc.110.078352 [pii]

10.1105/tpc.110.078352

Barta, A., Sommergruber, K., Thompson, D., Hartmuth, K., Matzke, M. A., & Matzke, A. J. M. (1986). The expression of a nopaline synthase - human growth hormone chimaeric gene in transformed tobacco and sunflower callus tissue. *Plant Mol Biol, 6*, 347-357.

Batsche, E., Yaniv, M., & Muchardt, C. (2006). The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol, 13*(1), 22-29. doi: 10.1038/nsmb1030

Behm-Ansmant, I., Kashima, I., Rehwinkel, J., Sauliere, J., Wittkopp, N., & Izaurralde, E. (2007). mRNA quality control: an ancient machinery recognizes and degrades mRNAs with nonsense codons. *FEBS Lett, 581*(15), 2845-2853. doi: S0014-5793(07)00555-8 [pii]

10.1016/j.febslet.2007.05.027

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., . . . Dreyfuss, G. (2012). U1 snRNP determines mRNA length and regulates isoform expression. *Cell, 150*(1), 53-64. doi: 10.1016/j.cell.2012.05.029

Berget, S. M. (1995). Exon recognition in vertebrate splicing. *J Biol Chem, 270*(6), 2411-2414.
Beyer, A. L., & Osheim, Y. N. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev, 2*(6), 754-765.
Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem, 72*, 291-336.
Bothwell, A. L., Ballard, D. W., Philbrick, W. M., Lindwall, G., Maher, S. E., Bridgett, M. M., . . . Garcia-Blanco, M. A. (1991). Murine polypyrimidine tract binding protein. Purification, cloning, and mapping of the RNA binding domain. *J Biol Chem, 266*(36), 24657-24663.
Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R., & Blencowe, B. J. (2013). Dynamic integration of splicing within gene regulatory pathways. *Cell, 152*(6), 1252-1269. doi: 10.1016/j.cell.2013.02.034
Brett, D., Pospisil, H., Valcarcel, J., Reich, J., & Bork, P. (2002). Alternative splicing and genome complexity. *Nat Genet, 30*(1), 29-30. doi: 10.1038/ng803
ng803 [pii]
Brown, J. W., Feix, G., & Frendewey, D. (1986). Accurate in vitro splicing of two pre-mRNA plant introns in a HeLa cell nuclear extract. *EMBO J., 5*, 2749-2758.
Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell, 46*(6), 871-883. doi: 10.1016/j.molcel.2012.05.039
Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M., & Buell, C. R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics, 7*, 327.
Carpenter, C. D., Kreps, J. A., & Simon, A. E. (1994). Genes encoding glycine-rich Arabidopsis thaliana proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm. *Plant Physiol, 104*(3), 1015-1025.
Chang, Y. F., Imam, J. S., & Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem, 76*, 51-74. doi: 10.1146/annurev.biochem.76.050106.093909
Cosson, B., Couturier, A., Chabelskaya, S., Kiktev, D., Inge-Vechtomov, S., Philippe, M., & Zhouravleva, G. (2002). Poly(A)-binding protein acts in translation termination via eukaryotic release factor 3 interaction and does not influence [PSI(+)] propagation. *Mol Cell Biol, 22*(10), 3301-3315.
Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., . . . dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res, 16*(6), 738-749. doi: 10.1101/gr.4825606
David, C. J., Boyne, A. R., Millhouse, S. R., & Manley, J. L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes Dev, 25*(9), 972-983. doi: 10.1101/gad.2038011
David, C. J., & Manley, J. L. (2011). The RNA polymerase C-terminal domain: a new role in spliceosome assembly. *Transcription, 2*(5), 221-225. doi: 10.4161/trns.2.5.17272
Drechsel, G., Kahles, A., Kesarwani, A. K., Stauffer, E., Behr, J., Drewe, P., . . . Wachter, A. (2013). Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell, 25*(10), 3726-3742. doi: 10.1105/tpc.113.115485
Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., . . . Blencowe, B. J. (2012). Tissue-specific alternative splicing remodels protein-

protein interaction networks. *Mol Cell, 46*(6), 884-892. doi: 10.1016/j.molcel.2012.05.037

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., . . . Mockler, T. C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res, 20*(1), 45-58. doi: gr.093302.109 [pii]
10.1101/gr.093302.109

Fresco, L. D., & Buratowski, S. (1996). Conditional mutants of the yeast mRNA capping enzyme show that the cap enhances, but is not required for, mRNA splicing. *RNA, 2*(6), 584-596.

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods, 8*(6), 469-477. doi: 10.1038/nmeth.1613

Garcia-Blanco, M. A., Jamison, S. F., & Sharp, P. A. (1989). Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev, 3*(12A), 1874-1886.

Ghosh, S., Marchand, V., Gaspar, I., & Ephrussi, A. (2012). Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA. *Nat Struct Mol Biol, 19*(4), 441-449. doi: 10.1038/nsmb.2257

Gohring, J., Jacak, J., & Barta, A. (2014). Imaging of Endogenous Messenger RNA Splice Variants in Living Cells Reveals Nuclear Retention of Transcripts Inaccessible to Nonsense-Mediated Decay in Arabidopsis. *Plant Cell*. doi: 10.1105/tpc.113.118075

Golovkin, M., & Reddy, A. (1999). An SC35-like protein and a novel serine/arginine-rich protein interact with Arabidopsis U1-70K protein. *J Biol Chem, 274*(51), 36428-36438.

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

Goodall, G. J., & Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell, 58*, 473-483.

Gornemann, J., Kotovic, K. M., Hujer, K., & Neugebauer, K. M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol Cell, 19*(1), 53-63. doi: 10.1016/j.molcel.2005.05.007

Grabowski, P. J. (2007). RNA-binding proteins switch gears to drive alternative splicing in neurons. *Nat Struct Mol Biol, 14*(7), 577-579. doi: 10.1038/nsmb0707-577

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. *RNA, 6*(9), 1197-1211.

Guan, Q., Zheng, W., Tang, S., Liu, X., Zinkel, R. A., Tsui, K. W., . . . Culbertson, M. R. (2006). Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast. *PLoS Genet, 2*(11), e203. doi: 10.1371/journal.pgen.0020203

Gullerova, M., Barta, A., & 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*(4), 631-643.

Guo, M., Lo, P. C., & Mount, S. M. (1993). Species-specific signals for the splicing of a short Drosophila intron in vitro. *Mol Cell Biol, 13*(2), 1104-1118.

Guo, S., Zheng, Y., Joung, J. G., Liu, S., Zhang, Z., Crasta, O. R., . . . Fei, Z. (2010). Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics, 11*, 384. doi: 10.1186/1471-2164-11-384

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., . . . Regev, A. (2013). De novo transcript sequence reconstruction from RNA-

seq using the Trinity platform for reference generation and analysis. *Nat Protoc, 8*(8), 1494-1512. doi: 10.1038/nprot.2013.084

Han, J., Xiong, J., Wang, D., & Fu, X. D. (2011). Pre-mRNA splicing: where and when in the nucleus. *Trends Cell Biol, 21*(6), 336-343. doi: 10.1016/j.tcb.2011.03.003

Han, S. P., Tang, Y. H., & Smith, R. (2010). Functional diversity of the hnRNPs: past, present and perspectives. *Biochem J, 430*(3), 379-392. doi: 10.1042/BJ20100396

Harada, Y., Funatsu, T., Murakami, K., Nonoyama, Y., Ishihama, A., & Yanagida, T. (1999). Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys J, 76*(2), 709-715. doi: 10.1016/S0006-3495(99)77237-1

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

Hattori, D., Millard, S. S., Wojtowicz, W. M., & Zipursky, S. L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. *Annu Rev Cell Dev Biol, 24*, 597-620. doi: 10.1146/annurev.cellbio.24.110707.175250

He, F., Li, X., Spatrick, P., Casillo, R., Dong, S., & Jacobson, A. (2003). Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. *Mol Cell, 12*(6), 1439-1452.

Heintzen, C., Melzer, S., Fischer, R., Kappeler, S., Apel, K., & Staiger, D. (1994). A light- and temperature-entrained circadian clock controls expression of transcripts encoding nuclear proteins with homology to RNA-binding proteins in meristematic tissue. *Plant J, 5*(6), 799-813.

Hertel, K. J. (2008). Combinatorial control of exon recognition. *J Biol Chem, 283*(3), 1211-1215. doi: 10.1074/jbc.R700035200

Hertel, K. J., Lynch, K. W., & Maniatis, T. (1997). Common themes in the function of transcription and splicing enhancers. *Curr Opin Cell Biol, 9*(3), 350-357.

Hiller, M., Zhang, Z., Backofen, R., & Stamm, S. (2007). Pre-mRNA secondary structures influence exon recognition. *PLoS Genet, 3*(11), e204. doi: 10.1371/journal.pgen.0030204

Hillman, R. T., Green, R. E., & Brenner, S. E. (2004). An unappreciated role for RNA surveillance. *Genome Biol, 5*(2), R8. doi: 10.1186/gb-2004-5-2-r8

gb-2004-5-2-r8 [pii]

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., & Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science, 325*(5940), 626-628. doi: 10.1126/science.1172926

Hoshino, S., Imai, M., Kobayashi, T., Uchida, N., & Katada, T. (1999). The eukaryotic polypeptide chain releasing factor (eRF3/GSPT) carrying the translation termination signal to the 3'-Poly(A) tail of mRNA. Direct association of erf3/GSPT with polyadenylate-binding protein. *J Biol Chem, 274*(24), 16677-16680.

House, A. E., & Lynch, K. W. (2006). An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol, 13*(10), 937-944. doi: 10.1038/nsmb1149

Hsin, J. P., & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev, 26*(19), 2119-2137. doi: 10.1101/gad.200303.112

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., . . . Shinozaki, K. (2004). Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Res, 32*(17), 5096-5103. doi: 10.1093/nar/gkh845

32/17/5096 [pii]

Isken, O., & Maquat, L. E. (2007). Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev, 21*(15), 1833-1856. doi: 21/15/1833 [pii]

10.1101/gad.1566807
Iwata, H., & Gotoh, O. (2011). Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics, 12*, 45. doi: 10.1186/1471-2164-12-45
Izaurralde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E., & Mattaj, I. W. (1994). A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell, 78*(4), 657-668.
Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F., & Ratsch, G. (2010). RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics, Chapter 11*, Unit 11 16. doi: 10.1002/0471250953.bi1106s32
Johansson, M. J., He, F., Spatrick, P., Li, C., & Jacobson, A. (2007). Association of yeast Upf1p with direct substrates of the NMD pathway. *Proc Natl Acad Sci U S A, 104*(52), 20872-20877. doi: 10.1073/pnas.0709257105
Johansson, M. J., & Jacobson, A. (2010). Nonsense-mediated mRNA decay maintains translational fidelity by limiting magnesium uptake. *Genes Dev, 24*(14), 1491-1495. doi: 10.1101/gad.1930710
Kalyna, M., & Barta, A. (2004). A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? *Biochem Soc Trans, 32*(Pt 4), 561-564.
Kalyna, M., Lopato, S., & Barta, A. (2003). Ectopic expression of atRSZ33 reveals its function in splicing and causes pleiotropic changes in development. *Mol Biol Cell, 14*(9), 3565-3577.
Kalyna, M., Lopato, S., Voronin, V., & Barta, A. (2006). Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res, 34*(16), 4395-4405.
Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., . . . Brown, J. W. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res, 40*(6), 2454-2469. doi: 10.1093/nar/gkr932
Kataoka, N., Yong, J., Kim, V. N., Velazquez, F., Perkinson, R. A., Wang, F., & Dreyfuss, G. (2000). Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. *Mol Cell, 6*(3), 673-682.
Kebaara, B. W., & Atkin, A. L. (2009). Long 3'-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in Saccharomyces cerevisiae. *Nucleic Acids Res, 37*(9), 2771-2778. doi: 10.1093/nar/gkp146
Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet, 11*(5), 345-355. doi: 10.1038/nrg2776
Koncz, C., Dejong, F., Villacorta, N., Szakonyi, D., & Koncz, Z. (2012). The spliceosome-activating complex: molecular mechanisms underlying the function of a pleiotropic regulator. *Front Plant Sci, 3*, 9. doi: 10.3389/fpls.2012.00009
Kornblihtt, A. R. (2005). Promoter usage and alternative splicing. *Curr Opin Cell Biol, 17*(3), 262-268. doi: 10.1016/j.ceb.2005.04.014
Kornblihtt, A. R. (2007). Coupling transcription and alternative splicing. *Adv Exp Med Biol, 623*, 175-189.
Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J., & Nogues, G. (2004). Multiple links between transcription and splicing. *RNA, 10*(10), 1489-1498. doi: 10.1261/rna.7100104
Kurihara, Y., Matsui, A., Hanada, K., Kawashima, M., Ishida, J., Morosawa, T., . . . Seki, M. (2009). Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proc Natl Acad Sci U S A, 106*(7), 2453-2458. doi: 0808902106 [pii]
10.1073/pnas.0808902106

Kyburz, A., Friedlein, A., Langen, H., & Keller, W. (2006). Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Mol Cell, 23*(2), 195-205. doi: 10.1016/j.molcel.2006.05.037

Lambermon, M. H., Simpson, G. G., Wieczorek Kirk, D. A., Hemmings-Mieszczak, M., Klahre, U., & Filipowicz, W. (2000). UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation. *EMBO J, 19*(7), 1638-1649.

Lareau, L. F., Green, R. E., Bhatnagar, R. S., & Brenner, S. E. (2004). The evolving roles of alternative splicing. *Curr Opin Struct Biol, 14*(3), 273-282. doi: 10.1016/j.sbi.2004.05.002

S0959440X04000764 [pii]

Lazar, G., & Goodman, H. M. (2000). The Arabidopsis splicing factor SR1 is regulated by alternative splicing. *Plant Mol Biol, 42*(4), 571-581.

Le Hir, H., Izaurralde, E., Maquat, L. E., & Moore, M. J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J, 19*(24), 6860-6869. doi: 10.1093/emboj/19.24.6860

Lewis, B. P., Green, R. E., & 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*(1), 189-192.

Lewis, J. D., & Izaurralde, E. (1997). The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem, 247*(2), 461-469.

Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C., & Mattaj, I. W. (1996). A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev, 10*(13), 1683-1698.

Li, Y., Chen, Z. Y., Wang, W., Baker, C. C., & Krug, R. M. (2001). The 3'-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAs in vivo. *RNA, 7*(6), 920-931.

Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochem J, 417*(1), 15-27. doi: BJ20081501 [pii]

10.1042/BJ20081501

Lopato, S., Forstner, C., Kalyna, M., Hilscher, J., Langhammer, U., Indrapichate, K., . . . 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*(42), 39989-39998.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J., & Barta, A. (1999). A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol Biol, 39*(4), 761-773.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A. R., & Barta, A. (1999). atSRp30, one of two SF2/ASF-like proteins from Arabidopsis thaliana, regulates splicing of specific plant genes. *Genes Dev, 13*(8), 987-1001.

Lopato, S., Waigmann, E., & Barta, A. (1996). Characterization of a novel arginine/serine-rich splicing factor in Arabidopsis. *Plant Cell, 8*, 2255-2264.

Lorkovic, Z. J., Wieczorek Kirk, D. A., Klahre, U., Hemmings Mieszczak, M., & Filipowicz, W. (2000). RBP45 and RBP47, two oligouridylate-specific hnRNP-like proteins interacting with poly(A)+ RNA in nuclei of plant cells. *RNA, 6*(11), 1610-1624.

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., & Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science, 327*(5968), 996-1000. doi: 10.1126/science.1184208

Marquez, Y., Brown, J. W., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res, 22*(6), 1184-1195. doi: 10.1101/gr.134106.111

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., . . . Wang, Z. (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics, 11*, 663. doi: 10.1186/1471-2164-11-663

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

Martinez-Contreras, R., Fisette, J. F., Nasim, F. U., Madden, R., Cordeau, M., & Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol, 4*(2), e21. doi: 10.1371/journal.pbio.0040021

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N., & Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev, 19*(13), 1512-1517. doi: 10.1101/gad.1302205

May, G. E., Olson, S., McManus, C. J., & Graveley, B. R. (2011). Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. *RNA, 17*(2), 222-229. doi: 10.1261/rna.2521311

McCracken, S., Lambermon, M., & Blencowe, B. J. (2002). SRm160 splicing coactivator promotes transcript 3'-end cleavage. *Mol Cell Biol, 22*(1), 148-160.

Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F., & Dietz, H. C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet, 36*(10), 1073-1078. doi: 10.1038/ng1429

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science, 338*(6114), 1593-1599. doi: 10.1126/science.1228186

Millevoi, S., Geraghty, F., Idowu, B., Tam, J. L., Antoniou, M., & Vagner, S. (2002). A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3'-end processing. *EMBO Rep, 3*(9), 869-874. doi: 10.1093/embo-reports/kvf173

Moore, R. C., & Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol, 8*(2), 122-128. doi: 10.1016/j.pbi.2004.12.001

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods, 5*(7), 621-628. doi: nmeth.1226 [pii]

10.1038/nmeth.1226

Muhlrad, D., & Parker, R. (1999). Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA, 5*(10), 1299-1307.

Nesic, D., & Maquat, L. E. (1994). Upstream introns influence the efficiency of final intron removal and RNA 3'-end formation. *Genes Dev, 8*(3), 363-375.

Neves, G., Zucker, J., Daly, M., & Chess, A. (2004). Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet, 36*(3), 240-246. doi: 10.1038/ng1299

Nicholson, P., Yepiskoposyan, H., Metze, S., Zamudio Orozco, R., Kleinschmidt, N., & Muhlemann, O. (2010). Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell Mol Life Sci, 67*(5), 677-700. doi: 10.1007/s00018-009-0177-1

Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends Plant Sci, 10*(9), 444-449. doi: 10.1016/j.tplants.2005.07.007

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet, 40*(12), 1413-1415. doi: ng.259 [pii]
10.1038/ng.259
Query, C. C., Moore, M. J., & Sharp, P. A. (1994). Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev, 8*(5), 587-597.
Rebbapragada, I., & Lykke-Andersen, J. (2009). Execution of nonsense-mediated mRNA decay: what defines a substrate? *Curr Opin Cell Biol, 21*(3), 394-402. doi: S0955-0674(09)00061-1 [pii]
10.1016/j.ceb.2009.02.007
Reddy, A. S. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol, 58*, 267-294.
Reddy, A. S., Marquez, Y., Kalyna, M., & Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell, 25*(10), 3657-3683. doi: 10.1105/tpc.113.117523
Reed, R. (1996). Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev, 6*(2), 215-220.
Reed, R. (2003). Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol, 15*(3), 326-331.
Reed, R., & Cheng, H. (2005). TREX, SR proteins and export of mRNA. *Curr Opin Cell Biol, 17*(3), 269-273.
Rehwinkel, J., Letunic, I., Raes, J., Bork, P., & Izaurralde, E. (2005). Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA, 11*(10), 1530-1544. doi: 10.1261/rna.2160905
Rigo, F., & Martinson, H. G. (2008). Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. *Mol Cell Biol, 28*(2), 849-862. doi: 10.1128/MCB.01410-07
Rigo, F., & Martinson, H. G. (2009). Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. *RNA, 15*(5), 823-836. doi: 10.1261/rna.1409209
Risso, G., Pelisch, F., Quaglino, A., Pozzi, B., & Srebrow, A. (2012). Regulating the regulators: serine/arginine-rich proteins under scrutiny. *IUBMB Life, 64*(10), 809-816. doi: 10.1002/iub.1075
Robberson, B. L., Cote, G. J., & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol, 10*(1), 84-94.
Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., . . . Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nat Methods, 7*(11), 909-912. doi: 10.1038/nmeth.1517
Ruby, S. W., & Abelson, J. (1991). Pre-mRNA splicing in yeast. *Trends Genet, 7*(3), 79-85.
Ruhl, C., Stauffer, E., Kahles, A., Wagner, G., Drechsel, G., Ratsch, G., & Wachter, A. (2012). Polypyrimidine tract binding protein homologs from Arabidopsis are key regulators of alternative splicing with implications in fundamental developmental processes. *Plant Cell, 24*(11), 4360-4375. doi: 10.1105/tpc.112.103622
Sanford, J. R., Gray, N. K., Beckmann, K., & Caceres, J. F. (2004). A novel role for shuttling SR proteins in mRNA translation. *Genes Dev, 18*(7), 755-768. doi: 10.1101/gad.286404
Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., . . . Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell, 101*(6), 671-684.
Schwartz, S., Meshorer, E., & Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol, 16*(9), 990-995. doi: 10.1038/nsmb.1659

Schwer, B., & Shuman, S. (1996). Conditional inactivation of mRNA capping enzyme affects yeast pre-mRNA splicing in vivo. *RNA, 2*(6), 574-583.
Sharp, P. A. (1994). Split genes and RNA splicing. *Cell, 77*, 805-815.
Shepard, P. J., & Hertel, K. J. (2008). Conserved RNA secondary structures promote alternative splicing. *RNA, 14*(8), 1463-1469. doi: 10.1261/rna.1069408
Shepard, P. J., & Hertel, K. J. (2009). The SR protein family. *Genome Biol, 10*(10), 242. doi: 10.1186/gb-2009-10-10-242
Shyu, A. B., Wilkinson, M. F., & van Hoof, A. (2008). Messenger RNA regulation: to translate or to degrade. *EMBO J, 27*(3), 471-481. doi: 7601977 [pii]
10.1038/sj.emboj.7601977
Silverman, I. M., Li, F., & Gregory, B. D. (2013). Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Sci, 205-206*, 55-62. doi: 10.1016/j.plantsci.2013.01.009
Simpson, C. G., Fuller, J., Maronova, M., Kalyna, M., Davidson, D., McNicol, J., . . . Brown, J. W. (2008). Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J, 53*(6), 1035-1048.
Staiger, D. (2001). RNA-binding proteins and circadian rhythms in Arabidopsis thaliana. *Philos Trans R Soc Lond B Biol Sci, 356*(1415), 1755-1759. doi: 10.1098/rstb.2001.0964
Staiger, D., & Brown, J. W. (2013). Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. *Plant Cell*. doi: 10.1105/tpc.113.113803
Staiger, D., Zecca, L., Wieczorek Kirk, D. A., Apel, K., & 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*(2), 361-371.
Stalder, L., & Muhlemann, O. (2008). The meaning of nonsense. *Trends Cell Biol, 18*(7), 315-321. doi: 10.1016/j.tcb.2008.04.005
Stauffer, E., Westermann, A., Wagner, G., & Wachter, A. (2010). Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant J, 64*(2), 243-255. doi: 10.1111/j.1365-313X.2010.04321.x
Sterner, D. A., Carlo, T., & Berget, S. M. (1996). Architectural limits on split genes. *Proc Natl Acad Sci U S A, 93*(26), 15081-15085.
Streitner, C., Koster, T., Simpson, C. G., Shaw, P., Danisman, S., Brown, J. W., & Staiger, D. (2012). An hnRNP-like RNA-binding protein affects alternative splicing by in vivo interaction with transcripts in Arabidopsis thaliana. *Nucleic Acids Res, 40*(22), 11240-11255. doi: 10.1093/nar/gks873
Stutz, F., Bachi, A., Doerks, T., Braun, I. C., Seraphin, B., Wilm, M., . . . Izaurralde, E. (2000). REF, an evolutionary conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export. *RNA, 6*(4), 638-650.
Sugliani, M., Brambilla, V., Clerkx, E. J., Koornneef, M., & Soppe, W. J. (2010). The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *Plant Cell, 22*(6), 1936-1946. doi: 10.1105/tpc.110.074674
Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., . . . Yaspo, M. L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science, 321*(5891), 956-960. doi: 10.1126/science.1160342
Sun, H., & Chasin, L. A. (2000). Multiple splicing defects in an intronic false exon. *Mol Cell Biol, 20*(17), 6414-6425.
Swartz, J. E., Bor, Y. C., Misawa, Y., Rekosh, D., & Hammarskjold, M. L. (2007). The shuttling SR protein 9G8 plays a role in translation of unspliced mRNA

containing a constitutive transport element. *J Biol Chem, 282*(27), 19844-19853. doi: 10.1074/jbc.M701660200
Talerico, M., & Berget, S. M. (1990). Effect of 5' splice site mutations on splicing of the preceding intron. *Mol Cell Biol, 10*(12), 6299-6305.
Talerico, M., & Berget, S. M. (1994). Intron definition in splicing of small Drosophila introns. *Mol Cell Biol, 14*(5), 3434-3445.
Tange, T. O., Damgaard, C. K., Guth, S., Valcarcel, J., & Kjems, J. (2001). The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *EMBO J, 20*(20), 5748-5758. doi: 10.1093/emboj/20.20.5748
Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcarcel, J., & Guigo, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol, 16*(9), 996-1001. doi: 10.1038/nsmb.1658
Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi: 10.1093/bioinformatics/btp120
Vagner, S., Ruegsegger, U., Gunderson, S. I., Keller, W., & Mattaj, I. W. (2000). Position-dependent inhibition of the cleavage step of pre-mRNA 3'-end processing by U1 snRNP. *RNA, 6*(2), 178-188.
Vagner, S., Vagner, C., & Mattaj, I. W. (2000). The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes Dev, 14*(4), 403-413.
Viegas, M. H., Gehring, N. H., Breit, S., Hentze, M. W., & Kulozik, A. E. (2007). The abundance of RNPS1, a protein component of the exon junction complex, can determine the variability in efficiency of the Nonsense Mediated Decay pathway. *Nucleic Acids Res, 35*(13), 4542-4551. doi: 10.1093/nar/gkm461
Wachter, A., Ruhl, C., & Stauffer, E. (2012). The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. *Front Plant Sci, 3*, 81. doi: 10.3389/fpls.2012.00081
Walters, B., Lum, G., Sablok, G., & Min, X. J. (2013). Genome-wide landscape of alternative splicing events in Brachypodium distachyon. *DNA Res, 20*(2), 163-171. doi: 10.1093/dnares/dss041
Wang, B. B., & Brendel, V. (2004). The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol, 5*(12), R102. doi: 10.1186/gb-2004-5-12-r102
Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., . . . Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature, 456*(7221), 470-476. doi: 10.1038/nature07509
Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., . . . Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res, 38*(18), e178. doi: 10.1093/nar/gkq622
Wang, Z., & Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA, 14*(5), 802-813. doi: 10.1261/rna.876308
Weatheritt, R. J., & Gibson, T. J. (2012). Linear motifs: lost in (pre)translation. *Trends Biochem Sci, 37*(8), 333-341. doi: 10.1016/j.tibs.2012.05.001
Welch, E. M., & Jacobson, A. (1999). An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. *EMBO J, 18*(21), 6134-6145. doi: 10.1093/emboj/18.21.6134
Will, C. L., & Luhrmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem, 386*(8), 713-724. doi: 10.1515/BC.2005.084
Will, C. L., & Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb Perspect Biol, 3*(7). doi: 10.1101/cshperspect.a003707

Wittmann, J., Hol, E. M., & Jack, H. M. (2006). hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay. *Mol Cell Biol, 26*(4), 1272-1287. doi: 10.1128/MCB.26.4.1272-1287.2006

Yepiskoposyan, H., Aeschimann, F., Nilsson, D., Okoniewski, M., & Muhlemann, O. (2011). Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA, 17*(12), 2108-2118. doi: 10.1261/rna.030247.111

Yoine, M., Ohto, M. A., Onai, K., Mita, S., & Nakamura, K. (2006). The lba1 mutation of UPF1 RNA helicase involved in nonsense-mediated mRNA decay causes pleiotropic phenotypic changes and altered sugar signalling in Arabidopsis. *Plant J, 47*(1), 49-62.

Zeng, C., & Berget, S. M. (2000). Participation of the C-terminal domain of RNA polymerase II in exon definition during pre-mRNA splicing. *Mol Cell Biol, 20*(21), 8290-8301.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., . . . Delledonne, M. (2010). Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. *Plant Physiol, 152*(4), 1787-1795. doi: 10.1104/pp.109.149716

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., . . . Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res, 20*(5), 646-654. doi: gr.100677.109 [pii]
10.1101/gr.100677.109

Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum Mol Genet, 7*(5), 919-932.

Zhang, X. H., Arias, M. A., Ke, S., & Chasin, L. A. (2009). Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA, 15*(3), 367-376. doi: 10.1261/rna.1498509

Zhong, X. Y., Wang, P., Han, J., Rosenfeld, M. G., & Fu, X. D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell, 35*(1), 1-10. doi: S1097-2765(09)00429-8 [pii]
10.1016/j.molcel.2009.06.016

Zhu, J., Mayeda, A., & Krainer, A. R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell, 8*(6), 1351-1361.

Zhu, W., Schlueter, S. D., & Brendel, V. (2003). Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. *Plant Physiol, 132*(2), 469-484. doi: 10.1104/pp.102.018101
132/2/469 [pii]

Appendix

Alternative splicing in plants—Coming of age.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W. (2012).

Trends Plant Sci. 17: 616–623.

Review

Alternative splicing in plants – coming of age

Naeem H. Syed[1], Maria Kalyna[2], Yamile Marquez[2], Andrea Barta[2] and John W.S. Brown[1,3]

[1] Division of Plant Sciences, University of Dundee at the James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK
[2] Max F. Perutz Laboratories, Medical University of Vienna, Dr Bohr-Gasse 9/3, A-1030 Vienna, Austria
[3] Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK

More than 60% of intron-containing genes undergo alternative splicing (AS) in plants. This number will increase when AS in different tissues, developmental stages, and environmental conditions are explored. Although the functional impact of AS on protein complexity is still understudied in plants, recent examples demonstrate its importance in regulating plant processes. AS also regulates transcript levels and the link with nonsense-mediated decay and generation of unproductive mRNAs illustrate the need for both transcriptional and AS data in gene expression analyses. AS has influenced the evolution of the complex networks of regulation of gene expression and variation in AS contributed to adaptation of plants to their environment and therefore will impact strategies for improving plant and crop phenotypes.

Frequency and consequences of alternative splicing (AS)

AS (see Glossary) produces multiple mRNAs from the same gene through variable selection of splice sites during pre-mRNA splicing. It plays a key regulatory role in modulating gene expression during development and in response to environmental signals [1–4]. Regulation of AS in different cell types and under different conditions depends on sequence elements in pre-mRNAs and the interactions of RNA-binding proteins which vary in their concentration and activity. The phenotype of a cell is determined by transcriptional, post-transcriptional, and post-translational networks, which include as a key component the regulated AS of thousands of genes. In humans, where >95% of genes are alternatively spliced, extensive protein diversity is largely a result of AS [5]. Next generation sequencing has revolutionized research into AS and global mapping of human splicing regulatory proteins to their target RNA sequences has led to the development of a splicing code that will allow prediction of tissue-dependent AS [6]. AS not only contributes to proteome diversity but also can generate truncated proteins that are potentially regulatory or detrimental to the cell. It also plays a role in gene expression by regulating transcript levels through production of isoforms which are degraded by the nonsense-mediated decay (NMD) pathway.

In plants, computational analyses of AS events based on expressed sequence tags and more recently high-throughput transcriptome sequencing have examined the frequency of occurrence of AS in different species [42% and 33% of intron-containing genes in *Arabidopsis* (*Arabidopsis thaliana*) and rice (*Oryza sativa*), respectively] and of different types of AS events [7–10]. In *Arabidopsis*, the frequency of occurrence has risen significantly over the past 10 years (Figure 1). Recently, an extensive RNA-seq analysis has significantly increased the observed frequency of AS in *Arabidopsis* to more than 61% of intron-containing genes showing AS. This estimate of 61% of AS is based on analysis of plants grown under normal growth conditions [11] and it is likely that this level will increase further as different tissues at various developmental stages and growth conditions are analyzed [11]. Of the most common types of AS (Figure 2b), intron retention (IR) has been shown to be the most frequent AS event in plants [7–10]. However, some IR events were recently shown to be more likely to represent partially spliced transcripts due to their

Glossary

Alternative 5′ and 3′ splice sites: use of alternative splice sites at either end of the intron (in the intron or exon sequence) adds or removes sequences.
Alternative splicing: precursor mRNAs (pre-mRNAs) are spliced differently to generate different mRNA isoforms.
Exon skipping/inclusion: an exon can be removed in a single splicing event or included by two splicing events. Such exons are called alternative exons or cassette exons.
Heterogeneous nuclear ribonucleoproteins (hnRNP): RNA-binding proteins which bind RNA in the cell.
Intron retention (IR): one or more introns is/are not removed from a pre-mRNA. It is often difficult to determine whether intron retention is due to DNA contamination or partial splicing of pre-mRNAs at the time of RNA extraction or are actively retained.
Micro-protein (miP): micro-proteins are usually generated by translation of a transcript containing a premature termination codon, lack one or more functional domains, and have fewer than 100 amino acids [82].
Nonsense-mediated decay (NMD): a cellular quality control mechanism that recognizes mRNA transcripts containing PTCs and targets them for degradation.
Polypyrimidine tract-binding protein (PTB): an hnRNP protein which binds pyrimidine-rich sequences in RNA to regulate alternative splicing and other mRNA biogenesis processes.
Precursor messenger RNA (pre-mRNA): the primary transcript of a gene which is processed to mRNA.
Premature termination codon (PTC): a translational stop codon found in transcripts upstream of the authentic stop codon; PTCs can be generated by mutations in DNA, errors in transcription or splicing, or alternative splicing.
Serine/arginine-rich (SR) proteins: constitutive and alternative splicing factors containing RNA-binding motifs and a RS-rich domain.
Small-interfering peptides (siPEPs): short peptides which interfere with cellular processes.

Corresponding author: Brown, J.W.S. (j.w.s.brown@dundee.ac.uk), (John.Brown@hutton.ac.uk).

 1360-1385/$ – see front matter http://dx.doi.org/10.1016/j.tplants.2012.06.001

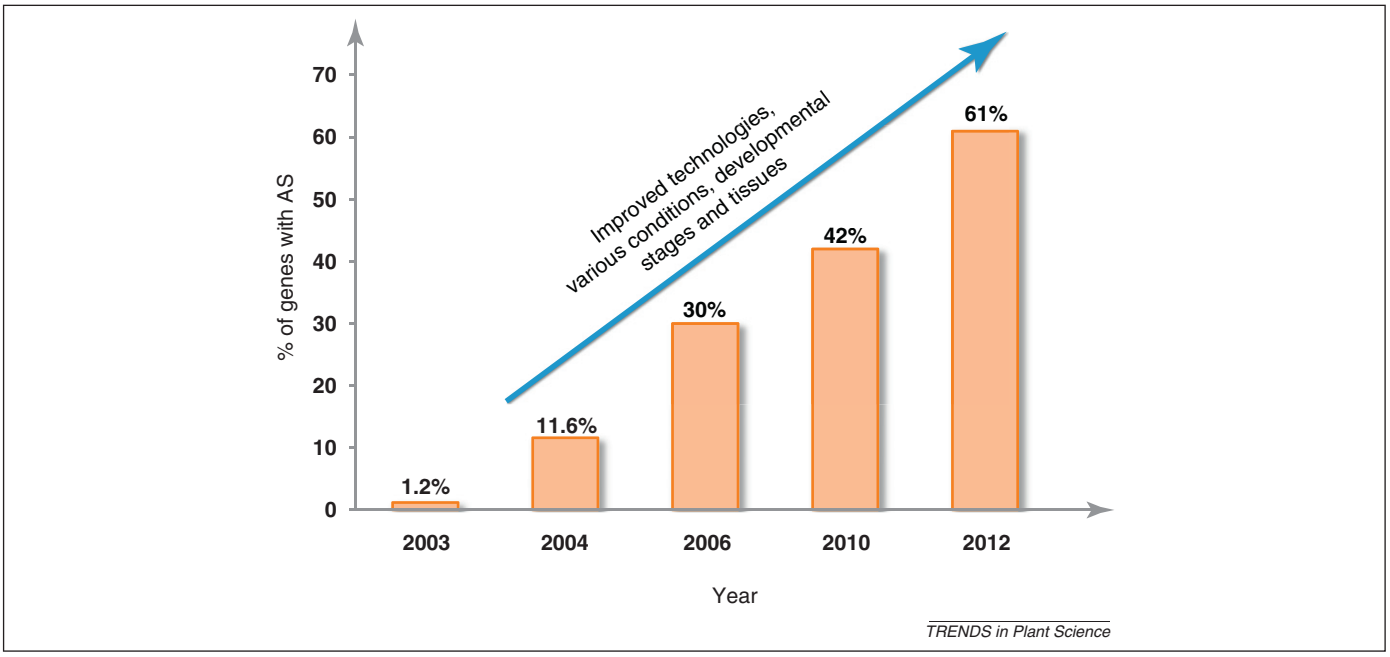


Figure 1. Increasing frequency of occurrence of alternative splicing (AS) in *Arabidopsis* with time. In 2003, a study using EST (expressed sequence tag) libraries estimated that only 1.2% of the genes in *Arabidopsis* undergo AS [101]. Subsequently, greater coverage of ESTs and cDNAs libraries allowed the discovery of many more AS events (2004–2006, [7,102–104]). The advent of high-throughput technologies [9,11] has resulted in significant increases in the frequency of AS (almost 60-fold over the past 10 years).

low abundance [11]. In addition, in the genome-wide analysis above, IR was still the most frequent AS event (40%) but it only occurred in assembled AS transcripts of 23% of the genes providing a more reasonable estimate of the impact of IR to AS plants (Figure 2c) [11]. More importantly, 51% of intron-containing genes utilize alternative 5′ or 3′ splice sites or exon skipping events which can affect the protein coding sequence or generate unproductive mRNAs to affect transcript levels [11]. The widespread occurrence of AS and the range of functional gene groups which it affects supports an essential role for AS in plant development, physiology, metabolism, and responses to environmental conditions and pathogens, all of which have important consequences on plant/crop phenotypes [7,12–18].

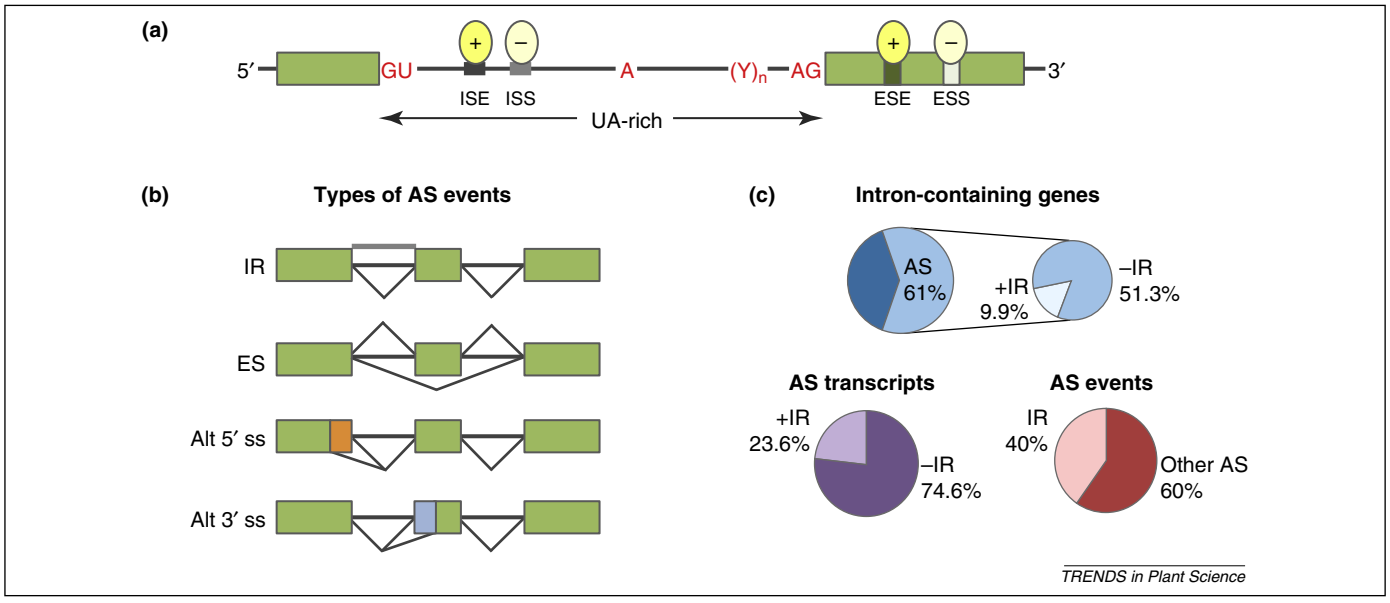


Figure 2. Main types of alternative splicing (AS) events and frequency in *Arabidopsis*. **(a)** Splicing of pre-mRNA is directed by *cis* elements which include splice sites, branch point, and polypyrimidine tract sequences. Selection of alternative splice sites is affected by *trans*-acting factors binding to auxiliary exonic and intronic *cis* elements, termed splicing enhancers and silencers. **(b)** Types of AS events. **(c)** Frequency of occurrence of intron retention in *Arabidopsis*. Intron retention is the most frequent AS event in *Arabidopsis* (40%) but its contribution to transcript diversity is much lower [11]. Of the 61% of *Arabidopsis* intron-containing genes with AS, 51% produce AS transcripts which do not involve intron retention (–IR). Among alternatively spliced transcripts, 23.6% contain one or more retained introns (+IR), whereas the rest (74.6%) are produced by other AS events. Colored boxes, exons; lines, introns; GU, 5′ splice site which includes highly conserved GU dinucleotide; AG, 3′ splice site which includes highly conserved AG dinucleotide; A, branch point adenosine; $(Y)_n$, polypyrimidine tract; ovals, positive and negative splicing regulators; carets, splicing events; thick gray line, unspliced (retained) intron. Abbreviations: ESE, exonic splicing enhancers; ESS, exonic splicing silencers; ISE, intronic splicing enhancers; ISS, intronic splicing silencers; Alt 3′ ss, alternative 3′ splice sites; Alt 5′ ss, alternative 5′ splice sites; ES, exon skipping; IR, intron retention.

Regulation of alternative splicing

Assembly of the spliceosome during intron removal and ligation of exons is directed by sequence features of the pre-mRNA. The *cis* sequences in the pre-mRNA include splice sites, branch point, polypyrimidine tract, enhancer, and suppressor sequences (Figure 2a). In addition, it is well documented in plants that UA-richness of introns contributes to their recognition and is essential for efficient splicing. The splice sites situated at the transition between UA-rich introns and GC-rich exons are preferentially selected for splicing (for review see [19]). The compositional bias for UA-richness has been always considered as a distinguishing feature of plant introns; however, recently it has been shown also for animals that exons have higher GC content than introns [20,21]. Regulation of AS depends on *cis* signals and their recognition by *trans*-acting splicing factors. Serine/arginine-rich (SR) and heterogeneous nuclear RNP (hnRNP) proteins function as constitutive and AS splicing factors controlling splice site choice in a concentration-dependent manner [22]. SR proteins are highly conserved in metazoa and plants and, typically, contain one or two RNA recognition motifs (RRMs) and a C-terminal domain (CTD) rich in serine and arginine residues (RS domain). Interestingly, plants possess plant-specific SR proteins and in general they have almost a double number of SR proteins of that in nonphotosynthetic organisms [23]. Many of them have different spatiotemporal expression patterns, implicating diverse target specificities and biological functions [24]. hnRNP proteins, by contrast, constitute a structurally diverse group of RNA-binding proteins associated with nascent pre-mRNA molecules and, in addition to their role in splicing, are involved in a variety of molecular processes [25]. SR and hnRNP proteins bind splicing signals and intronic and exonic enhancer/silencer sequences and through multicomponent interactions with other splicing factors (including cell- and tissue-specific factors) determine splice site selection and where the spliceosome assembles. In general, SR proteins promote splicing and the hnRNP proteins inhibit splice site selection. However, both SRs and hnRNPs can also have opposite functions with, for example, SRSF10 (also known as SRp38) [26] being a negative regulator and polypyrimidine tract-binding protein (PTB; also known as hnRNPI) being a positive regulator [27].

Variation in abundance and activity of splicing factors determines the AS profiles of target genes and therefore their differential expression under different growth conditions and during development [28–31]. Importantly, different growth conditions modulate AS of SR and hnRNP genes causing dynamic changes in the splicing factor profile, further impacting expression of target genes (Figure 3). For example, AS of many *Arabidopsis* SR genes is affected by temperature, light, salt, hormones, etc. [30–32]. In addition, the activity or localization of SR proteins can be affected by phosphorylation [4,33,34] and protein kinases which phosphorylate plant SR proteins have been identified [35–37]. Not surprisingly therefore, SR protein overexpression and knockout lines show a variety of developmental and growth phenotypes, demonstrating the importance of these proteins to normal growth and development and broad effects on gene expression [28]. The best-studied plant

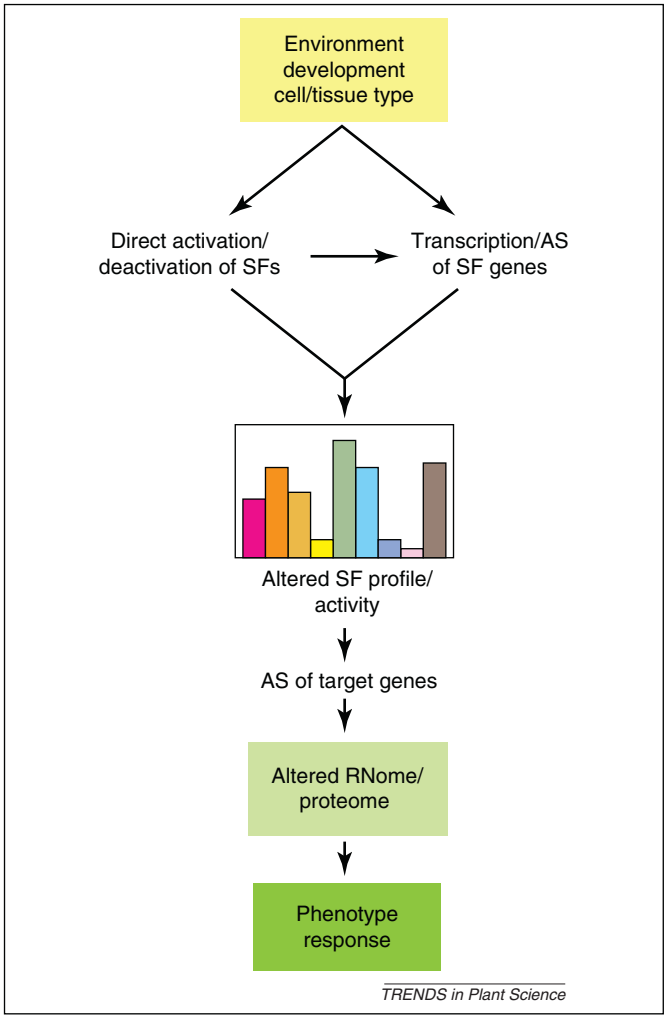


Figure 3. Dynamic regulation of expression by alternative splicing. Developmental or environmental cues activate signaling pathways which can directly modulate splicing factor activity by post-translational modifications, relocalization, etc. Signaling also directs changes in transcription of splicing factor genes and alternative splicing of these genes changes the abundance, composition, and activity of the splicing factor population. Expression of other target genes including transcription factors is also modulated by alternative splicing responding to the dynamic changes in splicing factor profiles. Changes in the proteome feedback to transcription and alternative splicing (and other post-transcriptional mechanisms) ultimately generating the cellular and organismal phenotype and response. Boxes of colored bars, abundance and/or composition of SFs. Abbreviations: AS, alternative splicing; SF, splicing factor.

hnRNP proteins to date are the *Arabidopsis* orthologs of the animal negative splicing regulator, PTB [38], and the glycine-rich RNA-binding proteins, GRP7 and GRP8, components of a slave oscillator coupled to the circadian clock [39–41]. The *Arabidopsis* PTBs auto- and crossregulate the AS within the family [38]. GRP7 and GRP8 also autoregulate their own AS and crossregulate each other's AS to generate unproductive mRNAs which are targeted by NMD to reduce mRNA and protein levels [39,41]. Finally, the nuclear cap-binding complex (the CBC) consists of two subunits, AtCBP20 and AtCBP80, and AtCBP20 contains a canonical RNA-binding domain (RBD). Mutation of these subunits showed the CBC to be involved in AS of at least some *Arabidopsis* genes and to preferentially influence AS of the first intron, particularly at the 5′ splice site [42].

Although the abundance and activity of splicing factors determines the AS profiles of downstream genes, very few

examples of the biological relevance of AS-derived protein isoforms of splicing factors have been published. Functional diversity of AS isoforms was elegantly demonstrated when two isoforms of SR45 (which differ by only eight amino acids and include putative phosphorylation sites), differentially complemented petal or root developmental phenotypes in the *sr45* mutant [43]. Hence, isoforms with very similar sequences can have substantially different morphological outcomes which will reflect a host of gene expression changes in different organs of the plant.

Epigenetic control of alternative splicing

An extensive body of evidence from human and yeast shows that splicing is often coupled to transcription [44–46]. The CTD of RNA polymerase II serves as a landing pad for recruitment of proteins involved in capping, splicing, polyadenylation, and export [47–52]. The rate of transcription elongation by RNA polymerase II may affect splice site choice and thereby AS outcomes [45,46]. A slower rate of transcription tends to favor inclusion of weak upstream exons before the splicing complex is committed to splicing of a stronger downstream exon [45,46]. Importantly, efficiency of the splicing process may also influence the rate of transcription elongation, where, for example, transient pausing of RNAP II at the 3′ splice site of an intron coincides with the appearance of spliced product [53]. The rate of transcript elongation may depend on the chromatin state [50]. For example, nucleosome occupancy varies along a gene with GC-rich exons being relatively nucleosome-rich compared with GC poor introns [20,21,52,54] and transcription through nucleosome-rich regions with compact chromatin tends to be slower [50]. Furthermore, nucleosome occupancy is also lower in alternatively spliced exons compared with constitutively spliced exons [20]. The relations between chromatin state, nucleosome occupancy, RNAP II elongation rates, splicing efficiency, and AS outcomes are key to understanding AS regulation. Indeed, a direct link between histone modifications and AS was demonstrated recently [55]. The splicing of two PTB-dependent mutually exclusive exons in the human fibroblast growth factor receptor 2 (FGFR2) gene depended on histone modifications H3K36me3 and H3K4me1 acting in one direction, and H3K27me3, H3K4me3, and H3K9me1 acting in the opposite direction. The changes in chromatin state are read by the chromatin-binding adapter protein MRG15 which then recruits the splicing factor PTB to the pre-mRNA and affects splicing outcomes [45,46,55]. In plants, the direct link between AS and either chromatin state or RNAP II elongation rates (transcription) has not yet been demonstrated. However, clearly the impact of environmental cues and signaling pathways on chromatin and transcription to generate different AS variants has the potential to further our understanding how plants respond to their ever-changing environment.

Alternative splicing affects protein complexity and transcript levels and stability

The different forms of AS (Figure 2b) and, in particular, alternative 5′ and 3′ splice site selection and exon skipping often lead to changes in protein sequences from the inclusion or removal of a few amino acids (see above for SR45 [43]) to large regions of proteins affecting protein domains, or generating changes in N-terminal or C-terminal regions [56,57]. Different protein isoforms have the potential for differential functions as highlighted by several recent studies. For example, cold-induced sweetening is a serious problem in potatoes where starch is converted to glucose and fructose by vacuolar acid invertase: lines showing resistance to cold-induced sweetening have higher expression of two splice variants (INH2α and INH2β) of the invertase inhibitor gene (INH2) [58]. The flavin-dependent monooxygenase gene, *YUCCA4*, involved in auxin biosynthesis, undergoes tissue-specific AS to generate isoforms with different intracellular localization. One isoform is expressed in all tissues and is distributed throughout the cytosol, whereas a second is restricted to flowers and is attached to the endoplasmic reticulum [59]. *In silico* analysis of MADS-box MIKC-type (MADS, Intervening, Keratin-like and C-terminal domain) transcription factors in *Arabidopsis* predicted protein isoforms which affect dimerization properties or higher order protein complex formation [57]. The potential for AS to influence function was shown by the differential effects on flowering time and floral development of overexpression of isoforms of *SHORT VEGETATIVE PHASE*, consistent with their different protein–protein interactions [57]. This systematic analysis of a large gene family illustrates the potential of AS to affect key protein domains and function as well as the impact of AS in the evolution of gene families and protein interaction networks.

AS can regulate mRNA levels through the production of AS isoforms containing premature termination codons (PTCs) which are targeted for degradation by NMD [60,61]. Plants possess orthologs of the key eukaryotic NMD proteins, UPF1, UPF2, UPF3, and SMG-7 (but not SMG-1, SMG-5, or SMG-6) and these have been shown to be involved in degrading mRNAs with PTCs [62–68]. Rules for NMD in plants have been established mainly by studying mutations in a small number of model transcripts. The mechanisms of recognition of NMD substrates in plants appear to be fundamentally similar to those in other eukaryotes relying on the distance from a PTC to the 3′ end of the transcript (long 3′UTR) or downstream splice junctions (splicing-dependent) [18,64,69–71]. Recently, by analyzing a large population of endogenous *Arabidopsis* transcripts, coupled AS, and NMD has been shown to be a widespread mechanism for regulating gene expression with 11–18% of alternatively spliced transcripts being turned over by NMD [18]. This study also showed that transcripts containing PTCs which are NMD substrates are often readily detectable and can contribute significantly to steady-state transcript levels of genes. In addition, some PTC-containing transcripts were not turned over by NMD. For example, some transcripts containing retained introns or parts of introns were unaffected in NMD mutants, suggesting that not all NMD triggering signals or transcript arrangements are understood [18]. Thus, the generation of unproductive AS transcripts can influence the levels of functional mRNAs (full-length protein coding), as has been observed in the regulation of human SR and

hnRNP proteins through AS and ultraconserved elements [72–74]. Similarly in plants, SR and PTB genes are regulated by AS [28,38,75] which gives rise to PTC-containing transcripts, suggesting a regulatory function via unproductive mRNAs.

Alternative splicing in the plant circadian clock

Regulation of expression by AS generating unproductive transcripts has recently been demonstrated for core circadian clock genes in *Arabidopsis*. Circadian clocks have approximately 24-h rhythms and allow organisms to anticipate the day–night cycle and coordinate their genetic, biochemical, and physiological responses [76–78]. Core clock gene expression is regulated at multiple different levels: transcription, protein degradation, and modification with AS being an emerging theme in regulation of the clock [77,78]. Until recently, examples of AS in clock genes and their functional significance were rare. For example, an IR event in CCA1 was conserved in at least four plant species and levels of IR-containing transcripts increased in high light and decreased in the cold [9]. In addition, a mutant in PROTEIN ARGININE METHYL TRANSFERASE 5 (PRMT5) showed a longer circadian period and dramatic changes in the levels of unproductive AS transcripts of PRR9 [79,80]. Recently, extensive AS in the majority of the core clock genes in *Arabidopsis* was identified and dynamic changes in AS profiles for many AS events were observed in response to changes in temperature and particularly to lower temperatures [81]. AS events were either induced by low temperatures or increased in abundance to 10–50% of the transcripts; the majority of these events were nonproductive resulting in a reduction of functional mRNAs potentially impacting protein levels [81]. Furthermore, the partially redundant gene pairs, LHY and CCA1, and PRR7 and PRR9 behaved differently with respect to AS, implying functional differences between these related genes. Thus, temperature-associated AS modulating the balance between productive and nonproductive mRNA isoforms is an additional mechanism involved in the operation and control of the plant circadian clock.

Alternative splicing and regulation by small interfering peptides/micro-proteins

The fate of alternatively spliced transcripts containing PTCs is expected to be degradation by the NMD pathway but some PTC-containing transcripts are stable and appear to avoid the NMD machinery [18]. PTC-containing transcripts also have the potential to be translated into truncated proteins or peptides. In plants, intron-containing mRNA transcripts were found associated with polysomes and recently ribosome profiling in mouse has identified novel upstream open reading frames (ORFs) and ORFs in long noncoding RNAs [82,83]. In animal and plant systems, small interfering peptides (siPEPs) or micro-proteins (miPs) named after their analogy with siRNAs and miRNAs have been described [84,85]. Such peptides can have altered functionality by only containing particular domains (e.g., DNA binding, transcriptional activators) and can act as both positive and negative regulators and affect regulatory feedback loops [86]. For example, in animals, an AS isoform generates a miP of the ETS1 transcription factor which regulates growth and development responses, lacks the transactivation domains, and interacts physically with ETS1 blocking ETS1-mediated expression of target genes in a dominant negative manner [86–88]. Similarly, miP AS protein isoforms of the animal transcription factor MEIS2 also interact in a dominant negative manner [89]. Interestingly, one of the splice variants (MEIS2E) is structurally similar to a plant protein 'KNATM', which is a member of the TALE homeodomain transcription regulators (controlling meristem formation, organ position and morphogenesis, and some aspects of reproductive phase) [86]. KNATM, however, lacks a homeodomain and by forming nonfunctional heterodimers with the BELL TALE protein regulates leaf pattern [90,91]. It is intriguing that a protein generated via AS in animals appears to exist as a miP equivalent in plants.

Genome-wide analysis of AS in *Arabidopsis* suggested that 78% of alternative transcripts introduced in-frame PTCs [9], providing a huge potential for production of miPs. Examples of miPs in plants with functional consequences are rare but a recent study showed that the *Arabidopsis* transcription factor gene, *IDD14*, produces a splice variant (IDD14β) which lacks the DNA-binding domain but interacts with the functional IDD14α isoform to produce heterodimers. The IDD14α/β heterodimer has reduced binding affinity for the promoter of the Qua-Quine Starch (QQS) gene which regulates starch accumulation by initiating starch degradation [84]. Starch accumulation is one response of plants to cold and as the IDD14β splice variant is only expressed under cold conditions, starch degradation is reduced providing an AS/miP-dependent strategy for maintaining starch reserves at low temperatures. Interestingly, the core clock proteins CCA1 and LHY can both homo- and heterodimerize [92,93], and different combinations have different binding affinity to their target sequences [94]. It is interesting to speculate whether the extensive production of PTC-containing transcripts in core clock genes by AS could add a further level of regulation by miPs. Furthermore, siPEPs/miPs offer another mechanism to modify or knock-down expression of endogenous genes. Artificial siPEPs/miPs encoding dimerization domains could be transformed into plants to reduce the activity of a target gene. As they function at the protein level and depend on homo- or heterodimerization, this should improve specificity and reduce off-target silencing [85].

Alternative splicing diversity in ecotypes and polyploids

Extensive AS occurs under altered growth or stress conditions in plants. Similarly, extensive variation in AS is expected in diverse ecotypes adapted to very different climates thereby achieving environmental and phenotypic plasticity. To address such diversity, genomic and transcriptomic sequencing is being performed on geographically and phenotypically diverse accessions of *Arabidopsis*. Sequencing of the genomes and transcriptomes of 18 *Arabidopsis* accessions identified extensive single nucleotide polymorphism and indel variation among the genotypes [95]. When compared with Col-0 (TAIR10) one-third of protein coding genes were disrupted/altered in at least

one accession although re-annotation restored coding potential in most cases. Sequence variations affected translation start and stop sites, introduced PTCs, or changed the frame of the coding sequence, or potentially generated protein isoforms in different accessions. Two-thirds of 2572 genes with disrupted splice sites when compared to TAIR10 had new splice sites and a quarter of these sites were close to the splice sites in Col-0 [95]. Clearly, naturally occurring sequence variation can disrupt splice sites and RNA-binding motifs for splicing factors. Such mutations impact protein expression and activity and provide a basis for selection for adaptation of different ecotypes to their environments [7,8,10,57].

Extensive duplication or polyploidization has occurred in the evolutionary history of many plant species [96]. Theoretically, such events could generate immediate and drastic changes in the abundance, composition, and activity of splicing factors which in turn could affect splice site choice among variable analogous splice site sequences. Further mutation, gene loss, or changes in expression or protein functionality provide the basis for selection and continued evolution of the species. Recently, AS patterns have been studied among natural and synthetic polyploids of *Brassica napus*. Interestingly, two independently synthesized lines showed parallel loss of AS events after polyploidy [97]. This is intriguing because it shows that even in two independent events of polyploidy, using the same species, results in identical pattern of AS loss, pointing towards a non-random response of the so-called 'genomic shock' after two genomes physically interact with each other. The same study also showed that 26–30% of the duplicated genes show changes in AS, compared with the parents, with some showing organ specificity or response to abiotic stress [97]. It is likely that AS has played an important role in the evolution and adaptation of cultivated crops to different environmental conditions and niches, because many crop species are polyploids and this whole area will be one of great interest in the future, particularly with the power of high-throughput sequencing.

Evolutionary aspects of alternative splicing factor diversity

The *Arabidopsis* genome encodes 18 SR proteins which is nearly double the number found in humans. At least 12 of the 18 SR genes are located in duplicated genomic regions and the current data indicate that most of them have different spatiotemporal expression patterns, suggesting functional diversification [24]. In different plant lineages, the number of paralogous SR genes is highly variable. For example, the plant-specific RS subfamily of SR proteins is encoded by four genes in *Arabidopsis*, two in rice, at least five in *Pinus taeda* and by one gene both in *Physcomitrella patens* and *Chlamydomonas reinhardtii* [98]. Differential or common, redundant functions of SR paralogs in different species remain to be determined. However, it is interesting that several SR paralogs and orthologs are regulated by AS events conserved from unicellular green algae to land plants. These events occur in the analogous long introns situated in the RRM coding regions [98,99] and, moreover, they involve unusually highly conserved sequences around alternative splice sites [98], suggesting an important biological function of such regulation. A recent systematic survey of SR genes in 27 eukaryotic genomes showed that flowering plants on average possess nearly double the number of SR genes than nonphotosynthetic species [23]. Moreover, most of the plant SR genes are under purifying selection, ensuring that paralogous genes which originated due to the duplication events maintain their structure and function, whereas redundancy is reduced via diversification of gene expression [23].

Future challenges in alternative splicing research in plants

AS research in plants has made substantial progress in the past 4–5 years. The ever-increasing number of plant genes with AS and the processes in which they are involved point to the importance of understanding the mechanisms and regulation of AS and the functions of AS. The functional impact of AS is one of the most important questions – this is largely due to the relatively small number of examples of AS for which differential functions have been demonstrated for different protein isoforms. However, we draw a parallel with research to discover the extent of AS in plants. Around 10 years ago the first estimate was only 1.2% of plant genes showing AS! With massively improved technologies this number has now grown to more than 61% (Figure 1). Similarly, the increasing number of plant genome sequences and the generation of vast transcriptome data will allow computational analyses to identify conservation of AS events across species and tissue-, developmental-stage and environment-specific regulation of AS providing evidence of functionality. The number of functional examples of AS, whether at the mRNA transcript stability level or protein function, continues to grow and in turn is stimulating wider interest in AS in the plant community. High-throughput sequencing will also address dynamic changes in AS in development and under different environmental conditions and stresses, and how variation in AS patterns in different ecotypes and polyploids contributes to plasticity and adaptation of plant species. Furthermore, we need to understand how signaling pathways affect splicing factor activity directly or via chromatin modification and how transcriptional and AS networks interact [100]. AS is a major mechanism by which plants modulate and fine-tune expression of their genes. The next 5 years will see an explosion of knowledge of the functional significance of AS and understanding its contribution to the complexity of gene expression will offer new opportunities in approaches to modifying plant function for improved phenotypes.

Acknowledgments

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/G024979/1 (ERA-NET Plant Genomics (PASAS)]; the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS); the Austrian Science Fund (FWF) [SFB 1710, 1711; DK W1207; ERA-NET Plant Genomics (PASAS) I254; SFB RNAreg F43-P10]; the Austria Genomic Program (GENAU III) [ncRNAs]; and the EU FP6 Program Network of Excellence on Alternative Splicing (EURASNET) [LSHG-CT-2005-518238].

References

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

2 Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107

3 Lareau, L.F. *et al.* (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* 14, 273–282
4 Stamm, S. *et al.* (2005) Function of alternative splicing. *Gene* 344, 1–20
5 Pan, Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415
6 Barash, Y. *et al.* (2010) Deciphering the splicing code. *Nature* 465, 53–59
7 Wang, B.B. and Brendel, V. (2006) Genome-wide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7175–7180
8 Wang, B.B. *et al.* (2008) Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol.* 8, 17
9 Filichkin, S.A. *et al.* (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20, 45–58
10 Barbazuk, W.B. *et al.* (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.* 18, 1381–1392
11 Marquez, Y. *et al.* (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* 22, 1184–1195
12 Ali, G.S. and Reddy, A.S. (2008) Regulation of alternative splicing of pre-mRNAs by stresses. *Curr. Top. Microbiol. Immunol.* 326, 257–275
13 Chen, F.C. *et al.* (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
14 Halterman, D.A. *et al.* (2003) Powdery mildew-induced *Mla* mRNAs are alternatively spliced and contain multiple upstream open reading frames. *Plant Physiol.* 131, 558–567
15 Lorkovic, Z.J. (2009) Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci.* 14, 229–236
16 Quesada, V. *et al.* (2003) Autoregulation of *FCA* pre-mRNA processing controls *Arabidopsis* flowering time. *EMBO J.* 22, 3142–3152
17 Zhu, J. *et al.* (2007) Interplay between cold-responsive gene regulation, metabolism and RNA processing during plant cold acclimation. *Curr. Opin. Plant Biol.* 10, 290–295
18 Kalyna, M. *et al.* (2011) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* 40, 2454–2469
19 Lorkovic, Z.J. *et al.* (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci.* 5, 160–167
20 Schwartz, S. *et al.* (2009) Chromatin organization marks exon–intron structure. *Nat. Struct. Mol. Biol.* 16, 990–995
21 Tilgner, H. *et al.* (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* 16, 996–1001
22 Matlin, A.J. *et al.* (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* 6, 386–398
23 Richardson, D.N. *et al.* (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
24 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
25 Martinez-Contreras, R. *et al.* (2007) hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.* 623, 123–147
26 Feng, Y. *et al.* (2008) Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator. *Nat. Struct. Mol. Biol.* 15, 1040–1048
27 Xue, Y. *et al.* (2009) Genome-wide analysis of PTB–RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell* 36, 996–1006
28 Barta, A. *et al.* (2008) Plant SR proteins and their functions. *Curr. Top. Microbiol. Immunol.* 326, 83–102
29 Lazar, G. and Goodman, H.M. (2000) The *Arabidopsis* splicing factor SR1 is regulated by alternative splicing. *Plant Mol. Biol.* 42, 571–581
30 Palusa, S.G. *et al.* (2007) Alternative splicing of pre-mRNAs of *Arabidopsis* serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J.* 49, 1091–1107
31 Tanabe, N. *et al.* (2007) Differential expression of alternatively spliced mRNAs of *Arabidopsis* SR protein homologs, at*SR30* and at*SR45*a, in response to environmental stress. *Plant Cell Physiol.* 48, 1036–1049
32 Duque, P. (2011) A role for SR proteins in plant stress responses. *Plant Signal. Behav.* 6, 49–54
33 Stamm, S. (2002) Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum. Mol. Genet.* 11, 2409–2416
34 Stamm, S. (2008) Regulation of alternative splicing by reversible protein phosphorylation. *J. Biol. Chem.* 283, 1223–1227
35 de la Fuente van Bentem, S. *et al.* (2006) Phosphoproteomics reveals extensive in vivo phosphorylation of *Arabidopsis* proteins involved in RNA metabolism. *Nucleic Acids Res.* 34, 3267–3278
36 de la Fuente van Bentem, S. *et al.* (2003) The subcellular localization of plant protein phosphatase 5 isoforms is determined by alternative splicing. *Plant Physiol.* 133, 702–712
37 Savaldi-Goldstein, S. *et al.* (2003) Alternative splicing modulation by a LAMMER kinase impinges on developmental and transcriptome expression. *Plant Cell* 15, 926–938
38 Stauffer, E. *et al.* (2010) Polypyrimidine tract-binding protein homologues from *Arabidopsis* underlie regulatory circuits based on alternative splicing and downstream control. *Plant J.* 64, 243–255
39 Schoning, J.C. *et al.* (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
40 Schuttpelz, M. *et al.* (2008) Changes in conformational dynamics of mRNA upon AtGRP7 binding studied by fluorescence correlation spectroscopy. *J. Am. Chem. Soc.* 130, 9507–9513
41 Schoning, J.C. *et al.* (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
42 Raczynska, K.D. *et al.* (2010) Involvement of the nuclear cap-binding protein complex in alternative splicing in *Arabidopsis thaliana*. *Nucleic Acids Res.* 38, 265–278
43 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
44 Carrillo Oesterreich, F. *et al.* (2010) Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol. Cell* 40, 571–581
45 Luco, R.F. *et al.* (2011) Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26
46 Luco, R.F. and Misteli, T. (2011) More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.* 21, 366–372
47 de Almeida, S.F. and Carmo-Fonseca, M. (2008) The CTD role in cotranscriptional RNA processing and surveillance. *FEBS Lett.* 582, 1971–1976
48 Fong, N. *et al.* (2003) A 10 residue motif at the C-terminus of the RNA pol II CTD is required for transcription, splicing and 3′ end processing. *EMBO J.* 22, 4274–4282
49 Goldstrohm, A.C. *et al.* (2001) The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol. Cell. Biol.* 21, 7617–7628
50 Howe, K.J. (2002) RNA polymerase II conducts a symphony of pre-mRNA processing activities. *Biochim. Biophys. Acta* 1577, 308–324
51 McCracken, S. *et al.* (1997) The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* 385, 357–361
52 Schwartz, S. and Ast, G. (2010) Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO J.* 29, 1629–1636
53 Alexander, R.D. *et al.* (2010) Splicing-dependent RNA polymerase pausing in yeast. *Mol. Cell* 40, 582–593
54 Spies, N. *et al.* (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* 36, 245–254
55 Luco, R.F. *et al.* (2010) Regulation of alternative splicing by histone modifications. *Science* 327, 996–1000
56 Airoldi, C.A. *et al.* (2010) Single amino acid change alters the ability to specify male or female organ identity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18898–18902

57 Severing, E.I. *et al.* (2012) Predicting the impact of alternative splicing on plant MADS domain protein function. *PLoS ONE* 7, e30524
58 Brummell, D.A. *et al.* (2011) Induction of vacuolar invertase inhibitor mRNA in potato tubers contributes to cold-induced sweetening resistance and includes spliced hybrid mRNA variants. *J. Exp. Bot.* 62, 3519–3534
59 Kriechbaumer, V. *et al.* (2012) Alternative splicing of the auxin biosynthesis gene *YUCCA*4 determines its subcellular compartmentation. *Plant J.* 70, 292–302
60 Rebbapragada, I. and Lykke-Andersen, J. (2009) Execution of nonsense-mediated mRNA decay: what defines a substrate? *Curr. Opin. Cell Biol.* 21, 394–402
61 Nicholson, P. *et al.* (2010) Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell. Mol. Life Sci.* 67, 677–700
62 Arciga-Reyes, L. *et al.* (2006) UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in *Arabidopsis*. *Plant J.* 47, 480–489
63 Hori, K. and Watanabe, Y. (2005) UPF3 suppresses aberrant spliced mRNA in *Arabidopsis*. *Plant J.* 43, 530–540
64 Kerenyi, Z. *et al.* (2008) Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J.* 27, 1585–1595
65 Riehs, N. *et al.* (2008) *Arabidopsis* SMG7 protein is required for exit from meiosis. *J. Cell Sci.* 121, 2208–2216
66 Wu, J. *et al.* (2007) Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. *Plant J.* 51, 693–706
67 Yoine, M. *et al.* (2006) *Arabidopsis* UPF1 RNA helicase for nonsense-mediated mRNA decay is involved in seed size control and is essential for growth. *Plant Cell Physiol.* 47, 572–580
68 Yoine, M. *et al.* (2006) The *lba1* mutation of *UPF1* RNA helicase involved in nonsense-mediated mRNA decay causes pleiotropic phenotypic changes and altered sugar signalling in *Arabidopsis*. *Plant J.* 47, 49–62
69 Kertesz, S. *et al.* (2006) Both introns and long 3′-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res.* 34, 6147–6157
70 Nyiko, T. *et al.* (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol. Biol.* 71, 367–378
71 Schwartz, A.M. *et al.* (2006) Stability of plant mRNAs depends on the length of the 3′-untranslated region. *Biochemistry (Mosc)* 71, 1377–1384
72 Lareau, L.F. *et al.* (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446, 926–929
73 Ni, J.Z. *et al.* (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* 21, 708–718
74 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
75 Reddy, A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.* 58, 267–294
76 Doherty, C.J. and Kay, S.A. (2010) Circadian control of global gene expression patterns. *Annu. Rev. Genet.* 44, 419–444
77 Petrillo, E. *et al.* (2011) Alternative splicing adds a new loop to the circadian clock. *Commun. Integr. Biol.* 4, 284–286
78 Staiger, D. and Green, R. (2011) RNA-based regulation in the plant circadian clock. *Trends Plant Sci.* 16, 517–523
79 Hong, S. *et al.* (2010) Type II protein arginine methyltransferase 5 (PRMT5) is required for circadian period determination in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21211–21216
80 Sanchez, S.E. *et al.* (2010) A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature* 468, 112–116
81 James, A.B. *et al.* (2012) Alternative splicing mediates responses of the *Arabidopsis* circadian clock to temperature changes. *Plant Cell* 24, 961–981
82 Ingolia, N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802
83 Ner-Gaon, H. *et al.* (2004) Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J.* 39, 877–885
84 Seo, P.J. *et al.* (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. *Nat. Commun.* 2, 303
85 Seo, P.J. *et al.* (2011) Competitive inhibition of transcription factors by small interfering peptides. *Trends Plant Sci.* 16, 541–549
86 Staudt, A.C. and Wenkel, S. (2011) Regulation of protein function by 'microProteins'. *EMBO Rep.* 12, 35–42
87 Laitem, C. *et al.* (2009) Ets-1 p27: a novel Ets-1 isoform with dominant-negative effects on the transcriptional properties and the subcellular localization of Ets-1 p51. *Oncogene* 28, 2087–2099
88 Lee, G.M. *et al.* (2005) The structural and dynamic basis of Ets-1 DNA binding autoinhibition. *J. Biol. Chem.* 280, 7088–7099
89 Yang, Y. *et al.* (2000) Three-amino acid extension loop homeodomain proteins Meis2 and TGIF differentially regulate transcription. *J. Biol. Chem.* 275, 20734–20741
90 Hamant, O. and Pautot, V. (2010) Plant development: a TALE story. *C. R. Biol.* 333, 371–381
91 Magnani, E. and Hake, S. (2008) KNOX lost the OX: the *Arabidopsis* KNATM gene defines a novel class of KNOX transcriptional regulators missing the homeodomain. *Plant Cell* 20, 875–887
92 Lu, S.X. *et al.* (2009) CIRCADIAN CLOCK ASSOCIATED1 and LATE ELONGATED HYPOCOTYL function synergistically in the circadian clock of *Arabidopsis*. *Plant Physiol.* 150, 834–843
93 Yakir, E. *et al.* (2009) Posttranslational regulation of CIRCADIAN CLOCK ASSOCIATED1 in the circadian oscillator of *Arabidopsis*. *Plant Physiol.* 150, 844–857
94 O'Neill, J.S. *et al.* (2011) Circadian clock parameter measurement: characterization of clock transcription factors using surface plasmon resonance. *J. Biol. Rhythms* 26, 91–98
95 Gan, X. *et al.* (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423
96 Adams, K.L. and Wendel, J.F. (2005) Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids. *Genetics* 171, 2139–2142
97 Zhou, R. *et al.* (2011) Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16122–16127
98 Kalyna, M. *et al.* (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res.* 34, 4395–4405
99 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
100 Blencowe, B.J. (2006) Alternative splicing: new insights from global analyses. *Cell* 126, 37–47
101 Zhu, W. *et al.* (2003) Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiol.* 132, 469–484
102 Iida, K. *et al.* (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res.* 32, 5096–5103
103 Xiao, Y.L. *et al.* (2005) Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2 reveals numerous transcript variants. *Plant Physiol.* 139, 1323–1337
104 Campbell, M.A. *et al.* (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* 7, 327

Identification of RNA targets for the nuclear multidomain cyclophilin atCyp59 and their effect on PPIase activity

Bannikova, O., Zywicki, M., Marquez, Y., Skrahina, T., Kalyna, M., and Barta, A. (2013).

Nucleic Acids Res. 41: 1783–1796.

Published online 16 December 2012 Nucleic Acids Research, 2013, Vol. 41, No. 3 1783–1796
doi:10.1093/nar/gks1252

Identification of RNA targets for the nuclear multidomain cyclophilin atCyp59 and their effect on PPIase activity

Olga Bannikova[1], Marek Zywicki[2], Yamile Marquez[1], Tatsiana Skrahina[1], Maria Kalyna[1] and Andrea Barta[1,*]

[1]Max F. Perutz Laboratories, Medical University of Vienna, Vienna and [2]Innsbruck Biocenter, Medical University Innsbruck, Division of Genomics and RNomics, Innsbruck, Austria

Received August 3, 2012; Revised and Accepted November 1, 2012

ABSTRACT

AtCyp59 is a multidomain cyclophilin containing a peptidyl-prolyl *cis/trans* isomerase (PPIase) domain and an evolutionarily highly conserved RRM domain. Deregulation of this class of cyclophilins has been shown to affect transcription and to influence phosphorylation of the C-terminal repeat domain of the largest subunit of the RNA polymerase II. We used a genomic SELEX method for identifying RNA targets of AtCyp59. Analysis of the selected RNAs revealed an RNA-binding motif (G[U/C]N[G/A]CC[A/G]) and we show that it is evolutionarily conserved. Binding to this motif was verified by gel shift assays *in vitro* and by RNA immunopreciptation assays of AtCyp59 *in vivo*. Most importantly, we show that binding also occurs on unprocessed transcripts *in vivo* and that binding of specific RNAs inhibits the PPIase activity of AtCyp59 *in vitro*. Surprisingly, genome-wide analysis showed that the RNA motif is present in about 70% of the annotated transcripts preferentially in exons. Taken together, the available data suggest that these cyclophilins might have an important function in transcription regulation.

INTRODUCTION

Cyclophilins are ubiquitous proteins with a peptidyl-prolyl *cis*/*trans* isomerase (PPIase) activity and have important functions in protein folding (1). Typically they are small single-domain proteins but some of them have accessory domains. AtCyp59 is a member of the cyclophilin family which consists of 29 genes in *Arabidopsis* (2). AtCyp59 is unusually complex as it consists of a PPIase domain, an RRM motif, a Zn-knuckle and a charged C-terminal domain with RS/RD repeats (arginine/serine and arginine/aspartate) (3). It is an evolutionarily highly conserved protein present from *Schizosaccharomyces pombe* to humans, but the Zn-knuckle is a plant-specific addition. It was first described in *Paramecium tetraurelia* as a protein involved in cell morphogenesis (4).

The *Arabidopsis* protein Cyp59 was isolated in a yeast two-hybrid screen with plant SR (serine, arginine) proteins which are an important and conserved family of splicing factors (3). Deletion analysis showed that the C-terminal domain is indispensable for interacting with the SR proteins *in vitro*. AtCyp59 is a nuclear protein, but it does not significantly co-localize with SR proteins in nuclear speckles. Instead, its punctuate localization pattern resembles transcription initiation sites. In line with these observations, it was shown that AtCyp59 resides within a complex with the C-terminal repeat domain (CTD) of the largest subunit of the RNA polymerase II. These results suggested a possible function for AtCyp59 at the interface of transcription and splicing (3). It is now widely accepted that most splicing events occur co-transcriptionally whereby the CTD domain of RNA polymerase II plays very important roles in both transcription and RNA processing (5) and recently reviewed in (6). In general, the CTD acts as a binding platform for various protein factors during transcription, and at the same time recruits pre-mRNA processing proteins to the nascent transcripts (7,8). The CTD of most eukaryotes consists of a variable number of heptapeptide repeats (YSPTSPS) which undergo

*To whom correspondence should be addressed. Tel: +43 1 4277 61640; Fax: +43 1 4277 9616; Email: andrea.barta@meduniwien.ac.at
Present addresses:
Marek Zywicki, Laboratory of Computational Genomics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland.
Tatsiana Skrahina, Max Planck Institut for Infection Biology, Berlin, Germany.

dynamic phosphorylation/dephosphorylation events on serine residues and thereby determine the course of transcription and binding of RNA processing factors. These include capping, splicing and polyadenylation factors whose dynamics are tightly coordinated with transcription (9,10). The severe growth effect upon overexpression of AtCyp59 and the fact that no mutants are available point to an essential function for this protein (3). The involvement of this class of cyclophilins in the transcription process was corroborated by experiments with the *S. pombe* orthologue Rct1, which showed that Rct1 is an essential gene, is recruited to actively transcribed genes and its deregulation affected CTD phosphorylation and RNA polymerase II transcription (11). However, the mechanism of how these cyclophilins function in the transcription cycle is unknown. Possible scenarios are that they might act directly on the CTD structure and therefore influence phosphorylation/dephosphorylation of the heptapeptide or they might act on kinases/phosphatases which are regulating the CTD. It is worth mentioning that other smaller PPIases have been shown to be important for correct CTD conformation and phosphorylation thereby influencing both transcription and RNA processing (12–16).

One of the most interesting features of AtCyp59 is its complex domain structure. PPIases are usually small proteins, but a few other complex cyclophilins have been described. For example, in the splicing complex several RS-domain-containing cyclophilins have been investigated in *Arabidopsis* and mammals suggesting a role in RNA processing (12,17,18). Other RRM-containing cyclophilins have been found but their functions are mostly not determined. Among them, the best described protein is hCyp33, a regulator of a histone acetyl transferase; however, the function of RNA binding is not well defined (19,20). Interestingly, the RRM domain of the multidomain AtCyp59 is evolutionarily highly conserved and is in fact the most conserved feature of this protein. This RRM has been shown to bind RNA with preferences to G and C bases (3). These data indicated an important contribution of the RRM to the activity of AtCyp59 raising the question of what are its RNA target(s) and what influence RNA binding has on cyclophilin activity. Approaches to these questions are not trivial as the tight regulation of AtCyp59 strongly hindered *in vivo* approaches for determining RNA targets.

In this article, we describe the identification of an RNA-binding motif for AtCyp59 by a genomic SELEX method using an *Arabidopsis* genomic RNA library. Binding to this motif was verified by electrophoretic mobility shift analyses (21) *in vitro* and by RNA immunoprecipitation experiments *in vivo*. Interestingly, the identified motif is present in about 70% of the annotated genes in *Arabidopsis* indicating that binding of AtCyp59 to mRNA might be a general feature of most RNA polymerase II transcripts. This is supported by the conservation of this motif in *S. pombe*. In addition, we have shown activity of the PPIase domain of AtCyp59 and importantly its regulation by binding to specific RNAs. Considering the evolutionary conservation of the RNA-binding motif and the known characteristics of this protein, these data indicate a function for AtCyp59 in the transcription cycle.

MATERIALS AND METHODS

Genomic SELEX

Using random priming, a representative library of the *Arabidopsis thaliana* genome was constructed containing overlapping sequences from 50 to 300 nt in length. Library fragments were generated using the method and adaptors described in (22). Each library fragment contained fixed primers suitable for PCR amplification and preceding T7 promoter sequence at 5′-end for *in vitro* RNA transcription.

For selection of binding RNAs, we incubated recombinant AtCyp59_RRM_Zn domain protein with the *in vitro* transcribed genomic RNA pool for 30 min at room temperature using neutral PBS buffer conditions (2 mM $MgCl_2$, 0.5 mM DTT and 0.135 M NaCl, 27 mM KCl, 8 mM Na_2HPO_4, 2 mM NaH_2PO_4, pH = 7.5). Separation of bound and unbound fractions was performed via GST-tagged AtCyp59_RRM_Zn on 4B glutathione sepharose blocked with 100 µg tRNA (Sigma-Aldrich). Recovery of AtCyp59_RRM_Zn-binding RNAs was performed via urea-mediated denaturation followed by phenol/chloroform extraction. Selected sequences were amplified via reverse transcriptase-polymerase chain reaction (RT–PCR) as described (23). For the next SELEX cycle, obtained PCR products were again *in vitro* transcribed into RNA. In total, 10 cycles of *in vitro* selection were performed using molar ratio RNA:protein = 3:1 (or 10:1 on later cycles). On the ninth cycle the control (anti-GST selection) was performed using 10:1 molar ratio of recombinant GST over RNA library. Binding reaction was done as described above with the only change of keeping the RNA fraction unbound to the beads instead of beads fraction. Next, 1/10 of the selected library was cloned via T/A cloning to the pGEM-T easy vector and 386 clones were sequenced using Sanger sequencing method. The rest of the selected library was sequenced using 454 sequencing technology. Plant material preparation, plasmids and protein purification are described in Supplementary Methods.

Computational analysis of the SELEX library

The raw sequencing reads have been submitted to analysis with APART pipeline (24). For the adaptor filtering step, the following sequences have been used: AGGGGAATTC GGAGCGGGGCAGC (5′-adaptor), CATCCCAGCCCC GAGGAT (3′-adaptor). For downstream steps, only sequences with both adaptors were analyzed. All parameters of APART have been set to default, except the minimum number of reads per contig which has been changed to 2. For identification of the AtCyp59-binding motif consensus sequences, the seven contigs with the highest read number (accounting for 79.5% of all mapped reads) have been used. The motif identification has been performed using the MEME program (25) with

the following non-default settings: motif width 7, minimum number of sites 4 and maximum number of sites 15.

Genome-wide distribution of the AtCyp59-binding motif

For the analysis of the AtCyp59-binding motif at a genome-wide scale, a motif descriptor based on experimentally verified motif variants has been constructed. The search has been performed using the glam2scan (26) software with the default parameter set and score cut-off 7. The sequences mapped to *A. thaliana* correspond to TAIR10 genome assembly (27). The occurrence of the 17 motifs experimentally validated in *Arabidopsis* was analysed in the transcription units (TUs). If more than one TU was annotated for a gene, the longest one was chosen for the motif search. Only the TUs of protein-coding genes were used in the analysis (TAIR10). As a control, the distribution of two scramble motifs ('TAGC GTC' and 'CATGTGC') was analyzed in the protein-coding genes of *Arabidopsis*. The same analysis was done for the TUs in *S. pombe*.

Additionally, a motif search was performed in TUs of different sizes in *Arabidopsis*. Small TUs were defined as TUs with a size between 500 and1000 nt. Medium TUs were defined with a size range between 1001 and 2000 nt and big TUs with a size range between 2001 and 4000 nt. Binomial tests were used for the *P*-value calculations.

Electrophoretic mobility shift assays

Synthetic 7 nt RNA oligonucleotide (Sigma-Aldrich) (see Table 1) (100 nM) or longer RNAs (see Supplementary Methods and Supplementary Table S3) was incubated on ice for 20 min in 10 mM HEPES-KOH, pH 7.9, 10 mM $MgCl_2$, 50 mM KCl, 1 mM DTT, 0.025% Nonindet P-40, supplemented with protease inhibitor cocktail (Roche) and RNAse inhibitor (Promega) with various amounts of recombinant AtCyp59, AtCyp59_RMM_Zn, AtCyp59_3M_RRM_Zn, Rct1 protein as indicated. RNA–protein complexes were separated on 6–10% native polyacrylamide gels at 3 V/cm. Gels were stained with SYBR GREEN II dye and visualized on a Phosphoimager (Thyphoon 900) and quantified using ImageQuant 1.4 software. Assay was performed three times with independent protein purifications.

Table 1. Range of dissociation constants of 17 different 7-nt RNA variants of the binding motif found in genomic SELEX and AtCyp59_RRM_Zn domain based on gel shift assay

Sequence	K_d AtCyp59_RRM_Zn (nM)	K_d AtCyp59 (nM)
GGUGCCG	40 ± 10	120 ± 25
GUGGCCG	40 ± 10	120 ± 25
GUCGCCG	40 ± 10	120 ± 25
GUAGCCA	105 ± 25	120 ± 25
GCUGCCG	105 ± 25	120 ± 25
GAUGCCA	105 ± 25	120 ± 25
GACGCCA	105 ± 25	120 ± 25
GCCGCCG	200 ± 35	120 ± 25
GUUGCCG	200 ± 35	120 ± 25
GUAGCCG	200 ± 35	120 ± 25
GCGGCCG	200 ± 35	120 ± 25
GUCGCCA	300 ± 40	120 ± 25
GCCGCCA	300 ± 40	120 ± 25
GAUGCCG	300 ± 40	120 ± 25
GUGGCCA	300 ± 40	120 ± 25
GGAGCCA	300 ± 40	120 ± 25
GCGGCCA	300 ± 40	120 ± 25

Preparation of whole-cell extracts from protoplasts and immunoprecipitation

Arabidopsis cell suspension protoplasts were isolated and transformed with pDEDH-AtCyp59-HA, pDEDH-Cyp59_3M-HA, pDEDH-GFP or pGREEN-MAPK6-HA as described (17). For double transformation experiments, *Arabidopsis* protoplasts were transformed with equal concentration of the following plasmid pairs: pDEDH-AtCyp59-HA/pDEDH-atSR34a; pDEDH-AtCyp59-HA/pDEDH-Mut_atSR34a; pDEDH-AtCyp59-HA/pDEDH-atRS2Z32 and pDEDHAtCyp59-HA/pDEDH-Mut_atRS2Z32. Transformed protoplasts were collected 24 h after transformation (15 min, 700*g*), frozen in liquid nitrogen, and resuspended in 300 µl (per 10 million protoplasts) protoplasts extraction buffer [50 mM HEPES-KOH pH 7.9, 2.5 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT, 1% sodium dodecyl sulphate (SDS)], supplemented with EDTA-free protease inhibitor cocktail (Roche Diagnostics) and RNase inhibitor (Roche Diagnostics). Suspension was sonicated three times for 6 s and incubated on ice for 20 min with occasional mixing. After centrifugation (15 min, 14 000 rpm, 4°C), concentration of SDS in extracts was adjusted to 0.1% with SDS-free lysis buffer. Extracts were incubated for 1 h with magnetic Dynal beads (m-270 epoxy; Invitrogen), coupled with anti-HA antibodies produced in mouse (HA-7, monoclonal; Sigma-Aldrich) on rotary shaker at 4°C and then were washed three times with protoplast extraction buffer without SDS and three times with washing buffer (10 mM HEPES-KOH, pH 7.9). RNA was extracted via protein digest with 100 µg proteinase K (Sigma-Aldrich) and 5 µl 10% SDS in 400 µl washing buffer for 30 min at 55°C followed by phenol/chloroform extraction and ethanol precipitation. RNA was purified by DNAse I (Qiagen) treatment followed by Qiagen RNeasy Plant Mini kit. Purified RNA was used directly in RT–PCR (33 PCR cycles) or stored at −80°C. RT reaction was performed with equal volume of RNA sample, 15-mer oligo-dT or pre-mRNA primers targeting a nascent pre-mRNA after the polyA signal and M-MLV reverse transcriptase (Promega) according to the manufacturing instruction and then was used in standard PCR with Phusion polymerase (Finnzyme). Primers for target genes used in PCR are listed in the Supplementary Table S3.

PPIase activity assay

The PPIase activity of recombinant GST-AtCyp59 or GST-AtCyp59_3M was performed as described (28) by using the tetrapeptide substrate Suc-AAPF-pNA (*N*-succinyl-Ala-Leu-Pro-Phe p-nitroanilide; Sigma-Aldrich). All reagents were pre-equilibrated until the temperature reached 4°C. In a 1-ml glass cuvette,

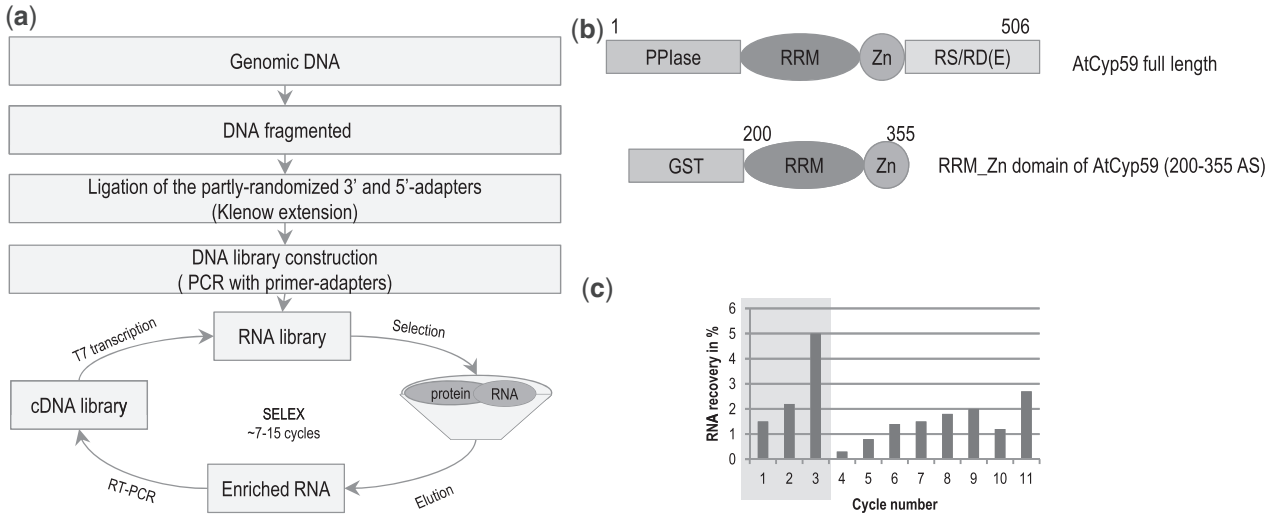


Figure 1. Genomic SELEX experiment with the multidomain cyclophilin AtCyp59_RRM_Zn domain. (**a**) Scheme of *Arabidopsis* RNA library preparation and selection process. (**b**) Schematic representation of the domain organization of the full-length AtCyp59 and GST-tagged construct of the RRM+Zn motif of AtCyp59. PPIase, peptidyl/prolyl *cis*/*trans* isomerase domain; RRM, RNA recognition motif; Zn, zinc finger motif class CCHC; RS/RD(E), domain enriched in arginine, serine, aspartate, glutamate; GST, glutathione *S*-transferase. (**c**) Enrichment of the RNA sequences from the *Arabidopsis thaliana* genome which bind AtCyp59. The RNA recovery was calculated as a percentage of the RNA material that was bound to the protein relatively to the amount of the initial RNA material in each cycle. During Cycles 1–3, the molecular ratio of the RNA to protein was 3:1, Cycles 4–9, 11 the ratio was 10:1, and Cycle 10 anti-GST selection.

800 nM GST-AtCyp59 or GST-AtCyp59_3M was mixed with 100 µl of α-chymotrypsin (Sigma-Aldrich; 60 mg/ml in 1 mM HCl), and the volume was adjusted to 975 µl with assay buffer (50 mM Hepes-KOH, pH 8.0 at 0°C, 100 mM NaCl, 2 mM $MgCl_2$, 1 mM EDTA). The reaction was initiated by the addition of 25 µl of substrate (4 mM tetrapeptide in 470 mM anhydrous LiCl dissolved in trifluoroethanol). Changes in absorbance due to released *p*-nitroaniline were monitored at 390 nm at 0°C over a 3-min period in a PerkinElmer Lambda 35 UV/VIS spectrophotometer with a thermostatically controlled cuvette holder. To check PPIase activity of proteins in presence of RNA, 800 nM GST-AtCyp59 or GST-AtCyp59_3M was pre-incubated in reaction buffer with equal concentration of 7 nt RNA (GUGGCCG), polyA+ fraction of total RNA (800 nM) isolated from 21-day-old wt Col-0 plants using Micropoly(A) purist kit (Ambion). Pre-incubated protein was similarly used in the assay above. The experiments were performed five times with different preparations of proteins.

RESULTS

RNA targets of AtCyp59 identified by Genomic SELEX

The systematic evolution of ligands by exponential enrichment (SELEX) is an approach to isolate high-affinity binding partners for a given molecule and usually uses a random nucleic acid library (29,30). In contrast, Genomic SELEX has the advantage of selecting only from the sequences available in a given genome which enhances the possibility of isolating natural RNA targets (31,22). *In vivo* studies on AtCyp59 are hindered by its tight gene regulation with the consequence that *Arabidopsis* overexpression and mutant lines are currently not available. We therefore used Genomic SELEX as an *in vitro* method to identify RNAs bound by AtCyp59 and to discover RNA motif(s) for this protein.

To create a genomic library 2 g of 3-week-old wild-type *A. thaliana* (Col-0) leaves were used to isolate genomic DNA. Thirty micrograms of this DNA was fragmented and used as template to create a DNA library by Klenow fragment reaction. The addition of primers with constant regions and T7 promoter sequences allowed the synthesis of an RNA library (Figure 1a). The sequences of primer adaptors were not present in the Arabidopsis genome. Library construction was performed in such a way as to create an RNA library with sizes of 30–300 nucleotides. Library construction and the selection of RNA targets were essentially done as described (22) and a more specific description for *Arabidopsis* is published elsewhere (32). For the selection procedure, the region encoding the RRM and Zn-knuckle domain of AtCyp59 was cloned and expressed as a recombinant GST-fusion protein. Binding to the RNA library was performed in a 3:1 molar excess of RNA over protein for the initial first three cycles of selection and then the ratio was changed to 10:1 to increase the stringency. Bound RNAs were separated on protein-GST beads and eluted with glutathione. To control for RNAs binding to the GST tag, a counter selection with GST protein was performed after the ninth cycle (Figure 1b). The selection was stopped after 11 rounds of SELEX when the recovery rate of RNA was about 3% of the input RNA. As a 10-fold molar excess of RNA over protein was used for competition, this value suggests that about 30% of the selected RNA pool is able to bind to AtCyp59.

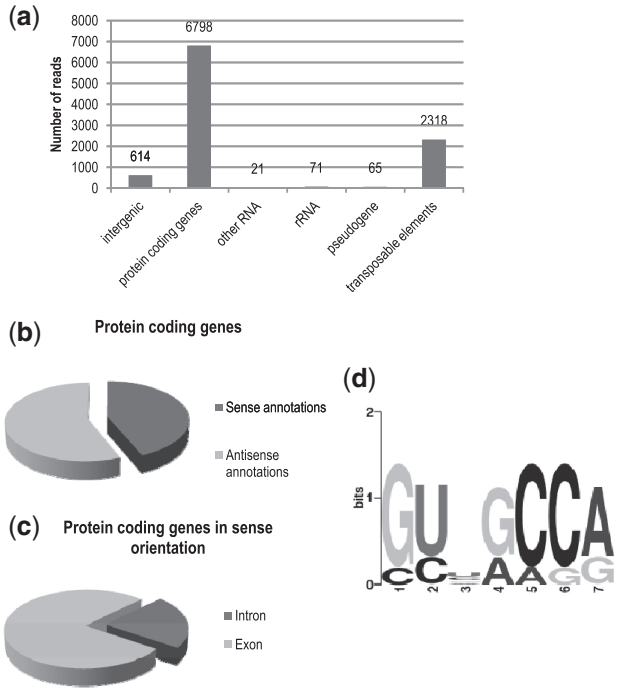


Figure 2. Bioinformatics analysis of the selected library. (**a**) Genomic distribution of the mapped reads generated after selection with AtCyp59_RRM_Zn domain. (**b**) Strand distribution of the aligned contigs according to the annotated gene. (**c**) Localization of the reads within the annotated gene in sense orientation. (**d**) Predicted binding motif based on sequences alignment of the seven most abundant clusters in the SELEX library using the MEME program tool.

To control the library construction and SELEX procedure, selected RNAs were reverse transcribed, cloned and sequenced by conventional sequencing. Most of the reads contained *Arabidopsis* sequences of about 40–50 nucleotides (Supplementary Figure S1b). Therefore, the selected library was sequenced by the 454 deep sequencing methods. We obtained a total of about 13 375 trimmed reads and about 70% mapped to the *Arabidopsis* genome (Supplementary Table S1). From these sequences almost 60% mapped uniquely. The sequences that are more than once mapped are probably affiliated to duplicated genomic regions. We also sequenced the genomic library with 454 technology and compared the abundance of genomic elements to the *Arabidopsis* genome and the SELEX library (Supplementary Figure S1c). All genomic elements were present in the genomic library albeit with a slightly different abundance, whereas the SELEX library showed a clear increase in exonic sequences and a decrease in intronic sequences.

Statistical examination of the selected sequences reads revealed that most of the targets of AtCyp59 reside in protein-coding genes (Figure 2a). Interestingly, the majority of reads were observed to map in antisense orientation to annotated genes, mostly due to the genomic location of two contigs with the highest read number (accounting for 50% of all reads). When reads were assembled into contigs, the ratio between contigs in sense and antisense annotations was 3:4 (Figure 2b). By investigating the gene structure of genes which contain selected sequences in the sense orientation, we discovered that binding occurs preferentially to exons of the annotated transcripts and only few hits corresponded to introns (Figure 2c). As the Genomic SELEX method samples the entire sequence of a given genome in an unbiased manner, the obtained data clearly show that the potential targets of AtCyp59 are mainly located in protein-coding genes with a high preference for exonic regions.

Bioinformatics analysis reveals an RNA-binding motif for AtCyp59

One of the purposes of any SELEX experiment is to find a common binding motif for a protein of interest within the pool of selected sequences. The seven most abundant clusters (accounting for 79.5% of all aligned reads) in our selection were the basis for the alignment with the MEME program, which searches for consensus binding motifs (33). The obtained consensus sequence G[U/C]N[G/A]CC[A/G] (Figure 2d) is clearly GC-rich which is in line with previously published data indicating that AtCyp59 prefers either G- or C-rich sequences (3). This G[U/C]N[G/A]CC[A/G] motif was highly enriched in the selected sequences (4.5 times/per 1000 nt) and about 50% of the sequenced reads contained the motif (Supplementary Table S2). As the motif discovery was based on 79.5% of the reads, there might be a possibility of another potential binding motif. It is well established that the composition of exons in general is more GC-rich and in particular the *Arabidopsis* exons of pre-mRNAs are biased towards a higher GC content (34). Therefore, our findings fitted well to the observation that sequencing reads were preferentially aligned to exons of protein-coding genes. Taken together, these results support the notion that the RNA targets of AtCyp59 are located within the exons of protein-coding genes.

In vitro verification of AtCyp59 binding to the selected motif

To verify the binding of AtCyp59 to its RNA targets, electrophoretic mobility shift assay with the GST-tagged version of the RRM and Zn-knuckle domain used in the selection process were performed. To control for unspecific binding, the RNP1 motif of the RRM was mutated in three conserved aromatic amino acids which are indispensable for RNA recognition (Y286D, F289D, F291D) (35). In addition, the full-length protein was expressed, purified and used along with the truncated version (Figure 3a).

An initial test used one of the sequences obtained in the genomic SELEX screen, a 180-nt RNA from the AT3G53500 gene (coding for the *Arabidopsis* SR protein At-RS2Z32) containing one of the binding motifs (GUCG CCG). Full-length AtCyp59 protein was added in increasing amounts and the binding reaction was separated on a native PAA gel. As shown in Figure 3b, prominent complex formation between AtCyp59 and the RNA with a K_D of 50 nM was observed suggesting that the sequence contained a functional binding motif.

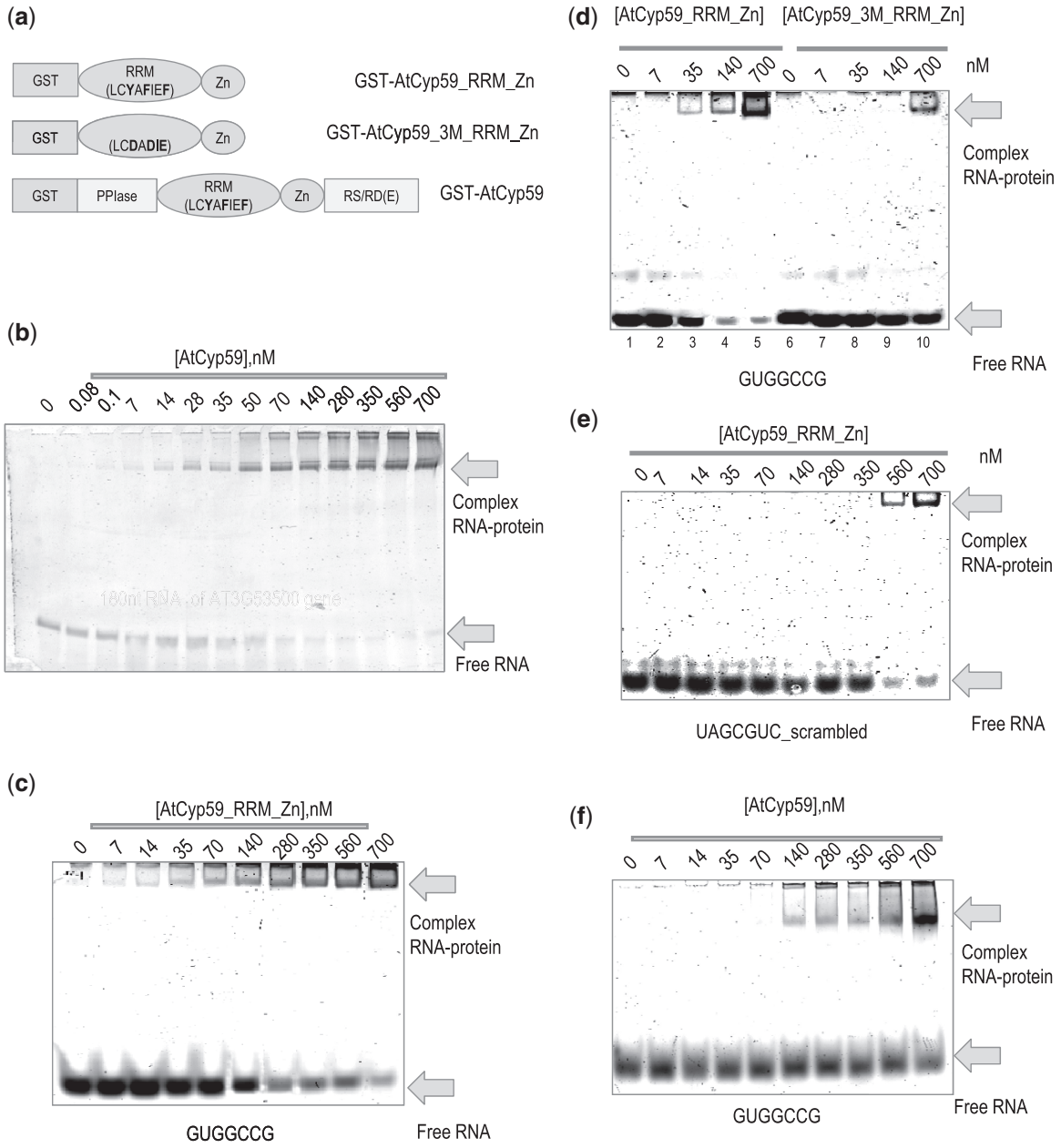


Figure 3. Binding of the AtCyp59 to its RNA targets *in vitro*. (**a**) Schematic representation of AtCyp59 protein variants used in the *in vitro* studies. The amino acids involved in RNA binding to the RRM (RNP1 motif) domain are marked in bold. These amino acids have been mutated in AtCyp59_3M_RRM_Zn. (**b**) 6% native polyacrylamide gel electrophoresis of a 180-nt RNA (gene At3G53500) containing one of the binding motifs (GCUGCCG). (**c**) 10% PAA gel electrophoresis of a 7-nt RNA sequence (GUGGCCG) incubated with an increasing concentration of the AtCyp59_RRM_Zn. (**d**) Comparison of binding to the 7-nt RNA (GUGGCCG) between wild-type AtCyp59_RRM_Zn protein (lanes 1–5) and the mutated RRM domain (lanes 6–10). (**e**) 10% native polyacrylamide gel electrophoresis of a 7-nt scrambled RNA (based on consensus sequence) incubated with an increasing concentration of the of AtCyp59_ RRM_Zn. (**f**) RNA–protein complex formation upon full-length AtCyp59 binding to the 7-nt RNA (GUGGCCG) on 6% PAA native gel. (b–f) The free RNA and RNA–protein complexes are indicated by arrows. Gels were stained with SYBR GREEN II dye. The concentration of the protein is shown in the nM range.

The binding motif for AtCyp59 G[U/C]N[G/A]CC[A/G] was used to search the sequencing reads and 17 sequence variations of the seven nucleotide motif were generated (Table 1, Supplementary Figure S1). All these 17 short RNA oligonucleotides were used for binding tests in an electrophoretic mobility shift assay on 10% native gels. Table 1 shows all tested RNA oligonucleotides listed according to their binding affinities with a K_D in the range of 40–300 nM. No significant preference for a particular nucleotide at the three variable positions could be observed, except that a G at the last position is preferable. By using one of these motifs, GUGGCCG, and the RRM_Zn-knuckle domain of AtCyp59, we observed strong binding with a K_D of 40 nM (Figure 3c). In contrast, when the mutated protein was used, binding dropped significantly to >700 nM (Figure 3d, lanes 6–10). This value is similar in range to that obtained with a scrambled RNA oligonucleotide (UAGCGUC)

bound to the non-mutated AtCyp59_RRM_Zn protein (Figure 3e). Thus, the RNP1 motif of the RRM domain is at least partially responsible for binding RNA. As the RRM is the most conserved domain in this cyclophilin family, we tested the *S. pombe* orthologue Rct1 in this binding assay using the Arabidopsis motif variant GUG GCCG. Interestingly, we observe similar binding of this oligo RNA to the Rct1 protein as to the *Arabidopsis* protein (Supplementary Figure S2c and d). These experiments strongly support an evolutionarily conserved function for the RRM of this family of cyclophilins.

To investigate the possible influence of the other domains of AtCyp59 on binding, the same experiments were performed using the full-length AtCyp59 protein. Interestingly, all the tested motif variants now showed a very similar dissociation constant of about 120 nM (Table 1 and Figure 3f). This levelling of the binding affinities to the 7-nt oligos by the other domains of AtCyp59 suggests that the motif variants might bind equally well to the full-length protein. However, as can be observed in the case of the GUCGCCG motif in the context of the AT3G53500 (At-RS2Z32) gene (Figure 3b and Supplementary Figure S2a and b) yielding a K_D of 50 nM, the RNA sequence context within the transcript might have an additional influence on binding to a particular motif. In summary, the *in vitro* experiments have verified the binding motif for AtCyp59 from sequences selected by the Genomic SELEX method and have shown that this motif is evolutionarily highly conserved.

RNA transcripts containing the selected binding motif are bound by AtCyp59 *in vivo*

The *in vivo* testing of RNAs with a binding motif for AtCyp59 proved difficult. All our efforts to establish stably transformed plants or tissue culture cell lines failed as even small changes in the level of AtCyp59 are detrimental to cell growth (3, and our unpublished data). We therefore used a transient expression system in plant protoplasts where we used HA-tagged WT and mutated Cyp59_3M (3 M in RNP1) proteins (Figure 4a) for RNA immunoprecipitation to show specific binding of AtCyp59 to endogenous mRNAs containing the selected motif. A plasmid expressing HA-tagged MAPK6 kinase was used as control and transformation efficiency was monitored by a 35S-GFP construct. Protoplasts isolated from an *Arabidopsis* cell suspension culture were transformed with DNA constructs and after 24-h extracts were prepared and used for immunoprecipitation with anti-HA antibody. Figure 4b shows a western blot of the total protein isolated from transformed plant protoplasts with anti-HA antibodies demonstrating that both the AtCyp59 and the mutated AtCyp59_3M were expressed at similar levels (top panel) and were efficiently immunoprecipitated with anti-HA antibody (bottom panel). To test if the transformation of the HA-tagged protein itself affects the levels of the tested endogenous RNAs, control RT–PCR experiments were performed. None of the transfected proteins influenced the mRNA expression level of any of the tested endogenous RNAs (Supplementary Figure S3a). RNAs co-precipitated with AtCyp59, Cyp59_3M and MAPK6 were isolated and analysed by RT–PCR using oligonucleotides corresponding to endogenous mRNAs containing the selected motif. The majority of the motif-containing RNAs (10 of 13) could be recovered by IP of the HA-tagged AtCyp59 but not by the mutated AtCyp59_3M or the MAPK6 kinase control (Figure 4c and Supplementary Figure S3b).

As AtCyp59 is involved in transcription regulation, we argued that it should bind to the nascent transcript. Therefore, we repeated the RT–PCR experiment for three target transcripts using an RT primer downstream of the polyA addition site. This should only capture RNAs before 3′-end cleavage and polyA addition. In the PCR analysis in Figure 4d, we observe recovery of pre-mRNAs, partially spliced RNAs as well as of spliced RNAs showing that binding of AtCyp59 must have occurred on an unprocessed transcript. To control for DNA contamination, PCRs were carried out without prior reverse transcription. Additionally, these experiments indicate that splicing can occur before 3′-end processing as suggested previously.

To investigate if binding indeed occurred to the established RNA-binding motif, we designed an experiment where we used wild-type AtCyp59 and transformed it together with a construct containing one of the IP mRNAs which was mutated at the ATG (TTG) to avoid protein overexpression and with a tag at the 3′-end to distinguish it from the endogenously expressed mRNA. In addition, we mutated the AtCyp59-binding motif to a U/A-stretch in this mRNA and used this construct in co-transformation experiments with AtCyp59. This was done for the mRNAs of AT3G49430 (At-SR34a) and AT3G53500 (At-RS2Z32) (Figure 4e). Figure 4f (upper panels) shows a western blot with anti-HA antibodies from the immunoprecipitation of the co-transformation experiments to control for AtCyp59 expression. Figure 4g (upper panels) shows RT–PCR of total RNA of the co-transformation extract indicating that all transfected target RNAs were expressed in equivalent amounts. RT–PCR analyses of the IP (Figure 4g, lower panels) show that the RNAs containing the selected motif are present in the IP in contrast to the mRNAs with the mutated binding motif.

Taken together, these experiments show that AtCyp59 binds to the selected binding motif *in vivo* and most importantly that binding likely occurs co-transcriptionally.

AtCyp59 is an active PPIase and its activity is inhibited by RNA binding

Having demonstrated that AtCyp59 indeed binds specific RNA sequences *in vivo*, it was of great interest to investigate if RNA binding could have an influence on the enzymatic activity (PPIase) of AtCyp59. However, it has never been shown that either the *A. thaliana* or the *S. pombe* protein possesses PPIase activity. We therefore used an *in vitro* PPIase assay and tested recombinant AtCyp59. This assay uses a substrate that has the

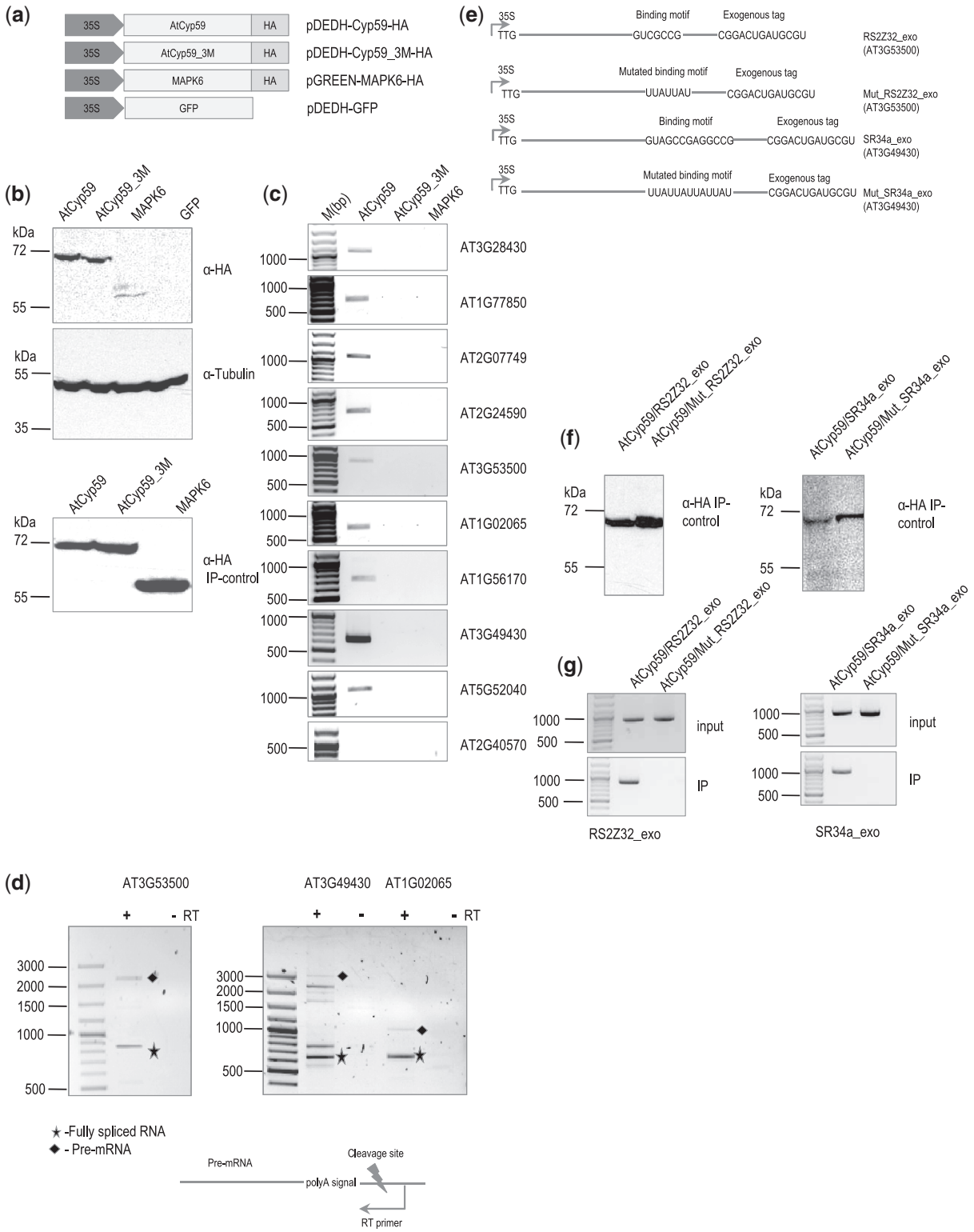


Figure 4. Transiently expressed AtCyp59 HA-tagged proteins in the *A. thaliana* protoplasts bind to endogenous mRNAs *in vivo*. (**a**) Schematic representation of the AtCyp59 constructs used in the transient transformation experiments. 35S, CaMV 35S RNA promoter sequence; HA, hemagglutinine antigen; 3M; mutations in the RRM domain (the same as in Figure 3a); GFP, green fluorescence protein; MAPK6, mitogen-activated protein kinase 6. (**b**) Western blot analyses of cells overexpressing the depicted constructs were performed using anti-HA antibody (upper panel) or anti-tubulin antibody (middle panel). Bottom panel is a western blot analysis of an HA immunoprecipitation used for RNA analysis shown in (**c**). Molecular weight markers are indicated on the left side. (c) Analysis of RNAs immunoprecipitated with the indicated proteins. RT–PCR with primers to the genes indicated on the right side (33 PCR cycles; primers used are listed in the Supplementary Table S3). Molecular weight marker is shown on the left side in bp range. (**d**) Analysis of selected pre-mRNAs immunoprecipitated with AtCyp59. Primers for RT reaction are designed to tag a pre-mRNA after the polyA signal; primers in the PCR are the same as in (c). +RT, PCR with reverse transcriptase added; -RT, PCR without reverse transcriptase; filled diamond, pre-mRNA product; star, fully spliced mRNA product. Other bands in the AT3G4930 lane are partially spliced

(continued)

peptide bond in *cis* conformation adjacent to a proline (Suc-Ala-Ala-Pro-Phe-p-nitroanilide) (36,28). This proline is followed by a *p*-nitro-Phe group which can be cleaved off by chymotrypsin if the peptide bond adjacent to the proline is in *trans* conformation. Conversion of substrate from *cis* to *trans* was measured by reading the absorbance of the released *p*-nitroaniline at 390 nm. Observed reaction rates were calculated as an average from four independent experiments. Figure 5a shows that addition of recombinant AtCyp59 to the substrate accelerated the spontaneous *cis*/*trans* isomerization considerably (blank: $K_{obs} = 2.37 \pm 0.39$;+Cyp59: $K_{obs} = 5.11 \pm 0.89$; Figure 5d). Addition of AtCyp59_3M in this assay resulted in only a small reduction in activity ($K_{obs} = 4.50 \pm 0.79$), indicating that the mutations in the RRM do not affect the PPIase activity of AtCyp59 significantly (Figure 5b and d). To test if binding of RNA influences PPIase activity, we first pre-incubated a polyA+ mRNA from total *Arabidopsis* RNA with wild-type AtCyp59 before adding it to the PPIase reaction. As shown in Figures 5a and d, we see a significant effect of reduced PPIase activity ($K_{obs} = 3.25 \pm 0.93$) on the polyA+ fraction. The effect on polyA+ RNA is in line with our observation that the binding motif of AtCyp59 mainly occurs in protein-coding genes. Furthermore, incubation with a 7-nt motif sequence (GUGGCCG) similarly decreased the activity of AtCyp59 PPIase ($K_{obs} = 3.17 \pm 0.29$) (Figure 5a, c and d), whereas incubation with a scrambled oligo (UAGCGUC) did not affect the activity ($K_{obs} = 4.89 \pm 1.02$). We also used the mutated AtCyp59_3M protein pre-incubated with polyA+ RNA. We observed a slight reduction of the PPIase activity compared with the AtCyp59_3M without RNA ($K_{obs} = 4.50 \pm 0.79$ versus $K_{obs} = 4.08 \pm 0.96$; Figure 5b and d). This observation corresponds to results from *in vitro* RNA-binding assays showing that the mutated protein still has a residual binding capacity for the RNA (Figure 3d and e). Taken together, these data show that AtCyp59 possesses an active PPIase domain and that specific binding of RNA to the RRM of AtCyp59 reduces the PPIase activity *in vitro*.

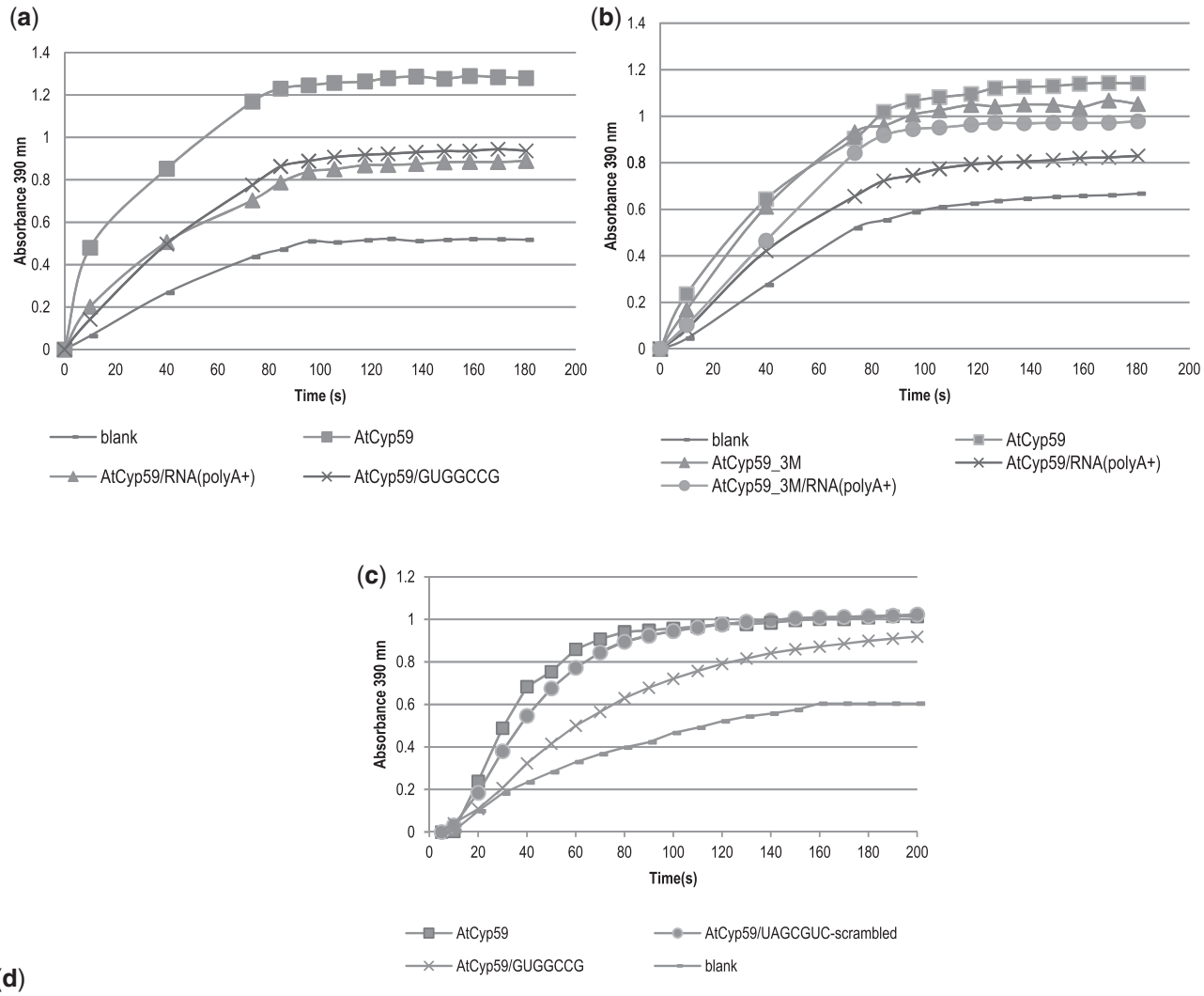
The AtCyp59-binding motif is present in most RNA polymerase II transcripts

Having verified the binding motif by *in vitro* and *in vivo* methods, it was interesting to get a more genomic view of the distribution of the G[U/C]NGCC[A/G] motif in the *Arabidopsis* genome. The 17 motif variants which were tested *in vitro* (Table 1) were used to construct an experimentally validated version of the binding motif descriptor. The analysis of the *Arabidopsis* genome (TAIR10) reveals that about 70% of the *Arabidopsis* protein-coding genes contained one of the AtCyp59-binding motifs (Figure 6a). Considering that we did not test all possible variants for this motif, the number of RNA polymerase II transcripts possessing this motif might be much higher. It is very interesting that the motif density (number of the motifs per 1000 nt) is equally distributed to sense and antisense transcripts of the annotated genes (Figure 6a, 0.5 in sense and 0.51 in antisense orientation), while for intergenic regions the number is much lower (0.27). Even more drastic changes have been observed comparing exons and introns of protein-coding genes. The motif densities varied significantly, from 0.83 in exons to 0.21 in introns. These data show that the predominant location of the binding motif is within exons of protein-coding genes. Furthermore, we aligned the motif to the cDNA sequences (corresponding to mature mRNAs) and found that the majority of the transcripts contain the binding motif only once or twice in a given transcript body (Figure 6b). The analysis of the 17 motifs in the *S. pombe* genome revealed a similar distribution as in *Arabidopsis* (Figure 6a). About 64% of protein coding transcripts possess a motif mostly in the coding region.

To examine if the motif had any general tendency for a location within the TU, we normalized motif occurrence to transcript length. Taking all 17 verified motif variants into account, we found a slightly higher proportion of motifs in the beginning and a decrease towards the end of the *Arabidopsis* transcript (Figure 6c). A similar analysis of the 17 motifs on the *S. pombe* genome showed a more prominent accumulation of the motifs towards the gene body (Figure 6d, grey bars). To determine a possible influence of gene length on the distribution of the motif, we performed the same analysis on small (750 ± 50 bp), medium (1500 ± 100 bp) and long (3000 ± 200 bp) genes (Supplementary Figure S4). Interestingly, the small genes profiles are similar to the *S. pombe* genomic distribution which is coherent with the smaller gene sizes in yeast. The bell–shaped curve is probably due to the 5′- and 3′-UTR having less motif variants. The medium- and long-size genes behave similar as in the genomic analysis (Figure 6c and d and Supplementary Figure S4) indicating that the observed motif peak at the beginning of genes comes from the lager genes as they contain more coding sequences in the first 10% of the nucleotide sequence.

Taken together, the fact that the AtCyp59-binding motif occurs in almost 70% of the protein-coding genes implicates that it might be important for the majority of RNA polymerase II TUs. This is supported by the

Figure 4. Continued
products. (**e**) Schematic representations of the RNA constructs used for double transformation experiment with AtCyp59. TTG, mutated translational start codon; line, coding sequence of the gene; exogenous tag, partial sequence of the hemagglutinine antigen. (**f**) Western blot analyses of immunoprecipitation experiments with cells overexpressing the depicted combinations of constructs. The molecular weight marker is displayed on the left side. (**g**) RT–PCR with primers to the genes indicated at the bottom (primers used are listed in the Supplementary Table S3). Molecular weight marker is shown on the left side in bp range. Upper panel is a RT–PCR from total RNA isolated from protoplasts double transformed with the depicted constructs. Lower panel is an RT–PCR of RNA isolated after IP. Bands shown are amplified using the primers to the synthetic RNA construct which were designed to differentiate from endogenously expressed RNA.



(d)

K_{obs} ($x10^{-3}$,s^{-1})						
Blank	AtCyp59	AtCyp59_3M	AtCyp59/RNA (polyA+)	AtCyp59_3M/ RNA(polyA+)	AtCyp59/GUGGCCG	AtCyp59/UAGCGUC
2.37±0.39	5.11±0.89	4.5±0.79	3.25±0.93	4.08±0.96	3.17±0.29	4.89±1.02

Figure 5. PPIase activity of AtCyp59 decreases upon binding to the mRNA (polyA+) or to a 7-nt RNA binding motif. (**a**) PPIase activity in the absence (−) or in the presence (filled square) of AtCyp59. PPIase activity is inhibited upon binding either to the polyA+ fraction of RNA (filled triangle) or to an RNA binding motif (×). (**b**) PPIase activity in the absence (−) or in the presence of AtCyp59 (filled square), or in the presence of AtCyp59_3M (filled triangle). PPIase activity of AtCyp59 upon binding to the polyA+ fraction of RNA (×). PPIase activity of AtCyp59_3M upon binding to the polyA+ fraction of RNA (filled circle). (**c**) PPIase activity in the absence (−) or in the presence (filled square) of AtCyp59. PPIase activity is inhibited upon binding to the specific RNA binding motif (×) but is not inhibited upon binding to the scramble RNA oligonucleotide (filled circle). (**d**) K_{obs} of four independent measurements of the PPIase activity. Kinetics is calculated as first-order Michaelis–Menten reaction.

bioinformatics analysis showing that it is equally abundant in antisense transcripts. In addition, the evolutionarily conserved abundance and location of this motif points to an important function of this protein in the transcription process.

DISCUSSION

In this article, we determined the RNA targets of an RRM-containing cyclophilin AtCyp59 by using a Genomic SELEX method. The selected sequences enabled us to identify the RNA-binding consensus motif for the evolutionarily highly conserved RRM domain of this multidomain protein. The binding of AtCyp59 to its RNA targets was confirmed *in vitro* by mobility shift assays and *in vivo* by RNA immunoprecipitation from protoplasts transiently expressing the HA-tagged protein. In addition, we showed that mutations in either the RRM domain of AtCyp59 or in the RNA motif sequence decrease the binding specificity of the AtCyp59 to its RNA targets. Furthermore, we show that recombinant AtCyp59 exhibits PPIase activity and that this activity is decreased upon binding to the specific RNA-binding motif. Genome-wide analysis of the RNA

(a)

	Motif density (motifs/1000nt)	Fraction of sequences with at least one motif (%)
A. thaliana		
Transcripts	0.66	67.68
Intergenic	0.27	-
Annotated genes in sense orientation	0.5	-
Antisense orientation to the annotated genes	0.51	-
Exons	0.83	21.22
Introns	0.21	3.34
S. pombe		
Transcripts	0.63	64.5

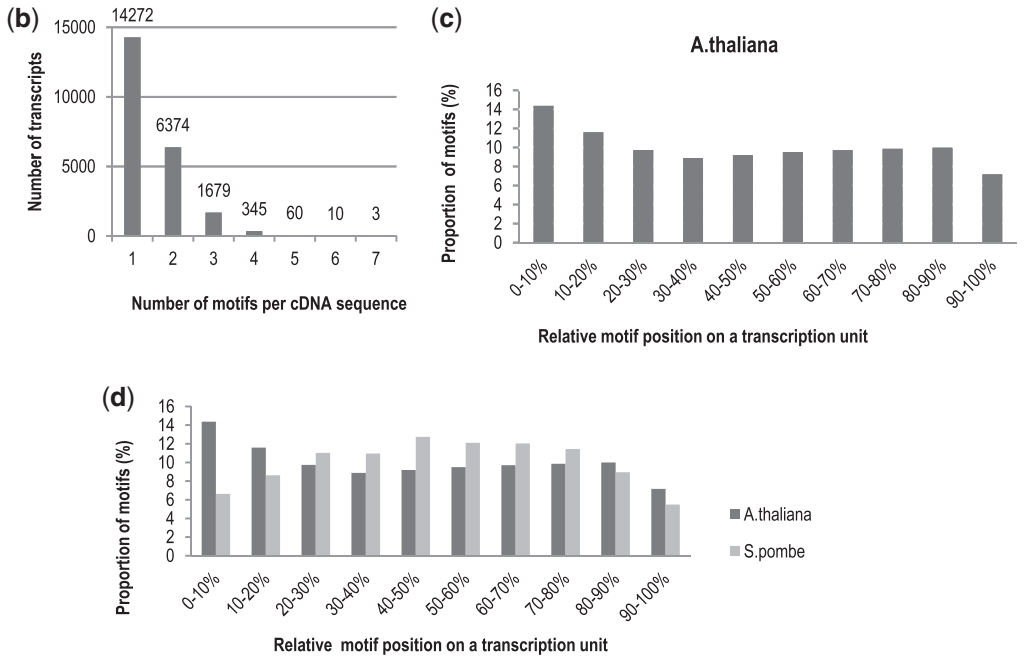


Figure 6. Computational analysis of the binding motif distribution in the A. thaliana genome. (**a**) Distribution of the AtCyp59 binding motif variants in the *A. thaliana* genome according to TAIR 10 annotation. (**b**) Frequency of occurrence of the binding motif variants per cDNA sequence in the *A. thaliana* genome**.** Each bar on the diagram represents the number of genes with the indicated number of the binding motifs per transcript sequence. (**c**) *Arabidopsis* AtCyp59 motif distribution in the TUs. Distribution of 17 motifs experimentally validated in *A. thaliana* is shown in 10 windows of the transcription units. (**d**) *A. thaliana* and *S. pombe* motif distribution. Distribution of the 17 motifs experimentally validated in *A. thaliana* is shown for *Arabidopsis* (dark grey) and *S. pombe* transcription units (light grey bars).

binding motif showed its presence in >70% of the *A. thaliana* mRNAs and its prevalent localization in the coding region. We also demonstrated that the RNA-binding motif and its genome-wide distribution are evolutionarily conserved.

Genomic SELEX is a method for the identification of an RNA-binding motif for a given protein. In our selection, the identified motif which consists of a 7-nt-long RNA consensus sequence G[U/C]N[G/A]CC[A/G], was consistent with previous data, which showed that AtCyp59 interacts with C- and G-rich RNA oligonucleotides *in vitro* (3). Sequence variations of the motif which were present in the SELEX sequences were tested *in vitro*. Interestingly, the binding motif variants showed variable binding affinities to the RRM-Zn domain of AtCyp59 which was used for selection. Furthermore, binding of the 7-nt binding motif variations to the full-length AtCyp59 showed an equal overall binding affinity to the motif variants, suggesting an influence of the other AtCyp59 domains in this interaction. AtCyp59 interacts with its RNA target sequence specifically, since mutations in the RRM domain of AtCyp59, which are known to be generally involved in RNA recognition (37), decreased binding activity. Similarly, mutations in the sequence of the 7-nt RNA consensus motif also decreased its affinity to the

RRM_Zn domain of AtCyp59. In addition, when we used longer RNA transcripts containing the binding motif, we observed an improved binding to the full-length AtCyp59. As we have indications that AtCyp59 possesses an RNA chaperone activity, this effect might be due to an impact on local RNA structure. Thus, full-length AtCyp59 regulates binding to RNA targets and might also change RNA structure.

The interaction between AtCyp59 and RNA targets was confirmed *in vivo* using transient expression of the AtCyp59 in protoplasts. We were restricted to transient expression for the experiments *in vivo* due to our inability to produce stable overexpression tagged lines of AtCyp59. This is most probably caused by the tight regulation of AtCyp59 levels *in vivo* (3,11). RNA immunoprecipitation analysis showed that wild-type AtCyp59 could immunoprecipitate endogenous RNA targets. Furthermore, mutations in the RRM domain of AtCyp59 abolished recovery of specific RNAs after immunoprecipitation. This suggests a direct recognition of RNA targets through the RRM domain of AtCyp59. When we performed a double transformation of the protoplasts with AtCyp59 and exogenous RNA targets, these RNAs could be also recovered by immunoprecipitation of AtCyp59. If the RNA-binding motif was mutated, AtCyp59 could no longer bind to the exogenous RNAs. These data show that AtCyp59 binds to motif containing mRNAs *in vivo* in a sequence-specific manner. Most importantly, these experiments also demonstrated that AtCyp59 binds to the transcript in the course of transcription as also unspliced or partially spliced RNAs were immunoprecipitated.

In addition, we could demonstrate PPIase activity of the recombinant AtCyp59 *in vitro*. The observed K_{obs} rates are in line with previously described PPIase activities of other isomerases (38). This also suggests that AtCyp59 could act as a PPIase enzyme *in vivo*. Most interestingly, the PPIase activity was reduced when the RRM domain of AtCyp59 was bound specifically to its RNA-binding motif or to polyA+ RNA but not to a scrambled RNA oligo. The mutations in the RRM domain of AtCyp59 slightly reduced its enzymatic activity similar to control mutations in the region between the RRM and the PPIase domain (data not shown). This minor decrease in activity might therefore come from changes in protein structure upon mutation in regions unrelated to the PPIase domain. However, if we added specific RNA to Cyp59_3M (mutated in the RRM), the PPIase activity did not change indicating that the decrease in activity upon RNA binding is exerted through the RRM domain. This suggests that binding of specific RNA to AtCyp59 causes structural changes that feed back to the PPIase domain resulting in a decrease in enzymatic activity. From the available data from AtCyp59 and spRct1 (3,11), we know that this protein is located in transcription/splicing complexes, but the mechanism of its action is unknown. There is one other RRM-containing cyclophilin, hCyp33, described in the literature which regulates a histone acetyl transferase. Contrary to AtCyp59, PPIase activity of hCyp33 is stimulated by binding AU-rich RNA (20) but neither a motif nor a natural RNA target has been identified (19). However, our results with AtCyp59 predict a possible negative feedback loop regulating PPIase activity upon RNA binding.

AtCyp59 is an interesting protein, because of its multidomain organization that includes a PPIase domain followed by an RRM domain, Zn knuckle and the C-terminal-charged domain. AtCyp59 is a nuclear protein and further analyses showed that AtCyp59 interacts *in vitro* and *in vivo* with the RNA polymerase II (3). Overexpression of the protein in cell culture was detrimental to cell growth and had an effect on the level of RNA polymerase II. Additional information about this highly conserved protein came from analysis of its *S. pombe* orthologue, Rct1, which was identified as an essential gene where overexpression of Rct1 led to a decrease in CTD phosphorylation. In contrast, lower levels of Rct1 upregulated CTD phosphorylation (11). This upregulation resulted in a decrease in RNA polymerase II activity as shown in run-on transcription assays. Furthermore, it was shown that Rct1 is associated with actively transcribed genes following the RNA polymerase II profile along the gene. These data suggested an important activity provided by this multidomain cyclophilin for transcription elongation. The RRM domain is evolutionarily the most conserved domain of this protein suggesting its important contribution to the overall activity of the protein. The experiments presented in this articles now show binding of AtCyp59 to the nascent transcript. In addition, they demonstrate that the selected binding motif, its abundance and genomic localization are also evolutionarily conserved.

We predicted that AtCyp59 might bind a small regulatory RNA (similarly to the P-TEFb transcription factor regulated by 7SK RNA (39)) or alternatively might bind a subset of RNA polymerase II transcripts. Surprisingly, our genome-wide analysis of the AtCyp59 RNA-binding motif showed that this motif is present in the majority of RNA polymerase II transcripts in sense and antisense orientation (Figure 6a). If one assumes a universal function for this cyclophilin in transcription, the binding motif should be also present in antisense transcripts as antisense transcription of protein coding genes is pervasive in many genomes (40,41). Furthermore, the binding affinities to the motif variants should be similar which is what we observe when using full-length AtCyp59 in binding assays. These data suggest that AtCyp59 might bind to many endogenous transcripts (sense and antisense) with similar strength. Most importantly, we have shown that binding of specific RNAs or polyA+ RNAs reduces the PPIase activity of AtCyp59 *in vitro*. Therefore, binding of AtCyp59 to its natural RNA targets might also change PPIase activity *in vivo* and thus send a signal to the transcription machinery. As AtCyp59 interacts with the RNA polymerase II complex, a possible scenario could be that it re-structures the heptapeptide repeats on the CTD thus influencing phosphorylation similarly to the activity of the PPIase Pin1 in humans and Ess1 in *Saccharomyces cerevisiae* (12,16,42,43). Alternatively, it might influence kinase or phosphatase activities which act on the CTD repeats. As the position of the binding motif is distributed

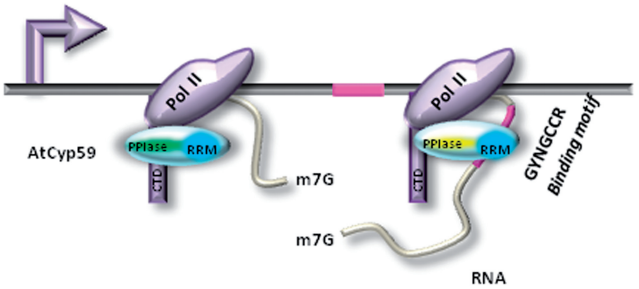


Figure 7. Model of how binding of AtCyp59 to the transcript might influence transcription. When the RRM domain (dark blue) of AtCyp59 (light blue) is not engaged in RNA–protein interaction, activity of its PPIase domain is unaffected (green). However, upon binding of AtCyp59 to the binding motif (purple) on the RNA transcript, the PPIase activity of AtCyp59 decreases (yellow). These changes might serve as a signal to modulate RNA polymerase II activity.

along the coding region of the gene, we suggest that AtCyp59 might bind to the nascent transcript in the elongation phase of transcription. The consequences of RNA binding on the function of this protein *in vivo* as well as on its temporal and spatial interaction with its partners are currently unknown. However, summarizing the available data for this cyclophilin, we propose a model (Figure 7) where in the course of transcription RNA-dependent inhibition of the PPIase activity of AtCyp59 influences RNA polymerase II activity. Our data indicate that this multidomain cyclophilin might have a role in transcription regulation, which is in line with the observation that it is an essential gene and its deregulation is detrimental to cell growth (3,11).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figures 1–4 and Supplementary Methods.

ACKNOWLEDGEMENTS

We thank Doris Chen for initial bioinformatics support and Zdravko Lorkovic, John WS Brown and Alwin Köhler for invaluable discussions. O.B. performed all experiments and analysed data; M.Z. and Y.M. did the computer analysis, motif finding and genome analysis; T.S. did the initial cycolphilin assays; M.K supervised experiments; A.B. designed and supervised the experiments; A.B, O.B. and M.K. wrote the article.

FUNDING

The Austrian Science Fund (FWF) [SFB 1710, 1711; DK W1207] and the Austria Genomic Program (GENAU III) [ncRNAs]; the EU FP6 Programme Network of Excellence on Alternative Splicing (EURASNET) [LSHG-CT-2005-518238]. Funding for open access charge: Austrian Science Fund, FWF.

Conflict of interest statement. None declared.

REFERENCES

1. Nicholson,L.K. and Lu,K.P. (2007) Prolyl cis-trans isomerization as a molecular timer in Crk signaling. *Mol. Cell.*, **25**, 483–485.
2. Romano,P.G., Horton,P. and Gray,J.E. (2004) The Arabidopsis cyclophilin gene family. *Plant Physiol*, **134**, 1268–1282.
3. 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.
4. Krzywicka,A., Keller,A.M., Cohen,J., Jerka-Dziadosz,M. and Klotz,C. (2001) KIN241: a gene involved in cell morphogenesis in *Paramecium tetraurelia* reveals a novel protein family of cyclophilin-RNA interacting proteins (CRIPs) conserved from fission yeast to man. *Mol. Microbiol.*, **42**, 257–267.
5. Bentley,D. (2002) The mRNA assembly line: transcription and processing machines in the same factory. *Curr. Opin. Cell Biol.*, **14**, 336–342.
6. Munoz,M.J., de la Mata,M. and Kornblihtt,A.R. (2010) The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem. Sci.*, **35**, 497–504.
7. Natalizio,B.J., Robson-Dixon,N.D. and Garcia-Blanco,M.A. (2009) The carboxyl-terminal domain of RNA polymerase II is not sufficient to enhance the efficiency of pre-mRNA capping or splicing in the context of a different polymerase. *J. Biol. Chem.*, **284**, 8692–8702.
8. Brody,Y., Neufeld,N., Bieberstein,N., Causse,S.Z., Bohnlein,E.M., Neugebauer,K.M., Darzacq,X. and Shav-Tal,Y. (2011) The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol.*, **9**, e1000573.
9. Proudfoot,N.J., Furger,A. and Dye,M.J. (2002) Integrating mRNA processing with transcription. *Cell*, **108**, 501–512.
10. Kornblihtt,A.R., de la Mata,M., Fededa,J.P., Munoz,M.J. and Nogues,G. (2004) Multiple links between transcription and splicing. *RNA*, **10**, 1489–1498.
11. Gullerova,M., Barta,A. and Lorkovic,Z.J. (2007) Rct1, a nuclear RNA recognition motif-containing cyclophilin, regulates phosphorylation of the RNA polymerase II C-terminal domain. *Mol. Cell Biol.*, **27**, 3601–3611.
12. Xu,Y.X., Hirose,Y., Zhou,X.Z., Lu,K.P. and Manley,J.L. (2003) Pin1 modulates the structure and function of human RNA polymerase II. *Genes Dev.*, **17**, 2765–2776.
13. Singh,N., Ma,Z., Gemmill,T., Wu,X., Defiglio,H., Rossettini,A., Rabeler,C., Beane,O., Morse,R.H., Palumbo,M.J. *et al.* (2009) The Ess1 prolyl isomerase is required for transcription termination of small noncoding RNAs via the Nrd1 pathway. *Mol. Cell*, **36**, 255–266.
14. Xu,Y.X. and Manley,J.L. (2007) Pin1 modulates RNA polymerase II activity during the transcription cycle. *Genes Dev.*, **21**, 2950–2962.
15. Poschmann,J., Drouin,S., Jacques,P.E., El Fadili,K., Newmarch,M., Robert,F. and Ramotar,D. (2011) The peptidyl prolyl isomerase Rrd1 regulates the elongation of RNA polymerase II during transcriptional stresses. *PLoS One*, **6**, e23159.
16. Ma,Z., Atencio,D., Barnes,C., Defiglio,H. and Hanes,S.D. (2012) Multiple roles for the Ess1 prolyl isomerase in the RNA polymerase ii transcription cycle. *Mol. Cell Biol*, **32**, 3594–3607.
17. Lorkovic,Z.J., Lopato,S., Pexa,M., Lehner,R. and Barta,A. (2004) Interactions of *Arabidopsis* RS domain containing cyclophilins with SR proteins and U1 and U11 small nuclear ribonucleoprotein-specific proteins suggest their involvement in pre-mRNA Splicing. *J. Biol. Chem.*, **279**, 33890–33898.
18. Hegele,A., Kamburov,A., Grossmann,A., Sourlis,C., Wowro,S., Weimann,M., Will,C.L., Pena,V., Luhrmann,R. and Stelzl,U. (2012) Dynamic protein-protein interaction wiring of the human spliceosome. *Mol, Cell*, **45**, 567–580.
19. Wang,Z., Song,J., Milne,T.A., Wang,G.G., Li,H., Allis,C.D. and Patel,D.J. (2010) Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression. *Cell*, **141**, 1183–1194.

20. Wang,Y., Han,R., Zhang,W., Yuan,Y., Zhang,X., Long,Y. and Mi,H. (2008) Human CyP33 binds specifically to mRNA and binding stimulates PPIase activity of hCyP33. *FEBS Lett*., **582**, 835–839.
21. Robertson,J.F., Semiglazov,V., Nemsadze,G., Dzagnidze,G., Janjalia,M., Nicholson,R.I., Gee,J.M. and Armstrong,J. (2007) Effects of fulvestrant 250mg in premenopausal women with oestrogen receptor-positive primary breast cancer. *Eur. J. Cancer*, **43**, 64–70.
22. Lorenz,C., von Pelchrzim,F. and Schroeder,R. (2006) Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nat. Protoc*., **1**, 2204–2212.
23. Zimmermann,B., Gesell,T., Chen,D., Lorenz,C. and Schroeder,R. (2010) Monitoring genomic sequences during SELEX using high-throughput sequencing: neutral SELEX. *PLoS One*, **5**, e9169.
24. Zywicki,M., Bakowska-Zywicka,K. and Polacek,N. (2012) Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis. *Nucleic Acids Res*., **40**, 4013–4024.
25. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*., **37(Web Server issue)**, W202–W208.
26. Frith,M.C., Saunders,N.F., Kobe,B. and Bailey,T.L. (2008) Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol*., **4**, e1000071.
27. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al*. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*., **40(Database issue)**, D1202–D1210.
28. Kofron,J.L., Kuzmic,P., Kishore,V., Colon-Bonilla,E. and Rich,D.H. (1991) Determination of kinetic constants for peptidyl prolyl cis-trans isomerases by an improved spectrophotometric assay. *Biochemistry*, **30**, 6127–6134.
29. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
30. James,T.D., Cashel,M. and Hinton,D.M. (2010) A mutation within the beta subunit of *Escherichia coli* RNA polymerase impairs transcription from bacteriophage T4 middle promoters. *J. Bacteriol*., **192**, 5580–5587.
31. Kim,S., Shi,H., Lee,D.K. and Lis,J.T. (2003) Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res*., **31**, 1955–1961.
32. Stamm,S., Smith,C. and Lührmann,R. (2012) *Alternative Pre-mRNA Splicing: Theory and Protocols*. Wiley, New York, p. 660.
33. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.
34. Carels,N. and Bernardi,G. (2000) The compositional organization and the expression of the *Arabidopsis* genome. *FEBS Lett*., **472**, 302–306.
35. Mayeda,A., Munroe,S.H., Caceres,J.F. and Krainer,A.R. (1994) Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins. *EMBO J*., **13**, 5483–5495.
36. Fischer,G., Bang,H. and Mech,C. (1984) [Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides]. (Translated from ger). *Biomed. Biochim. Acta*, **43**, 1101–1111, (in Germany).
37. 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.
38. Wiborg,J., O'Shea,C. and Skriver,K. (2008) Biochemical function of typical and variant *Arabidopsis thaliana* U-box E3 ubiquitin-protein ligases. *Biochem. J*., **413**, 447–457.
39. Eilebrecht,S., Benecke,B.J. and Benecke,A. (2011) 7SK snRNA-mediated, gene-specific cooperativity of HMGA1 and P-TEFb. *RNA Biol*., **8**, 1084–1093.
40. Chen,H.M. and Neiman,A.M. (2011) A conserved regulatory role for antisense RNA in meiotic gene expression in yeast. *Curr. Opin. Microbiol*., **14**, 655–659.
41. Ni,T., Tu,K., Wang,Z., Song,S., Wu,H., Xie,B., Scott,K.C., Grewal,S.I., Gao,Y. and Zhu,J. (2010) The prevalence and regulation of antisense transcripts in Schizosaccharomyces pombe. *PLoS One*, **5**, e15271.
42. Zhang,M., Wang,X.J., Chen,X., Bowman,M.E., Luo,Y., Noel,J.P., Ellington,F.A. and Zhang,Y. (2012) Structural and Kinetic analysis of prolyl-isomerization/phosphorylation cross-talk in the CTD code. *ACS Chem Biol*., **7**, 1462–1470.
43. Werner-Allen,J.W., Lee,C.J., Liu,P., Nicely,N.I., Wang,S., Greenleaf,A.L. and Zhou,P. (2011) cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. *J. Biol. Chem*., **286**, 5717–5726.

Complexity of the alternative splicing landscape in plants

Reddy, A. S., Marquez, Y., Kalyna, M., & Barta, A. (2013).

Plant Cell, 25(10), 3657-3683

The Plant Cell, Vol. 25: 3657–3683, October 2013, www.plantcell.org

REVIEW

Complexity of the Alternative Splicing Landscape in Plants[C][W][OPEN]

Anireddy S.N. Reddy,[a,1] Yamile Marquez,[b] Maria Kalyna,[b] and Andrea Barta[b]

[a] Department of Biology, Program in Molecular Plant Biology, Program in Cell and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523

[b] Max F. Perutz Laboratories, Medical University of Vienna, Vienna A-1030, Austria

Alternative splicing (AS) of precursor mRNAs (pre-mRNAs) from multiexon genes allows organisms to increase their coding potential and regulate gene expression through multiple mechanisms. Recent transcriptome-wide analysis of AS using RNA sequencing has revealed that AS is highly pervasive in plants. Pre-mRNAs from over 60% of intron-containing genes undergo AS to produce a vast repertoire of mRNA isoforms. The functions of most splice variants are unknown. However, emerging evidence indicates that splice variants increase the functional diversity of proteins. Furthermore, AS is coupled to transcript stability and translation through nonsense-mediated decay and microRNA-mediated gene regulation. Widespread changes in AS in response to developmental cues and stresses suggest a role for regulated splicing in plant development and stress responses. Here, we review recent progress in uncovering the extent and complexity of the AS landscape in plants, its regulation, and the roles of AS in gene regulation. The prevalence of AS in plants has raised many new questions that require additional studies. New tools based on recent technological advances are allowing genome-wide analysis of RNA elements in transcripts and of chromatin modifications that regulate AS. Application of these tools in plants will provide significant new insights into AS regulation and crosstalk between AS and other layers of gene regulation.

INTRODUCTION

Production of the right amount of protein in the right cells at the right time is crucial for growth and development of multicellular eukaryotes and their response to the environment. Hence, protein synthesis is tightly regulated with multiple layers of regulation. Transcriptional regulation of gene expression is a central component of this regulation. In recent years, it has become clear that regulation of cotranscriptional processes, such as splicing and polyadenylation, is also a major driving force of transcript complexity and abundance. Posttranscriptional gene regulation occurs at many levels, including transcript export, localization, mRNA stability, translation, posttranslational modifications of proteins, and protein stability and degradation, which ultimately dictate the amount and functionality of RNAs and proteins within the cell. The extent, complexity, and sophistication of posttranscriptional gene regulation are beginning to rival transcriptional regulation in eukaryotes. The first evidence for the significance of alternative splicing (AS) in plant development came from differential expression of Ser/Arg-rich (SR) protein splicing factors in different organs and during development (Lopato et al., 1996, 1999b; Kalyna et al., 2003; Palusa et al., 2007) indicating organ-specific regulation of AS in plants. Verification will come from genome-wide studies of AS in different organs and during development (Loraine et al., 2013) as has been shown in animals (Wang et al., 2008; Kalsotra and Cooper, 2011; Barbosa-Morais et al., 2012; Ellis et al., 2012). In addition, screens for mutants in various pathways have frequently identified splicing factors as modulators of functional proteins, indicating that these pathways are regulated via differential splicing (Lee et al., 2006; Monaghan et al., 2009; Sugliani et al., 2010; Fouquet et al., 2011; Koncz et al., 2012). There is an ever-growing body of literature on how alternative splicing (AS) influences important developmental and signaling pathways, and many important examples have been discussed in the accompanying review by Staiger and Brown (2013). This review focuses on the current knowledge on splicing and genome-wide AS in plants, its regulation and potential functions, and the important outstanding questions and tools to address these.

The mechanism of splicing has been elucidated mainly by in vitro assays and genetic studies in mammals and yeast. The lack of an in vitro splicing assay in plants has been a major limitation in studying the mechanisms involved in intron recognition and spliceosome assembly in plants (Barta et al., 2012). However, the advent of the genomic era and the availability of whole-genome sequences of several plants allowed the identification of orthologous proteins and small nuclear RNAs (snRNAs) of the core components of the spliceosome (Wang and Brendel, 2004; Barta et al., 2012; Koncz et al., 2012) (see Supplemental Table 1 online), suggesting that the main principles of intron processing are also applicable to plants. Nevertheless, the fact that animal introns cannot be processed in plants made it clear that there is some specificity in the plant spliceosomal machinery and in the plant intronic sequences for their successful splicing (Barta et al., 1986; Brown et al., 1986; Hartmuth and Barta, 1986). Even though some

[1] Address correspondence to reddy@colostate.edu.

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] Online version contains Web-only data.

[OPEN] Articles can be viewed online without a subscription.

www.plantcell.org/cgi/doi/10.1105/tpc.113.117523

animal introns are as small as or smaller than plant introns, plants do not have the long introns characteristic of many animal species. There is a clear difference in average size of introns between plants and animals; while animal introns can be quite long with an average size of 5 kb (Sakharkar et al., 2004), plant introns are relatively small with an average size of 160 bp (Iwata and Gotoh, 2011; Marquez et al., 2012), suggesting that recognition of exons and introns might differ in plants. The consensus sequences for the 5′ splice site (5′SS), 3′ splice site (3′SS), and branch point (BP) are comparable in animals and plants (Simpson et al., 2002; Reddy, 2007; Iwata and Gotoh, 2011), indicating that core splicing factors recognizing such sequences are acting in similar ways. However, in plants, the presence of uridine-rich sequences toward the 3′SS was found to be an important determinant of splicing efficiency as was a clear difference in the adenine-uridine (AU) and guanine-cytosine (GC) content of introns and exons especially in dicotyledons where introns are more than 60% AU rich (Goodall and Filipowicz, 1989). Differences can also be seen between different plant species; for example, there is a higher AU content in the introns of dicotyledons species in comparison with those in monocotyledons (Baek et al., 2008). Differences in the GC/AU content of exons and introns were lately reported to occur as well in mammals supporting a shared principle for exon and intron definition (Amit et al., 2012). The difference in GC/AU content might explain differences in nucleosome positioning of exons and introns that may influence RNA polymerase II elongation rates and have an effect on splicing (reviewed in Gómez Acuña et al., 2013).

SPLICEOSOME COMPOSITION AND ASSEMBLY

One of the major advances over the past years was the insight that most RNA splicing events occur cotranscriptionally and hence are dependent on DNA and chromatin structure and modifications (reviewed in Braunschweig et al., 2013). Up to 90% of protein-coding and many non-protein-coding genes contain introns in photosynthetic eukaryotes (Hirsch et al., 2006; Szarzynska et al., 2009; Labadorf et al., 2010). Generation of mature mRNAs from these genes requires precise removal of introns from precursor mRNAs (pre-mRNAs) and joining of exons. Four loosely conserved core sequence elements in introns, the 5′SS with a conserved GT, the 3′SS with a conserved AG, a BP with a conserved A residue 18 to 40 nucleotides upstream of the 3′SS, and a polypyrimidine tract following the BP, are necessary for spliceosome assembly. Splicing requires the precise recognition of splice sites and spliceosome assembly with numerous RNA–RNA, RNA–protein, and protein–protein interactions and conformational rearrangements of these interactions in a precise sequential manner. An in-depth discussion on spliceosome composition and dynamics during different stages of splicing in yeast and metazoans is presented in recent reviews (Will and Lührmann, 2011; Hoskins and Moore, 2012).

Plant spliceosomes have not been isolated so far; hence, their assembly and composition are not known. Extensive studies with yeast and human spliceosomes that are assembled in vitro on model pre-mRNA substrates have revealed that the spliceosome is a highly dynamic and complex macromolecular machine consisting of five snRNAs and over 300 proteins (Jurica and Moore, 2003; Will and Lührmann, 2011). The yeast *Saccharomyces cerevisiae* spliceosome has ~160 proteins (Fabrizio et al., 2009), and the homologs of all of these are present in the human spliceosome, indicating the conservation of core spliceosomal proteins across eukaryotes. However, human spliceosomes have over 100 additional proteins, which are thought to play a role in splice site selection and regulation of AS (Will and Lührmann, 2011).

The analyses of splicing homologs in *Arabidopsis* have led to the identification of almost twice the number of splicing regulatory factors in comparison to human (see Supplemental Table 1 online) (Wang and Brendel, 2004; Barta et al., 2012; Koncz et al., 2012). The presence of multiple paralogs in plants seems to be the result of whole-genome duplications followed by massive genomic rearrangements in the course of evolution (Kalyna and Barta, 2004; Adams and Wendel, 2005; Cui et al., 2006). The existence of multigene families of splicing factors including those involved in AS, like the multimember family of the SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) that share distinct levels of identity, raises the question of how many of the family members have redundant roles or have acquired new functions. The higher complexity of plant core spliceosomal proteins might be the reason that mutations in several proteins associated with plant small nuclear ribonucleoprotein particles (snRNPs) are not lethal but cause various defects in the development and response to environmental conditions (Staiger and Brown, 2013).

In metazoans, there are two types of spliceosomes: a major U2 type with U1, U2, U4, U5, and U6 snRNPs involved in splicing of most pre-mRNAs with GT-AG splice sites and a minor U12 type with U11, U12, U4atac, U5, and U6atac snRNPs that splice a small fraction of pre-mRNAs with so-called U12 type introns. These introns are characterized by a more conserved 5′SS and BP sequence and have both canonical (GT-AG) and noncanonical splice sites (Friend et al., 2008; Steitz et al., 2008). They are not very abundant with ~800 in human and mice and 300 in plants (Alioto, 2007); however, a recent deep-sequencing analysis resulted in identification of ~2000 introns with U12-type signatures in *Arabidopsis thaliana* (Marquez et al., 2012). These introns, which occur usually only once in a transcript, are spliced mainly cotranscriptionally with a slower kinetics. Furthermore, these are often evolutionary conserved and important for developmental processes (recently reviewed in Turunen et al., 2013). Interestingly, 20 U12-type introns are conserved in orthologous positions between human and *Arabidopsis* (Zhu and Brendel, 2003; Basu et al., 2008). Analysis of the large pool of newly discovered U12-type introns might change this number (Zhu and Brendel, 2003; Basu et al., 2008). Based on the presence of most of the RNA and protein homologs of mammalian spliceosomes (see Supplemental Table 1 online), it appears that plants, like animals, have both major and minor spliceosomes (Shukla and Padgett, 1999; Reddy, 2001; Wang and Brendel, 2004; Lorkovic et al., 2005; Ru et al., 2008; Simpson and Brown, 2008; Barta et al., 2012; Koncz et al., 2012).

CONSTITUTIVE SPLICING AND AS

A multiexon gene can produce a single mature mRNA by constitutive splicing (CS), where only one set of splice sites is used.

Failure to select or differential usage of some of these splice sites can lead to intron retention (IR) or exon skipping events. In addition, alternative splice sites (5′SS, 3′SS, or both) can be employed, and all these events generate multiple mRNA isoforms from the same gene (Figure 1) (reviewed in Reddy, 2007; Syed et al., 2012). A given splice variant may show more than one of these events. The long-standing question of how higher organisms achieve their complexity despite having similar gene numbers is at least partly explainable by the complexity of RNA transcripts created by AS. Deep sequencing of human transcriptomes from different organs, tissues, and cell lines indicates that almost every intron-containing gene generates multiple splice variants under certain conditions, and many AS events are regulated in a cell-, tissue-, or condition-specific manner (Pan et al., 2008; Wang et al., 2008; Merkin et al., 2012). Furthermore, recent studies indicate that the extent of AS increases as the complexity of tissues and species increases (Barbosa-Morais et al., 2012; Merkin et al., 2012).

The discovery of alternative transcripts in many organisms, including plant species, has been accomplished mainly using ESTs. The estimations of AS based on ESTs suggested that 11 to 30% of the multiexonic genes were alternatively spliced in *Arabidopsis* (Iida et al., 2004; Campbell et al., 2006; Barbazuk et al., 2008). Nevertheless, the discovery of new isoforms depended highly on the depth of the EST and cDNA libraries. Innovations in sequencing RNA (RNA-seq) led to the generation of millions of short sequence reads of cDNAs derived from poly(A)-enriched mRNA from different tissues or conditions in a cost effective manner, which were then mapped onto unique locations on the genome to predict AS events (Wang et al., 2009a; Filichkin et al., 2010; Hallegger et al., 2010; Marquez et al., 2012). This has been used to analyze the extent of AS and to identify novel introns and splice variants in *Arabidopsis* (Filichkin et al., 2010; Marquez et al., 2012; Syed et al., 2012), rice (*Oryza sativa*; Zhang et al., 2010b), cucumber (*Cucumis sativus*; Guo et al., 2010), grape (*Vitis vinifera*; Zenoni et al., 2010), and *Brachypodium distachyon* (Walters et al., 2013). These studies have revealed that up to 60% of intron-containing genes are alternatively spliced.

Analysis of different AS events in *Arabidopsis* and several other plants indicated that IR is the predominant mode of AS in land plants (Ner-Gaon et al., 2004; Wang and Brendel, 2006; Baek et al., 2008; Filichkin et al., 2010; Marquez et al., 2012; Syed et al., 2012). In metazoans, the predominant AS event is exon skipping and IR is least prevalent (Figure 1), suggesting differences in regulation of splice site recognition in metazoans and plants. Although IR is the most prevalent AS event in *Arabidopsis*, more than 51% of genes producing AS transcripts do not show IR events (Marquez et al., 2012). This indicates that the significance of IR for AS in plants has been overestimated. The analysis also showed that many IR events have very low sequence coverage and some annotated IR may not exist. A recent study has indicated that most of the IR events in *Arabidopsis* that are predicted targets of nonsense-mediated decay (NMD) escape this mechanism, provoking the question as to the fate of these transcripts and the function of gene expression regulation by IR events (see below) (Kalyna et al., 2012). One answer comes from investigations into gametophyte development in the fern *Marsilea vestia*, where it was shown that IR transcripts are stored in the spore until their splicing and translation are initiated upon spermatogenesis in a regulated temporal pattern (Boothby et al., 2013). In mammals, regulated IR in neuron-specific genes was shown to inhibit the export of partially spliced mRNAs to the cytoplasm and trigger their degradation in the nucleus independently of the NMD machinery and to play an important role during neurogenesis (Yap et al., 2012). It is possible that regulated retention of introns in the nucleus is widespread and serves to store precursor RNAs. These results are exciting as they indicate that IR might be used to regulate other developmental processes or to enhance the response to environmental signals.

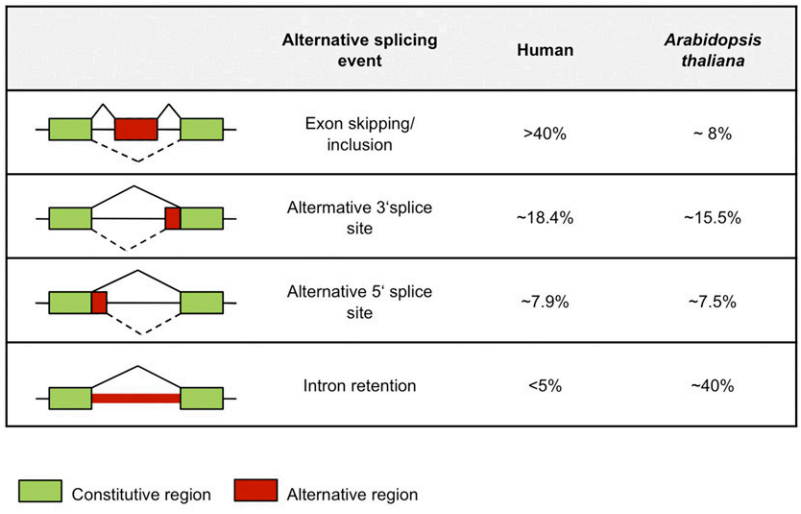


Figure 1. Frequency of Common Types of AS Events in Humans and *Arabidopsis*.

Proportion of common types of AS events in human (Keren et al., 2010; Reddy et al., 2012b) and *Arabidopsis* (Barta et al., 2012; Marquez et al., 2012). [See online article for color version of this figure.]

Cell-, tissue-, and condition-specific AS is highly prevalent in animals, and the physiological significance of many such events has been confirmed (Pan et al., 2008; Kalsotra and Cooper, 2011). So far, only limited global studies on AS with different cell types and tissues and under different conditions have been performed in plants that show regulated AS in different tissues or in response to stresses (Jiao and Meyerowitz, 2010; Loraine et al., 2013; reviewed in Staiger and Brown, 2013). As more RNA-seq studies are now being performed with different tissues and individual cell types and under different biotic and abiotic stresses, our knowledge of AS in plants is likely to further increase in the near future.

AS at Tandem Acceptor Sites

Numerous intron-containing genes in animals and land plants have tandem acceptor sites (3′SS) that are separated by three nucleotides (NAGNAG; N is any nucleotide) (Hiller et al., 2004; Iida et al., 2008; Schindler et al., 2008; Sinha et al., 2010). AS at these acceptor sites that preserve the reading frame generates mRNA isoforms that include or exclude a triplet thereby contributing to transcriptome and proteome diversity. A majority of NAGNAG splicing events lead to an insertion or a deletion of a single amino acid, mostly polar amino acids (Hiller et al., 2004; Iida et al., 2008; Bradley et al., 2012). Global RNA-seq studies in humans and mouse in 24 tissues have shown that ~20% of reading frame–preserving AS events are due to AS of tandem acceptors. Furthermore, about one-quarter of alternatively spliced NAGNAG sites show tissue-specific regulation in mammals, suggesting that NAGNAG AS contributes to differences in proteomes of different cell types. Furthermore, tissue-specific NAGNAG AS is highly conserved (Bradley et al., 2012). In *Arabidopsis*, close to 7000 introns in over 5000 genes contain tandem acceptors and occur predominantly in DNA binding proteins and splicing regulators such as SR and SR-like proteins (Iida et al., 2008; Schindler et al., 2008). There is evidence for both mRNA isoforms for several hundred genes that contained tandem acceptor sites in *Arabidopsis*, rice, and moss (Iida et al., 2008; Schindler et al., 2008; Sinha et al., 2010). Interestingly, NAGNAG AS in pre-mRNAs of *Arabidopsis* splicing regulators is regulated in an organ- and stress-specific manner (Schindler et al., 2008). The two isoforms of *Arabidopsis* U1-35K generated by alternative usage of a tandem acceptor differ in the presence or absence of Gln at position 124. The isoform lacking Gln showed altered binding affinity for SR proteins and U11 snRNA (Lorkovic et al., 2005). AS at a tandem acceptor site of ZINC-INDUCED FACILITATOR-LIKE1 produces two mRNA isoforms that differ in two nucleotides. One of these codes for a full-length protein that localizes to the plasma membrane and functions in auxin-regulated processes, whereas the second isoform codes for a truncated protein that localizes to the tonoplast membrane and functions in drought tolerance (Remy et al., 2013). These studies suggest the functional significance of AS as a result of tandem acceptor in plants.

EXPERIMENTAL AND COMPUTATIONAL APPROACHES FOR THE ANALYSIS OF AS

Mechanisms that regulate AS represent a complete black box in plants, as do the roles of the diverse set of RNA binding proteins (RBPs) in regulated AS. Numerous studies with metazoans and a few studies from plants suggest that multiple regulatory mechanisms determine the splicing outcome. Growing evidence points to elaborate crosstalk among multiple layers of posttranscriptional regulation and between transcription and chromatin. Chromatin landscape changes, such as nucleosome positioning, histone marks, and DNA methylation, together with RNA structural features and splicing regulatory elements (SREs) are important regulators of splicing. How important are RNA secondary and tertiary structural features in regulating CS and AS? There are a few examples in plants, but global studies of pre-mRNA/mRNA structures and correlating the information into AS should shed some light on this. Hence, to understand regulation of AS, it is necessary to integrate the impact of chromatin landscapes changes such as DNA methylation, nucleosome positioning, histone marks, RNA structural features, and SREs with AS. Genome-wide high-throughput tools to address these aspects have been developed recently (Figure 2, Table 1). Application of these tools to plant splicing research should provide novel insights into regulatory mechanisms and contribute to decoding of the plant splicing code. Performing such analyses with a single cell type or tissue would be ideal for such integrative studies. This would also require new computational tools to integrate data from these different sets of experiments.

The consequence of AS is the production of multiple mRNA isoforms from a single gene, which could be regulated by cell and tissue type, various developmental stages, and/or environmental conditions. However, as AS events cannot be easily predicted computationally from the genome sequence, a complete picture of all the possible mRNA isoforms generated from different genes of an organism requires a full exploration of AS in different tissues and diverse conditions and the availability of an accurately annotated genome. Since the first genome-wide analysis of AS, the estimates of AS have increased considerably (e.g., in humans, 5 to 35% [Lander et al., 2001] to now more than 95% [Mortazavi et al., 2008; Pan et al., 2008; Wang et al., 2008; Barbosa-Morais et al., 2012; Merkin et al., 2012]; in *Arabidopsis*, 1.2% (Zhu et al., 2003) to now 61% [Marquez et al., 2012]), indicating that estimates of AS mainly rely on the amount and diversity of the transcriptome data that are available from an organism.

The advent of high-throughput sequencing of RNA (RNA-seq) and large-scale microarray technologies (splice junction arrays and the tiling arrays) have allowed us to obtain a better picture of an organism's transcriptome complexity in tissues and under certain conditions (Johnson et al., 2003; Mortazavi et al., 2008; Pan et al., 2008; Sultan et al., 2008; Wang et al., 2008). The main difference of these two approaches is that with the arrays, it is necessary to have prior knowledge of the alternatively spliced variants, whereas RNA-seq allows de novo identification of transcripts at a single-base resolution (Figure 3). Despite the obvious advantages of these technologies in transcriptome analysis, several obstacles must be overcome, including the storage of the huge amount of data that is being generated (especially in the case of high-throughput sequencing) and, most importantly, a reliable pipeline for the analysis of the data and accurate prediction of AS events and splice variants (Reddy et al., 2012b). Sequencing a plant transcriptome poses another challenge: Just 10 highly expressed genes encoding proteins mainly involved in photosynthesis can generate about a quarter of all transcripts in a cell (Weber, 2007), tremendously compromising the representation of less abundant

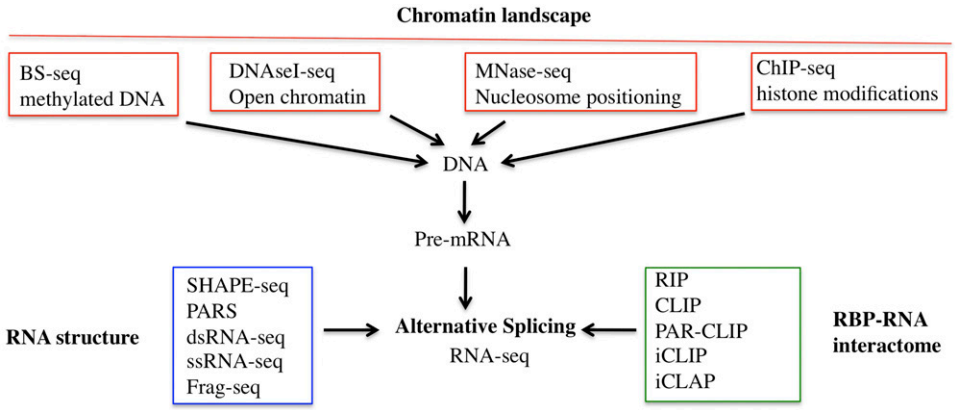


Figure 2. Integration of Chromatin Landscape, RNA Structure, and RBP-RNA Interactome into AS to Obtain a Comprehensive View of AS Regulation in Plants.

Genome-wide high-throughput sequencing tools that can provide insightful information on chromatin landscape, RNA structure, and RBP-RNA interactome are indicated. Integration of these results into AS will allow understanding of crosstalk between AS and other layers of gene regulation and mechanisms of AS. BS-seq, bisulphite sequencing; MNase-seq, micrococcal-nuclease sequencing; ChIP-seq, chromatin-immunoprecipitation sequencing; SHAPE-seq, selective 2′-hydroxyl acylation analyzed by primer-extension sequencing; dsRNA-seq, double-stranded RNA sequencing; ssRNA-seq, single-stranded RNA sequencing; Frag-seq, fragmentation sequencing; RIP-seq, RNA-immunoprecipitation sequencing; PAR-CLIP, photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; iCLIP, individual-nucleotide resolution cross-linking and immunoprecipitation; iCLAP, individual-nucleotide resolution UV-cross-linking and affinity purification. Table 1 provides a brief description of these tools.
[See online article for color version of this figure.]

transcripts in RNA-seq analysis. Thus, deeper sequencing with more reads is required to achieve good gene coverage throughout the genome. Using a normalized *Arabidopsis* cDNA library allowed reducing the abundance of highly expressed transcripts to ~1% and improved gene representation and discovery of AS events by RNA-seq (Marquez et al., 2012).

An RNA-seq experiment starts with the preparation of a cDNA library from mRNA extracted under certain conditions and/or tissue of interest followed by high-throughput sequencing, read alignment, transcript assembly, and transcript quantification (Figure 3). Sequencing results in the generation of millions of reads in the range of 36 to 400 bp, and they can be single-end (only one end of the mRNA fragment is sequenced) or paired-end (both ends of the mRNA fragment are sequenced) reads. The advantage of obtaining hundreds of millions of reads in one run is that it allows a high-resolution view of the transcriptome and the detection of low-abundance transcripts, but the disadvantage lies in the length of the sequenced fragments and the considerable error rates when compared with former sequencing methods. The short length of the sequencing reads is a critical difference in comparison to Sanger sequencing of ESTs or cDNAs that were used previously for predicting AS. The features of high-throughput sequencing (i.e., huge amounts of data and the high error rates and the short reads) created a computational challenge in terms of the alignment and prediction of full-length transcripts, especially when multiple AS events can co-occur in the same transcript or when highly similar paralogs are present in the genome, both of which are common in plant genomes.

Several computational tools have been developed in recent years for the prediction of splice junctions, which perform gapped alignment and detect exon-intron boundaries. These include QPALMA (Jean et al., 2010), Tophat (Trapnell et al., 2009), Genomic Short-Read Nucleotide Alignment Program (Wu and Nacu, 2010), MapSplice (Wang et al., 2010a), RNA-seq Unified Mapper (Grant et al., 2011), and SpliceMap (Au et al., 2010). Most of the programs available are based on splice junction prediction using a reference genome or gene annotation, and only a few tools are available that allow transcriptome prediction without relying on an annotated genome (e.g., TransABySS [Robertson et al., 2010], Trinity [Haas et al., 2013], and Rnnotator [Martin et al., 2010]). For those programs that use the genome as a scaffold, one of two main strategies is used. One is the exon-first approach, and the second one is called seed-extend approach (Garber et al., 2011). The exon-first approach starts with aligning reads to the genome, and then the remaining unaligned reads are chopped and realigned (e.g., Tophat [Trapnell et al., 2009], MapSplice [Wang et al., 2010a], RNA-seq Unified Mapper [Grant et al., 2011], and SpliceMap [Au et al., 2010]). The seed-extend method starts by splitting the reads followed by their alignment and seed extension (Genomic Short-Read Nucleotide Alignment Program and QPALMA). Both methods assume that most of the gaps that separate the spliced reads when aligned to the genome represent introns. At this point, it is easy to imagine that chopping of the reads into smaller pieces for their alignment represents a challenge in terms of uniqueness of their genomic location, especially when the reads are coming from paralogs with high sequence identity. For this reason, the use of shorter reads when sequencing a very complex transcriptome with a highly repeated genome can result in a very high level of false-positive splice junctions and/or make it impossible to detect splice junctions at certain genomic loci. Recent improvements in obtaining longer reads using high-throughput platforms will improve the spliced alignment reliability. Nevertheless, the use of the

Table 1. Next-Generation Sequencing Tools to Analyze Global Changes in Chromatin Landscape, AS, RNA Structure, and Targets of RBPs

Method	Description
A. AS	
RNA-seq	cDNA generated from polyadenylated mRNA is fragmented and sequenced by next-generation sequencing. Alignment of reads to the genome allows quantitative analysis of gene expression, AS, and relative quantification of mRNA isoforms (Wang et al., 2008).This tool is also used to sequence other types of RNA.
B. Chromatin landscape	
Bisulphite sequencing	Genomic DNA treated with bisulphite, which converts unmethylated cytosines to uracils, is amplified and sequenced using high-throughput methods. In the sequenced reads, unmethylated cytosines appear as thymines and methylated cytosines appear as cytosines (Cokus et al., 2008). Provides a genome-wide cytosine methylation map at single-base-pair resolution.
DNAseI-seq	Nuclei are treated with DNase I, and short sequence tags are obtained, size selected, and sequenced using next-generation sequencing. The short reads are then mapped to the genome and clustered computationally to identify DNAase I hypersensitive sites (Boyle et al., 2008; Hesselberth et al., 2009). Allows genome-wide mapping of open chromatin, which is hypersensitive to DNAase I.
Micrococcal-nuclease sequencing	Isolated chromatin is digested with MNase to remove DNA that is not wrapped around nucleosomes, and DNA from the nucleosomes is extracted and subjected to deep sequencing. Alignment of reads to the genome provides global nucleosome occupancy (Barski et al., 2007; Chodavarapu et al., 2010).
Chromatin-immunoprecipitation sequencing	Chromatin is immunoprecipitated with an antibody specific to a histone modification or a DNA binding protein, and the DNA from the precipitate is extracted, amplified, sequenced on a next-generation sequencing platform, and mapped to the genome (Furey, 2012). Used to analyze genome-wide protein–DNA interactions and histone modifications.
C. Pre-mRNA/mRNA structure*	
Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) sequencing	RNA is treated with a SHAPE reagent (e.g., 1-methyl-7-nitroisatoic anhydride, IM) that forms 2'-*O*-adducts with unpaired nucleotides and blocks the extension of cDNA during reverse transcription, resulting in termination of synthesis. All generated fragments are sequenced and mapped to RNA to determine its structure (Lucks et al., 2011). Allows simultaneous measurement of secondary and tertiary structures of large numbers of RNAs in vitro.
PARS sequencing	Isolated mRNAs are renatured in vitro, treated with structure-specific enzymes (RNAse V1, which preferentially cleaves phosphodiester bonds 3′ double-stranded RNA or RNAase S1 that cleaves single-stranded RNA) in parallel, and the resulting RNA fragments are sequenced using next-generation sequencing (Kertesz et al., 2010; Wan et al., 2013). Allows in vitro profiling of the secondary structures of RNA at single-nucleotide resolution globally.
Double-stranded RNA sequencing	RNAs are treated with single-strand- specific RNAases (RNAse ONE) to remove single-stranded regions and select for double-stranded RNA, which is then converted into cDNA for deep sequencing (Zheng et al., 2010). Reads are then used to obtain structural information.
Single-stranded RNA sequencing	RNAs are treated with double-strand–specific RNAases (RNAseV1) to remove double-stranded regions in RNA and select for single-stranded regions of RNA. These are then converted into cDNA for deep sequencing. These reads together with double-stranded RNA sequencing are then used to identify structural elements in RNA (Li et al., 2012).
Fragmentation sequencing	RNA mixture is treated with nuclease P1, which specifically cleaves single-stranded RNA followed by RNA-seq to deduce transcriptome-wide RNA structure (Underwood et al., 2010).
D. Targets of RBPs	
RIP sequencing	RNAs bound to a specific RBP are immunoprecipitated, and the RNA from IP is converted into cDNA and sequenced using a high-throughput sequencing platform (Zhao et al., 2010). The reads are then used to identify direct and indirect targets of that RBP.
HITS-CLIP	RBP–RNA interactions are first covalently cross-linked by exposing cells/tissues to UV (254 nm). An RBP-specific antibody is used to precipitate RNA bound to RBP, which is then treated with RNAse and used for RNA-seq. iCLIP (individual-nucleotide-resolution UV cross-linking and immunoprecipitation) is a variation of HITS-CLIP that exploits premature termination of reverse transcriptase at the cross-link site, thereby providing better resolution of binding sites (König et al., 2010). Allows genome-wide mapping of direct binding sites of an RBP on RNA (Licatalosi et al., 2008).
Individual-nucleotide-resolution UV-cross-linking and affinity purification	Double-tagged RBP is expressed in cells. Following UV cross-linking and RNAase digestion, the RBP-RNA complex is purified by a two-step affinity purification. RNA from these complexes is then deep sequenced (Wang et al., 2010b). Allows purification of RBP-RNA complexes.
Photoactivatable-ribonucleoside-enhanced CLIP	Similar to HITS-CLIP except that RNA is labeled with 4-thiouridine and cross-linking is done using UV light of 365 nm (Hafner et al., 2010b). This allows efficient cross-linking of RBPs with RNA, and identification of nucleotides involved in binding can be identified as the cross-linked regions show T-to-C mutations in the reads. A similar approach (photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) is also used to identify all RBPs bound to mRNAs and to create binding maps of multiple RBPs. In this case, mRNAs cross-linked with RBPs are isolated using oligo(dT) affinity purification and digested with RNAse followed by identification of proteins by mass spectrometry and RNA by RNA-seq (Baltz et al., 2012).

*These in vitro methods are based on high-throughput RNA-seq and can be applied to all types of RNAs.

pair-end sequencing platform adds additional information about the mRNA fragment, as it also provides the sequence of both ends of a fragment. Based on the fragmentation pattern of the RNA or cDNA prior to sequencing, it is possible to estimate the minimum and maximum distance between the paired reads, which in turn limits the separation of both reads when they are aligned to the genome. This distance restriction improves the accuracy of the read alignment, especially when one of the matched reads is coming from a repetitive sequence.

After aligning the reads, genes that are expressed in the sample and the possible transcripts that can be explained by the RNA-seq data are identified. Initially, few studies attempted to predict novel transcripts and most focused on quantifying the expression of known genes and isoforms and measuring specific types of AS events. Several programs were designed for quantifying only exon-skipping events, as this is the most predominant type of AS in mammalian systems (Keren et al., 2010). In the case of plants, such tools for AS quantification are not ideal because the distribution of AS events is completely different and exon skipping is rare (Figure 1). Nevertheless, using RNA-seq (Marquez et al., 2012) and high-resolution RT-PCR (Kalyna et al., 2012), numerous AS events have been identified and confirmed in *Arabidopsis*. The main problem with the approaches that focus on measuring specific types of AS events is the limitation in quantifying the complete set of AS events under certain condition(s), limiting the biological conclusions of AS regulation.

It is well established that the number of reads that align to a genomic sequence directly correlates to its expression. Three main problems arise when quantifying transcript levels: the existence of reads that do not align uniquely, the presence of alternatively spliced transcripts, and the determination of the functional outcome of the different isoforms. For the first problem, programs such as Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010) consider the multiple locations of different isoforms. However, a major difficulty is that most of the features can be shared between different transcripts (e.g., constitutive exons) and often very subtle differences are not shared (e.g., a three-nucleotide difference due to the presence of NAGNAG sites), which limits considerably the assembly of possible transcripts and their quantification. The use of paired end reads can alleviate to some extent the problem of exon connectivity, as there is additional information about the distance between the matched pairs. Moreover, assemblers like Cufflinks (Trapnell et al., 2010) and Scripture have also implemented the prediction of novel isoforms based on known gene models, which considerably improves transcript assembly. Recently, Rogers et al. (2012) developed SpliceGrapher, a method that can combine gene models, EST, and RNA-seq data to predict AS events and visualize splicing graphs in *Arabidopsis*, thereby improving the detection of novel splicing variants. Finally, a reliable determination of transcripts can considerably improve the prediction of isoforms' contribution to proteome diversity and to the regulation of mRNA levels. The combination of RNA-seq together with large-scale proteomic technologies will be valuable in addressing this question.

The ultimate goal is accurate detection and quantification of different mRNA isoforms and the investigation of the impact on the biological system in different conditions or developmental stages. The advent of high-throughput sequencing has provided the ability to detect novel and lowly expressed transcript isoforms and to explore transcriptome regulation across different samples. In mammals, the use of RNA-seq has been applied multiple times revealing tissue-specific AS (Mortazavi et al., 2008; Pan et al., 2008; Wang et al., 2008; Barbosa-Morais et al., 2012; Merkin et al., 2012), but only a few studies are available for plant systems (Jiao and Meyerowitz, 2010; Loraine et al., 2013). In the near future, the low sequencing costs will not only improve the exploration of plant transcriptomes and AS but also help in understanding the still unknown plant splicing code.

CONSEQUENCES OF AS

AS changes the sequence of an RNA transcript, which can influence gene expression on several levels. It either generates distinct protein isoforms with altered (related, distinct, or opposing) functions or determines the fate of the RNA transcript by regulation of mRNA stability through NMD- and microRNA (miRNA)–mediated mechanisms, or by influencing mRNA transport, localization, and translational efficiency. In plants, there are several examples where protein isoforms from alternatively spliced pre-mRNAs are targeted to different cellular locations (Dixon et al., 2009; Lamberto et al., 2010; Remy et al., 2013). mRNA isoforms may differ in the untranslated regions (UTRs) with functional implications in transcript localization, stability, or translation.

Creating Protein Isoforms

Even though increasing proteome diversity is one of the important postulated roles of AS, the extent of its impact is far from being settled (Severing et al., 2009). The consequences of producing protein isoforms are manifold depending on the AS event. In some instances, AS events can be in frame. It is interesting to note that gene ontology enrichment analysis of alternatively spliced genes with NAGNAG splice sites indicates that members of some functional groups, such as RNA binding splicing regulators and transcriptional factors, are overrepresented (Iida et al., 2008; Severing et al., 2012). Other AS events can introduce frame shifts and result in truncated protein isoforms. These protein isoforms might still have functional modules, might lack certain functional domains, or lack and gain sites of posttranslational modifications. As most transcriptional factors and splicing regulators are modular proteins with multiple functional modules (e.g., nucleic acid binding domain, protein–protein interaction domain, and transcription activation domain), it is likely that mRNA isoforms that produce truncated proteins containing some, but not all, modules may still perform part of their function and hence may act as dominant-negative regulators. Several recent studies provide evidence in support of such a dominant regulatory role for truncated proteins produced by splice variants in plants (Gagete et al., 2009; Penfield et al., 2010; James et al., 2012; Seo et al., 2012; Wang et al., 2012; Liu et al., 2013). For instance, in several cases where proteins form homo- or heteromers, splice variants produce small interfering peptides that participate in dimerization but lack other functional domains, which leads to formation of nonfunctional dimers that can compete with functional dimers resulting in dominant-negative regulation (Staudt and Wenkel, 2011; Seo et al., 2013).

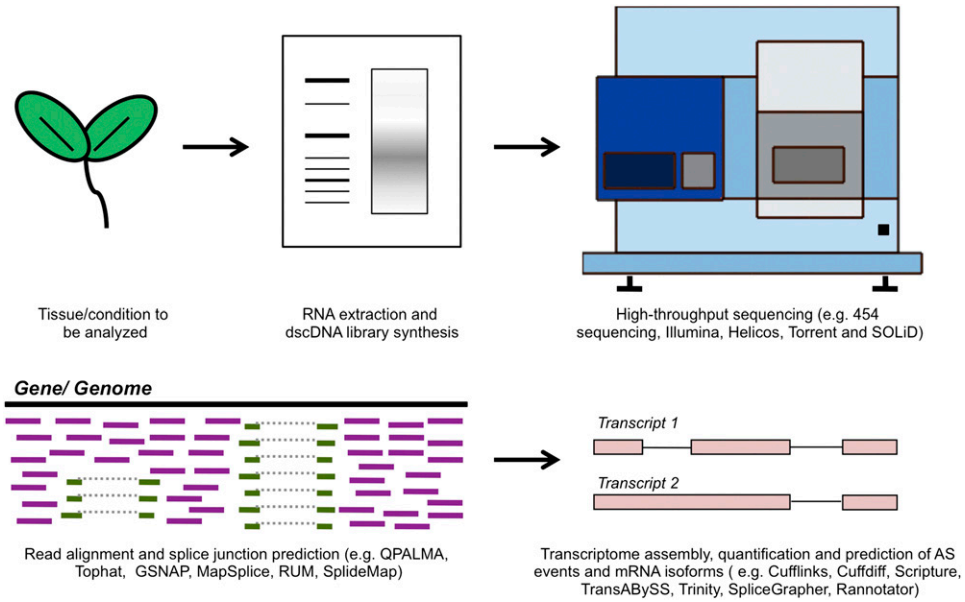


Figure 3. General Pipeline for the Analysis of AS Using High-Throughput Sequencing.

Although in-depth analyses of the consequences of AS isoforms as well as alternative transcription initiation or polyadenylation in plants are lacking, their potential is highlighted by recent global analyses in *S. cerevisiae* and animals. Global transcript isoform profiling in yeast revealed extensive transcriptional heterogeneity primarily due to alternative transcript initiation or polyadenylation with more than 26 major transcript isoforms per protein-coding gene (Pelechano et al., 2013). This study uncovered many short coding RNAs with a cap and poly(A) tail that may encode short peptides (Pelechano et al., 2013). Small bioactive peptides have been shown to perform important functions in morphogenesis by controlling the activity of a transcription factor (Kondo et al., 2010). Global studies in metazoans using computational and experimental methods investigating tissue-specific splice variants suggest that their protein products may rewire protein–protein interaction networks and impact signaling pathways (Buljan et al., 2012; Ellis et al., 2012; Weatheritt et al., 2012; Weatheritt and Gibson, 2012). It was found that alternative exons encode conserved intrinsically disordered regions that generally reside on protein surfaces and participate in protein–protein interaction, leading to rewiring of the protein–protein interaction network and signaling pathways (Buljan et al., 2012; Ellis et al., 2012; Weatheritt and Gibson, 2012). Furthermore, proteins from mRNA isoforms often gain or lose potential phosphorylation sites; hence, AS impacts protein kinase–mediated signaling pathways (Merkin et al., 2012). In *Arabidopsis*, SR and SR-like proteins RS41, CypRS64, and SR45a were found to have phosphorylation sites specific for single splice variants (de la Fuente van Bentem et al., 2006).

Coupling of AS to mRNA Stability through NMD

AS can affect transcript abundance by subjecting it to NMD. NMD is a cytoplasmic RNA degradation system, which occurs on ribosomes on the first round of translation. According to current models (Figure 4), NMD triggers transcript decay when translation termination is perturbed. UPF1, UPF2, and UPF3 are conserved core proteins of the NMD pathway, and their depletion results in the inactivation of NMD. They recruit mRNA decay enzymes to prematurely terminating mRNAs. For a detailed description and discussion of the NMD pathway, we direct readers to a number of excellent recent reviews (Behm-Ansmant et al., 2007; Isken and Maquat, 2007; Stalder and Mühlemann, 2008; Nicholson et al., 2010; Kervestin and Jacobson, 2012). The NMD machinery is largely conserved in plants, which have orthologs of the key NMD factors, including UPF1, UPF2, UPF3, and SMG-7, but not SMG-1, SMG-5, or SMG-6 (Hori and Watanabe, 2005; Arciga-Reyes et al., 2006; Yoine et al., 2006; Kerényi et al., 2008; Riehs et al., 2008). Though the rules of NMD in plants have been established mostly using artificial constructs and a handful of individual genes (Kertész et al., 2006; Schwartz et al., 2006; Hori and Watanabe, 2007; Kerényi et al., 2008; Nyikó et al., 2009), recent analysis of multiple endogenous NMD sensitive transcripts has shown that the majority comply with existing NMD rules (Kalyna et al., 2012). The classical features of NMD substrates include both long 3′UTRs and introns in 3′UTRs (Kertész et al., 2006; Schwartz et al., 2006; Hori and Watanabe, 2007; Wu et al., 2007), a premature termination codon (PTC) more than 50 to 55 nucleotides upstream of a splice junction (Kerényi et al., 2008), and upstream open reading frames (uORFs) (Nyikó et al., 2009) (Figure 5). All these features can be introduced or changed by AS, thereby altering the stability of a transcript. Therefore, AS transcripts that are subjected to NMD are often barely detectable. However, in mutants affecting NMD or in the presence translation inhibitors, there can be a considerable increase in these transcripts. The existence of PTC-containing transcripts might therefore significantly influence the level of gene expression by reducing fully spliced mRNA levels.

It is not commonly acknowledged that AS in the UTRs of a gene (which does not create a PTC) may trigger NMD. Recent data show that AS of introns in either the 3′UTR or 5′UTR can determine whether transcripts are targets of NMD (Figure 5). Examples of AS-NMD in 3′UTRs include NF-YB1/HAP3A transcription factor and PAUSED/Exportin-t genes, where it has been shown that only the splice variants with a distance >50 to 55 nucleotides from the authentic stop codon to the downstream splice junction were subjected to NMD (Kalyna et al., 2012). AS in 5′UTRs may affect the presence, size, and positions of uORFs, which may then trigger NMD of endogenous splice variants (Kalyna et al., 2012). Generally, stop codons of uORFs might be recognized as PTCs (additionally, a long 3′UTR is created and the downstream splice junctions exist in the case of intron-containing genes) and, thereby, targeting the transcript to the NMD pathway (Figure 5B). A novel feature of uORFs triggering NMD in plants has been identified recently where AS events in 5′UTRs can result in uORFs overlapping the authentic start codon, and such transcripts are NMD sensitive (Kalyna et al., 2012). By contrast, AS transcripts of the same genes with other uORFs are NMD resistant. Since ~20% of *Arabidopsis* genes contain uORFs (Nyikó et al., 2009) and AS frequently occurs in 5′UTRs, it is likely that AS-coupled NMD involving uORFs may emerge as an important mechanism for gene regulation.

A common presumption about NMD is to label an AS variant as an NMD candidate when AS introduces a PTC at a position ≥50 to 55 nucleotides upstream of a splice junction. Accordingly, based on EST and RNA-seq analyses of wild-type plants, 36 to 78% of *Arabidopsis* AS transcripts have been estimated to be potential candidates for NMD (Wang and Brendel, 2006; Filichkin et al., 2010). However, tiling and expression microarrays found that only ~1% of plant protein-coding genes were upregulated in NMD-deficient plants (Yoine et al., 2006; Kurihara et al., 2009). These arrays have a limited ability to distinguish AS transcripts, and the usual cutoff (>1.5-fold to twofold) puts a further limitation on the detection of AS-NMD transcripts. An attempt to overcome the limitations of the above methods using a high-resolution AS RT-PCR panel with 950 transcripts of 270 alternatively spliced genes showed that in the NMD-impaired backgrounds 11 to 18% of all transcripts and 16 to 25% of the AS transcripts analyzed were NMD sensitive (Kalyna et al., 2012). Extrapolating from this analysis around 13 to 18% of intron-containing genes in the *Arabidopsis* genome are potentially regulated by AS-NMD. These values were confirmed by a recent comprehensive genome-wide study on AS in NMD mutants in *Arabidopsis* (Drechsel et al., 2013). Taken together, these estimates correlate well with those reported for flies, worms, and human (Rehwinkel et al., 2005; Pan et al., 2006; Ramani et al., 2009; Weischenfeldt et al., 2012). Importantly, Kalyna et al. (2012) identified AS transcripts that contained NMD features but were immune to NMD. Unexpectedly, the majority of IR transcripts were not sensitive to NMD despite containing PTCs, downstream splice junctions, and long 3′UTRs. By contrast, transcripts from the same gene with other types of AS events in the same or nearby intron were NMD sensitive. In line with this, another study has shown that cold-regulated *Arabidopsis* IR transcripts containing PTC and predicted to be the targets of NMD are not affected in UPF3 mutant background (Leviatan et al., 2013). Finding that PTC+ IR transcripts escape the NMD machinery may explain the high frequency of IR events identified in plants.

Interestingly, splicing factors use AS-NMD to autoregulate their expression through a negative feedback loop. Metazoan SR proteins have been shown to activate the inclusion of a PTC+ exon using ultraconserved sequences in their own pre-mRNA, thus resulting in NMD (Morrison et al., 1997; Lareau et al., 2007; Ni et al., 2007; Saltzman et al., 2008). Conservation of AS events to produce PTC-containing transcripts has also been demonstrated for SR protein genes in lower and higher plants (Kalyna et al., 2006), and plant SR proteins are also subjected to AS-NMD regulation (Palusa and Reddy, 2010; Kalyna et al., 2012). An evaluation of all the splice products of 13 alternatively spliced *Arabidopsis* SR genes in a *UPF3* mutant revealed that about half of the 53 variants with a PTC are degraded by NMD (Palusa and Reddy, 2010). Interestingly, At-SR34a, an *Arabidopsis* homolog of human SFRS1/ASF/SF2, is regulated by AS-NMD in the 5′UTR, involving a uORF, which overlaps the AUG of the main ORF (Kalyna et al., 2012). Other plant splicing regulators and RBPs have been demonstrated to auto- and cross-regulate their mRNA levels using AS-NMD. An autoregulation feedback loop of *Arabidopsis* RBPs GRP8 and GRP7, components of a slave oscillator coupled to the circadian clock, is tied to NMD of alternatively spliced transcripts (Staiger et al., 2003; Schöning et al., 2008). AS in two splicing regulators, polypyrimidine tract binding protein (PTB)–like homologs, in *Arabidopsis* is also autoregulated; an increase in protein levels of the PTB-like proteins results in an increase of a PTC-containing alternatively spliced transcript (Stauffer et al., 2010; Wachter et al., 2012). A further level of AS-NMD complexity in *Arabidopsis* is evidenced by extensive AS coupled NMD in AFC2, encoding a highly conserved LAMMER kinase, which phosphorylates splicing factors, thus establishing a complex regulatory loop (Marquez et al., 2012). It appears that AS-NMD regulation occurs at multiple levels of RNA biology as multiple RBPs, splicing factors, RNA helicases, spliceosome, and exon junction complex proteins are subjected to this regulation (Kalyna et al., 2012).

Alternative polyadenylation, which results in differing 3′UTRs, is prevalent in plants and is thought to play an important role in regulating gene expression at multiple levels, especially mRNA stability through NMD (Simpson, 2004; Hunt, 2011; Wu et al., 2011). A genome-wide study indicates a link between AS with selection of polyadenylation sites, as over 4000 genes in *Arabidopsis* are predicted to be polyadenylated within introns (Wu et al., 2011). The extent of coupling of AS to alternative polyadenylation globally and its impact on gene regulation needs further investigations.

Coupling AS to mRNA Stability and Translation through MiRNA Regulation

miRNAs, a class of *trans*-acting short (20 to 24 nucleotides long) RNAs, are important regulators of gene expression at the post-transcriptional level. They bind to complementary sequences in mRNA targets and attenuate gene expression either by cleaving the mRNA or repressing its translation (Aukerman and Sakai, 2003; Bartel, 2009; Voinnet, 2009). In plants, transcripts of many transcription factors and other regulatory proteins that perform critical roles in plant growth, development, and stress responses are targets of miRNAs (Rogers and Chen, 2013b). Generation of

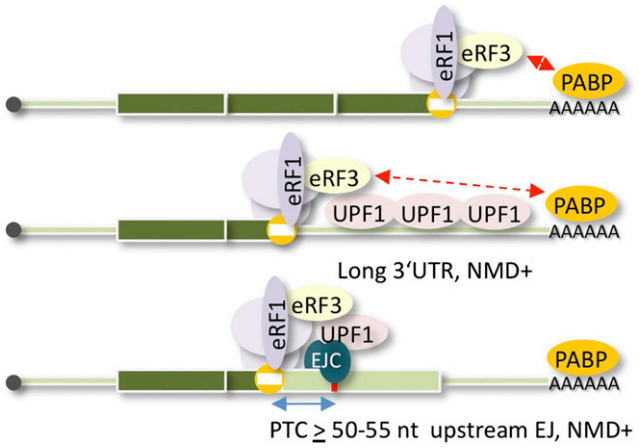


Figure 4. Current NMD Models.

Both authentic and premature stop codons are recognized by translation termination factors eRF1 and eRF3. In the course of canonical termination, interaction of eRF3 with poly(A) binding protein (PABP) leads to the efficient release of a peptide **(top)**. According to the long 3′UTR model (middle), this interaction does not occur when the distance between a stop codon (PTC) and poly(A) is extended. eRF3 then interacts with UPF1, which competes with poly(A) binding protein for binding sites. This perturbs translation termination and results in NMD. In the exon junction complex (EJC) model **(bottom)**, EJC is deposited 20 to 24 nucleotides (nt) 5′ of an exon-exon junction during pre-mRNA splicing (Singh et al., 2012). The EJCs are stripped off from mRNA during the first round of translation. In PTC-containing transcripts, termination of translation more than 50 to 55 nucleotides upstream of an EJC triggers RNA decay through NMD by recruiting UPF3 and its cofactors UPF1 and UPF2 (Schoenberg and Maquat, 2012).

short mature miRNAs from primary transcripts of miRNAs (pri-miRNAs), which are much longer (up to 3000 nucleotides), involves two sequential site-specific endonucleolytic events. The first event crops pri-miRNAs to produce pre-miRNAs with shorter stem-loop structure, and the second event results in release of an miRNA:miRNA* duplex (Rogers and Chen, 2013b, 2013a)

Recent studies suggest that AS can modulate miRNA-mediated regulation of gene expression in multiple ways. These include the generation of mRNA or noncoding RNA isoforms that either contain or lack target sites for miRNA (Figure 6A) and regulation of splicing of pri-miRNAs thereby modulating the levels of miRNAs, which in turn regulates the stability or translation of target mRNAs. Modulation of miRNA levels is also accomplished by regulating the splicing pattern of pri-miRNAs or pre-mRNAs encoding enzymes, such as dicer-like proteins, involved in miRNA biogenesis (Figure 6B). In plants, emerging studies are providing evidence in support of all of these mechanisms.

The presence of an miRNA target in some splice isoforms but not in others from a given gene links regulated AS to miRNA-mediated regulation of gene expression. In plants, miRNA binding sites (MBSs) in mRNAs are found in both the coding as well as UTRs. In *Arabidopsis*, using bioinformatic analysis, it was found that over 12% of identified MBSs (44 out of 354 mRNAs with MBSs) are affected by AS (Yang et al., 2012b), suggesting coupling of AS to miRNA-mediated gene regulation. Removal of introns in pre-mRNAs of some transacting small interfering RNA (*TAS*) genes (*TAS1A* and *TAS2*) results in a loss of MBSs and the levels of RNA isoforms of *TAS1A* and *TAS2* that do not contain MBS are predominant, suggesting a role of AS of *TAS* pre-mRNAs in regulating their transcript isoforms (Yang et al., 2012b). Members of *Squamosa-promoter binding protein* like (*SPL3*, *4*, and *5*) are key regulators of transition from vegetative to reproductive phase and contain an MBS for miR156 (Wu and Poethig, 2006). Expression of these SPLs is strongly inversely correlated with expression of miR156. *SPL4* generates three mRNA isoforms, but only one isoform (*SPL4-1*) contains an MBS for mIR156. Overexpression studies with miR156 have shown that the level of *SPL4-1* transcript but not other isoforms is regulated by this miRNA, suggesting a strong coupling of AS to miRNA-mediated gene regulation. Recently, it was found that introns in rice and *Arabidopsis* are targeted by miRNAs, suggesting that pre-mRNAs or spliced intron-containing transcripts could be cleaved by miRNAs (Meng et al., 2013).

Some studies show direct coupling of AS to miRNA biogenesis at the splicing of pre-miRNA level (Hirsch et al., 2006; Yan et al., 2012). A number of miRNA genes in plants reside within an intron of protein coding genes and are coexpressed and processed with the host protein-coding gene (Brown et al., 2008; Yan et al., 2012; Yang et al., 2012a). AS of primary transcripts due to IR and exon skipping of several plant miRNAs has been reported (Aukerman and Sakai, 2003; Kurihara and Watanabe, 2004; Hirsch et al., 2006; Szarzynska et al., 2009). In some cases, the secondary structure needed for miRNA biogenesis is affected (Hirsch et al., 2006) and hence may affect miRNA biogenesis. *Arabidopsis MIR400* resides within the intron of a protein-coding gene (At1g32853) and is cotranscribed with its host gene. Interestingly, heat regulates AS of this intron, resulting in an increased level of miR400 primary transcripts but a decreased level of mature miR400 (Yan et al., 2012). This coupling of regulated AS to biogenesis of miR400 appears to be important for heat stress tolerance as overexpression of miR400 resulted in decreased germination and seedling growth under heat stress (Yan et al., 2012). Interestingly, two recent studies have shown that introns in plant miRNA genes (*MIR161*, *MIR163*, and *MIR172a*) enhance miRNA production in a position-dependent manner (Bielewicz et al., 2013; Schwab et al., 2013). Removal of introns from these genes reduced the level of miRNAs. In addition to regulated splicing of pri-miRNAs and splicing-independent intron enhancement of miRNA production described above, AS is also coupled to miRNA biogenesis indirectly by regulating the splicing of pre-mRNAs that encode miRNA processing enzymes. *Dicer-like1* (*DCL1*), a key endonuclease in biogenesis of miRNA from pri-miRNA (Rogers and Chen, 2013b), has 20 exons and contains an MBS for miR162 that is interrupted by intron 12. AS in this intron generates four isoforms (*DCL1-1* to *DCL1-4*), and one of these four isoform contains an MBS (Yang et al., 2012a). Hence, regulated AS of DCL1 pre-mRNAs would affect regulation of DCL1 protein levels.

How Much of the Observed AS Is Biologically Meaningful?

How much of the observed AS is due to regulated splicing with biological consequences, and how much is due to splicing

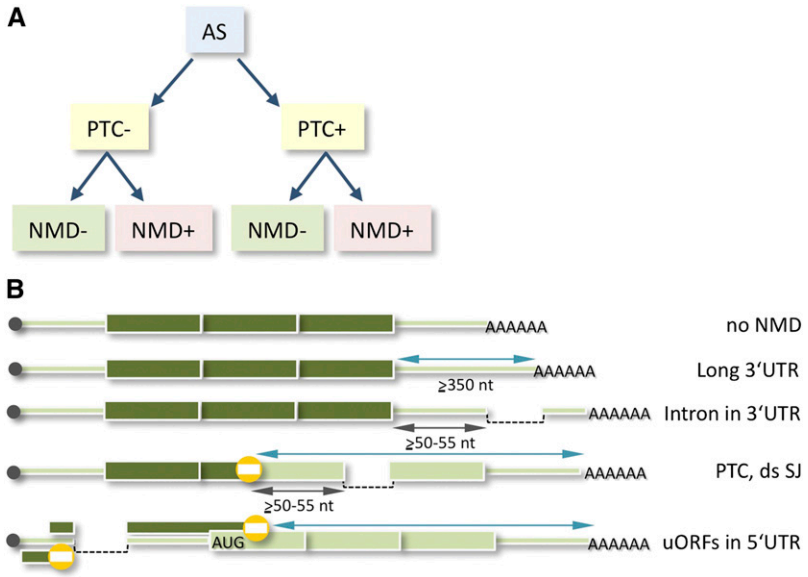


Figure 5. Coupling of AS to NMD.

(A) AS can result in both PTC− and PTC+ transcripts. Both PTC+ and PTC− transcripts can be subjected to NMD (NMD+). Not every PTC+ transcript is targeted by NMD (NMD−).

(B) Features of NMD sensitive transcripts: 3′UTRs longer than 350 nucleotides; introns in the 3′UTR with a distance >50 to 55 nucleotides (nt) from the authentic stop codon to the downstream splice junction (ds SJ); PTC more than 50 to 55 nucleotides upstream of a splice junction; uORFs in 5′UTR, including uORFs overlapping with authentic start codon (AUG) of the main ORF. PTCs are shown by yellow stop signs. Blue arrows above transcripts indicate long 3′UTRs that can serve as NMD signals.

errors is not known. Some studies argue that much of the AS represents splicing noise and is not physiologically relevant (Melamud and Moult, 2009; Severing et al., 2009; English et al., 2010; Pickrell et al., 2010). English et al. (2010) analyzed the AS events for genes that have ESTs and found that for most AS events, one major splicing choice was supported by a majority of the ESTs, with <10% of ESTs supporting the minor form. However, there was a subset of genes for which mRNA isoforms appeared about equally, but most (>80%) of these equal forms were affected in UTRs or the changes in the protein were small. An animal study found that a majority of all alternative isoforms resulted in unstable protein conformations, but only a subset of isoforms maintained protein structural integrity (Melamud and Moult, 2009). These studies suggest that AS represents splicing noise as most of it is not conserved, the level of minor isoforms are low, and many mRNA isoforms are predicted to produce truncated proteins (Melamud and Moult, 2009; Severing et al., 2009; English et al., 2010). The lack of conservation of an AS event across phylogenetically divergent species is not inconsistent with a biological role for AS, as global studies from a number of vertebrates suggest that lineage-specific AS is important for divergence of their organs' phenotype and functions (Barbosa-Morais et al., 2012; Merkin et al., 2012). Conservation of AS patterns in orthologous genes is not high in animals (Barbosa-Morais et al., 2012). Between human and mouse, only 10% of AS events are conserved in orthologous genes (Pan et al., 2005). However, a recent study in unicellular fungus *Schizosaccharomyces pombe* resulted in a striking finding of identical exon skipping events in plants, animals, and fungi, species that diverged more than one billion years ago (Awan et al., 2013). Some recent studies suggest that rapid divergence in AS may have contributed to organogenesis and speciation (Barbosa-Morais et al., 2012; Merkin et al., 2012). Unlike conserved AS events, mRNA isoforms generated by lineage-specific AS tend not to code for functional proteins due to disruption of reading frame (Merkin et al., 2012). Hence, these AS events are thought to be important for regulating gene expression. Furthermore, studies suggest that truncated proteins perform functions and a low level of some mRNA isoforms is likely due to AS-coupled regulation of levels of those isoforms. Although it is perhaps likely that most regulated AS has a biological role, future studies on functional significance of splice variants are needed to address this.

Evolution of AS Complexity

Soon after the discovery that eukaryotic genes are split, it was speculated that split gene organization could accelerate evolution (Gilbert, 1978). Recent AS studies in phylogenetically divergent organisms are lending support to this hypothesis. Analysis of AS in seven organs from 10 vertebrates ranging from frog to human and spanning ~350 million years of evolution revealed the highest AS complexity in human, and decreasing complexity as the evolutionary distance increased from primates (Barbosa-Morais et al., 2012). Based on this study, it was proposed that divergence in AS may have contributed to speciation. In another study, comparing gene expression in several tissues from mammals and

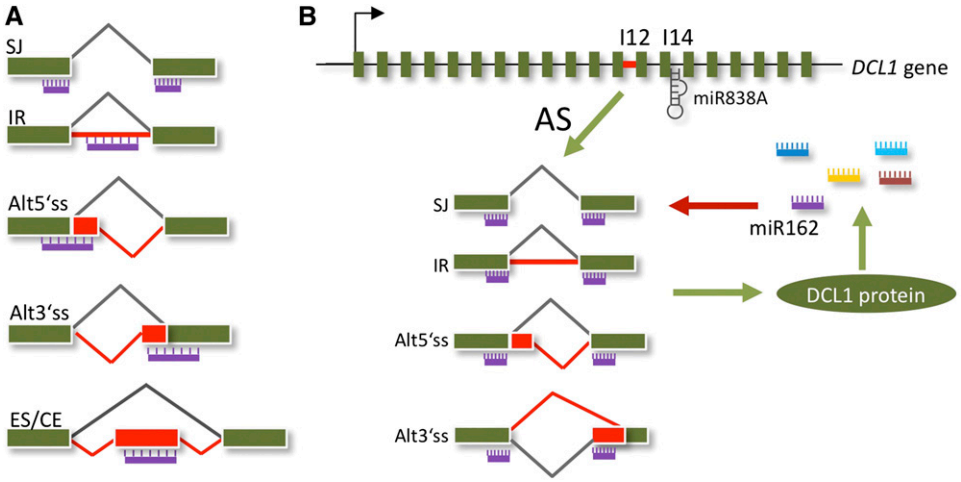


Figure 6. AS Can Modulate miRNA-Mediated Regulation of Gene Expression.

(A) Generation of RNA splice variants that either contain or lack target sites for miRNA. SJ, splice junction; ES/CE, exon skipping/cassette exon.
(B) Modulation of miRNA levels by regulating the splicing pattern of pre-mRNAs encoding enzymes, such as DCL1, involved in miRNA biogenesis. MBS for miR162 (shown in violet) is split between E12 and E13 of *DCL1* pre-mRNA. Only one of the four isoforms generated by AS contains a miRNA target site. In addition, miR838A is encoded in the intron 14 of *DCL1*. AS events and regions are marked red. *DCL1* gene structure is not to scale.

a bird it was found that tissue-specific gene expression is mostly conserved, whereas AS is highly diverged with many lineage-specific differences (Merkin et al., 2012). Phenotypic divergence in organs in different vertebrates is attributed to lineage-specific AS (Barbosa-Morais et al., 2012). Such in-depth comprehensive AS studies in land plants that diverged about ~450 million years ago should provide information on conservation of AS and evolution of AS over time.

AS REGULATION: DECODING THE SPLICING CODE IN PLANTS

Cis-Regulatory Elements

The basis for CS and regulated AS is differential usage of splice sites. Elucidating the mechanisms of CS and AS is of central importance to understanding how various signals influence AS (Chasin, 2007; Chen and Manley, 2009; Roca et al., 2013). Decoding of the splicing code requires comprehensive understanding of rules that dictate splice site choices to combine separate parts of a transcript and generate different genetic messages. The canonical splice site consensus sequences in introns are very short and degenerate; hence, they do not provide enough information for distinguishing correct pairs of splice sites from alternative splice sites. Research using various strategies has identified *cis*-acting sequence elements in exons and introns that are important for the correct recognition of pairs of splice sites (Blencowe, 2006; Chasin, 2007; Chen and Manley, 2009). These SREs are referred to as exonic splicing enhancers (ESEs)/silencers and intronic splicing enhancers/silencers (Chen and Manley, 2009; Huelga et al., 2012). The mechanisms that control the differential usage of splice sites have not been elucidated in plants. Studies on AS regulation in animals have uncovered that features in RNA sequences (predominantly *cis*-elements and in some cases RNA structures) and the RBPs (*trans*-factors that bind *cis*-elements), including SR proteins and hnRNPs, are important regulators of splice site choice and AS. This regulatory code of splicing is complex and involves numerous loosely conserved *cis*-elements, a plethora of RBPs, and an elaborate network of interactions among them.

SREs

SREs are short (5 to 10 nucleotides) conserved sequences that often occur in clusters (Chen and Manley, 2009; McManus and Graveley, 2011). Interactions between these *cis*-elements in pre-mRNA and RBPs either promote or suppress the use of a splice site. Studies with individual genes and several transcriptome-wide maps for *cis*-elements for different RNA binding splicing regulators in animals have identified numerous *cis*-elements and their effect on AS (Licatalosi et al., 2008; Chen and Manley, 2009; Barash et al., 2010; Hafner et al., 2010a). In fact, extensive analysis of such elements involved in exon skipping in animals has led to deciphering of the splicing code that is now used to predict exon skipping in animals (Barash et al., 2010). The fact that sequence motifs can be used to predict splicing outcomes in some cases suggest that they are the predominant *cis*-elements in regulated splicing (Barash et al., 2010; Zhang et al., 2010a). Splicing regulatory proteins bind to SREs and regulate CS and AS by recruiting/stabilizing spliceosomal components and/or antagonizing other splicing inhibitory proteins. Protein–protein interactions between SR proteins and proteins involved in both 5′SS and 3′SS aid in finding correct splice sites. Although numerous binding proteins have been identified in plants based on the presence of different types of RBPs (Lorković, 2009), the sequences that bind to these proteins are not known in most cases (see *trans*-factors below).

In animals, selective evolution of ligands through exponential enrichment (SELEX) (Coulter et al., 1997) and its variations (functional SELEX and genomic SELEX), RNA immunoprecipitation (RIP) (Barkan, 2009; Zhao et al., 2010), cross-linking and immunoprecipitation (CLIP) (Wang et al., 2009b) and its variations, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP) (Licatalosi and Darnell, 2010) and photoactivatable-ribonucleoside-enhanced CLIP (Hafner et al., 2010a) have been used to identify targets of RBPs (Figure 2, Table 1). The binding regions of many tissue-specific RNA binding splicing regulators in animals have been identified using these tools. The wealth of experimentally obtained information on sequences in RNA that bind to splicing regulators and the vast repertoire of mRNA isoforms with various AS events has paved the way to develop computational tools to predict *cis*-elements and AS events, especially exon skipping, in animals (Barash et al., 2010; Zhang et al., 2013a). In animals, at least 70 bases in introns flanking some of the alternative exons are conserved across organisms, suggesting intronic splice regulatory elements. Furthermore, analysis of intronic sequences downstream of broadly conserved alternative exons for 5mers identified sequences that match perfectly with consensus sequences of many tissue-specific splicing regulators (Merkin et al., 2012), suggesting the importance of intronic splicing regulatory elements in AS of broadly conserved exons.

Highly conserved intronic sequences flanking alternative splice sites have been found in plant-specific SR genes in mosses, monocots, and dicots, taxa separated by more than 400 million years of evolution implicating an ancient regulatory function (Kalyna et al., 2006). In plants, very little is known about the SREs that regulate AS. Genomic SELEX has been used to identify the sequences that bind to AtCyp59, which is known to interact with components of transcription and splicing (Bannikova et al., 2013). Using computational tools, Pertea et al. (2007) have identified over 80 ESEs in *Arabidopsis*, some of which are similar to human ESEs. Several of these were experimentally validated. In addition to these, only few other SREs have been identified experimentally in plants (Yoshimura et al., 2002; Lopato et al., 2006; Schöning et al., 2007, 2008; Reddy et al., 2012b; Thomas et al., 2012). The lack of information on experimentally identified SREs in plants is a major limitation in developing computational tools to predict RNA binding elements, AS, and the plant splicing code. However, accumulating data on various types of AS events and empirical identification of *cis*-elements in plants, as in animals, should pave the way for the development of computational tools to predict conserved patterns (motifs) associated with an AS event.

Regulation of AS by Secondary Structures

Regulatory information for splicing in RNAs can also be present in its secondary and tertiary structures that involve base pairing and structural conformations. Structural elements within the pre-mRNA can suppress splicing by masking splice cites or promote splicing either by unmasking splice sites or by bringing sequences that bind to splicing regulatory proteins to close proximity (Donahue et al., 2006; Warf and Berglund, 2010; Jin et al., 2011). In the Down Syndrome Cell Adhesion Molecule gene in *Drosophila melanogaster*, an extreme example of AS with 95 alternatively spliced exons, competing RNA structures were found to be required for mutual exclusion exons (May et al., 2011). In nonplant systems, there are well-documented examples demonstrating the roles RNA structures play in regulated splicing (Donahue et al., 2006; Warf and Berglund, 2010; Jin et al., 2011; Wan et al., 2011). In plants, it has been shown that insertion of an artificial secondary structure (stem-loop) in an intron strongly inhibits splicing in vivo in tobacco (*Nicotiana tabacum*) protoplasts (Goodall and Filipowicz, 1991). Furthermore, the inhibitory effect was observed only when the stem-loop was placed at the 5′SS or in the middle of an intron but not at the 3′SS (Liu et al., 1995), suggesting a regulatory role for secondary structures in pre-mRNA splicing in plants. So far, there are at least two well-characterized secondary structures (a riboswitch and a 5S rRNA mimic) in plants that are known to regulate splicing in vivo.

Secondary and tertiary structures in RNAs serve as a binding site for a metabolite (in the case of riboswitches) or an RBP (5S RNA mimic) and regulate AS. Riboswitches, which are widely used in prokaryotes, are metabolite binding structures in mRNA that function as genetic switches to turn on or off gene expression at the transcription or translation level in response to metabolites (Breaker, 2011; Serganov and Nudler, 2013). In bacteria, riboswitches regulate the biosynthesis of essential metabolites, including vitamins and amino acids. Riboswitches that regulate pre-mRNA splicing have recently been identified in land plants, algae, and fungi where a small molecule regulates splicing by modulating RNA structure. In these organisms, AS of pre-mRNAs encoding proteins involved in thiamine metabolism is modulated by a metabolite (thiamin-pyrophosphate [TPP]) binding riboswitch (Bocobza and Aharoni, 2008; Wachter, 2010). Three TPP switches were identified in the fungus *Neurospora crassa* (Cheah et al., 2007). Interestingly, it was found that one activates and two repress gene expression by controlling pre-mRNA splicing. The location of a TPP binding aptamer within the pre-mRNA can be in the 5′UTR, 3′UTR, or coding region (Wachter, 2010). In *Arabidopsis* and other plants, conserved TPP riboswitches in the 3′UTR of the *THIAMINE C SYNTHASE* (*THIC*) gene regulate thiamine metabolism (Bocobza et al., 2007; Wachter et al., 2007). The binding of TPP to an aptamer changes the splicing pattern of pre-mRNA by masking a splice site resulting in a splice variant with a retained intron, which is less stable. Therefore, an increase in TPP would lead to its binding to the aptamer, resulting in production of the less stable intron-retained variant of *THIC* and in a decrease in TTP production. This negative feedback regulation of thiamine biosynthesis at the AS level ensures a high level of transcripts under low cellular TPP concentrations. Plants expressing *THIC* with a disrupted TPP riboswitch in a *thic* mutant showed growth retardation, chlorosis, and delayed flowering, indicating the physiological significance of riboswitch-mediated regulation of *THIC* (Bocobza et al., 2013). It is not known to what extent riboswitches regulate AS of other pre-mRNAs in plants. There are speculations that riboswitch-mediated AS regulation may exist for other genes (Grojean and Downes, 2010). Synthesis of 5S rRNA (5S rRNA) by RNA polymerase III requires transcription factor IIIA (TFIIIA). In the case of plant TFIIIA pre-mRNA, a structural element that mimics 5S rRNA in its third exon regulates AS (Fu et al., 2009; Hammond et al., 2009). This exon is highly conserved in land plants. Skipping of this exon produces a functional

transcript, whereas its inclusion results in a nonfunctional PTC-containing transcript that is targeted for degradation; hence, it is referred to as a suicide exon. Interestingly, skipping of this exon is induced by ribosomal protein L5 that binds to the 5S rRNA mimic. These studies indicate that ribosomal protein L5 controls the synthesis of 5S rRNA by regulating AS of TFIIIA pre-mRNA via an exonic mimic of 5S rRNA binding site. It is likely that more of these will be discovered as global RNA secondary structure analyses are performed and those structures are integrated with AS data.

Recently, several novel in vitro high-throughput sequencing approaches to obtain genome-wide maps of mRNA secondary structures, referred to as the RNA structurome or RNA foldome have been developed (Wan et al., 2013). These include selective 2'-hydroxyl acylation analyzed by primer extension sequencing, fragmentation sequencing, parallel analysis of RNA structures (PARS) (Kertesz et al., 2010), double-stranded RNA-seq, and single-stranded RNA-seq (Table 1, Figure 2). These tools have been applied to yeast and animals systems (Kertesz et al., 2010; Warf and Berglund, 2010; Wan et al., 2011, 2013), while only two such global RNA structure analyses have been reported recently in plants (Zheng et al., 2010; Li et al., 2012). A recent in vivo genome-wide RNA structure study revealed secondary structure patterns at sites of AS (Ding et al., 2013). All of these are based on high-throughput sequencing of cDNAs prepared from RNAs that are either modified with specific chemicals or digested with specific types of nucleases. The data can be integrated into programs like SeqFold, which predicts structures, to obtain accurate RNA structure (Wan et al., 2013). Correlating the structural information obtained with these methods with AS in plants will likely uncover additional RNA structural elements involved in regulated splicing in plants (Figure 2). Understanding the roles of regulatory *cis*-elements in mRNA translation and identification of these in splice variants may also provide clues to roles of AS in translational regulation. Coupling of RIP methods with PARS should provide insights into the roles of RNA secondary structures in regulated splicing.

Modulation of splicing by secondary (e.g., hairpins and stem loops) and tertiary structural features in RNA that are responsive to ligands or altered cellular environments (e.g., osmolytes, ion concentration, and temperature) is appealing as it is rapid and may or may not require RBPs. RNA structures are highly sensitive to concentration and types of ions and osmolytes (Draper, 2004; Lambert and Draper, 2007; Mullen et al., 2012). Osmolytes have both stabilizing and destabilizing effects on RNA tertiary structure (Lambert and Draper, 2007). It is known that many abiotic stresses, such as high salinity, heat, and drought, impact cellular concentrations of ions and osmolytes in plants (Seki et al., 2007; Mullen et al., 2012; Chan et al., 2013), which might impact RNA secondary structure and regulate splicing and other aspects of posttranscriptional gene regulation. Hence, structural elements in pre-mRNA might function as sensors of abiotic stresses, such as temperature, drought, and salinity. Castiglioni et al. (2008) have shown that bacterial RNA chaperones (CspA or CspB) in maize (*Zea mays*) promote stress tolerance, indicating the importance of RNA secondary structures during stress responses. Recruitment of SR proteins by RNA secondary structure has also been reported (Buratti and Baralle, 2004; Buratti et al., 2004; Shepard and Hertel, 2008; McManus and Graveley, 2011). In prokaryotes, structures in mRNAs often referred to as RNA thermometers function as temperatures sensors, which occlude the ribosome binding site at normal temperature. High temperature unfolds these structures to enable translation (Kortmann and Narberhaus, 2012; Wan et al., 2012). In theory, temperature-sensitive structures in pre-mRNAs in eukaryotes may also mask or unmask binding sites for splicing factors and modulate splicing rapidly in response to temperature changes. Forty percent of introns in budding yeast (*S. cerevisiae*) contain a predicted secondary structure between the branch site and 3′SS (Meyer et al., 2011). Furthermore, a secondary structure at the 3′SS in an intron of *APE2* pre-mRNA was shown to function as an RNA thermosensor by detecting changes in temperature and modulating AS (Meyer et al., 2011). Since pre-mRNA splicing occurs cotranscriptionally, the rate of transcription can also have a significant effect on RNA structure-mediated splicing as slow folding structures may form only if the rate of transcription is low.

Trans-Acting Factors

There are over 800 proteins with different types of RNA binding domains in *Arabidopsis* (Silverman et al., 2013), and a vast majority of these contain one or more, or a combination of, RNA binding domains, such as an RNA recognition motif (RRM), a K homology domain, or a pentatricopeptide repeat domain that bind RNA. Many plant RRM-type RBPs have no orthologs in animals (Lorković, 2009). Of the non-snRNP proteins, there are many SR proteins or hnRNPs that regulate splice site selection by binding to RNA motifs (see Supplemental Table 1 online). The SR family of splicing regulators contains one or two RRMs in the N terminus and a Arg/Ser-rich (RS) domain at the C terminus and function in CS and AS (Barta et al., 2010). Pre-mRNAs of most SR proteins and several hnRNPs are alternatively spliced (Lopato et al., 1999b; Lazar and Goodman, 2000; Kalyna et al., 2003; Kalyna and Barta, 2004; Isshiki et al., 2006; Palusa et al., 2007; Schöning et al., 2007; Wachter et al., 2012). In plants, SR proteins have been investigated for their function in splicing (Lopato et al., 1999a, 1999b). In animal systems, SR proteins, apart from their role in pre-mRNA splicing, function in diverse processes associated with RNA metabolism and gene regulation, including mRNA export, stability and translation, chromatin binding, transcription elongation, subcellular localization of transcripts, genome stability and formation of cellular RNA granules, and miRNA processing structures (stress granules and processing bodies) (Long and Caceres, 2009; Shepard and Hertel, 2009; Risso et al., 2012; Yoon et al., 2013). Whether plant SR proteins, like their animal counterparts, have wide-ranging roles in other aspects of RNA metabolism remains to be seen.

The SR proteins interact with RNA through their RRM, which provides RNA binding specificity, and with other spliceosomal proteins through their RS domain, which aids in spliceosome assembly. In some instances, the RS domain interacts directly with pre-mRNA either at the BP or 5′SS (Shen et al., 2004). Some plant SR and SR-like proteins have been shown to bind alternatively spliced introns (Day et al., 2012; Thomas et al., 2012). Complex interaction networks among SR proteins and interaction with many other snRNP and non-snRNP proteins (Golovkin and Reddy, 1998; Golovkin and Reddy, 1999; Lopato et al., 2002;

Gullerova et al., 2006; Reddy, 2007; Barta et al., 2008) suggest an important role for this family of proteins in regulated splicing. Expression levels and AS patterns of SR proteins in plants vary in different tissues, implying their target specificity (Lopato et al., 1996, 1999a, 1999b, 2002; Golovkin and Reddy, 1998, 1999; Lazar and Goodman, 2000; Kalyna et al., 2003; Kalyna and Barta, 2004). The fact that some SR proteins have been shown to interact with components of the U12-dependent minor spliceosome suggests a role in splicing catalyzed by the minor spliceosome (Reddy, 2007; Barta et al., 2008).

The hnRNPs are a heterogeneous group of proteins involved in diverse functions in mRNA metabolism, mainly characterized by their ability to bind RNA due to the presence of RRM domains or other domains like K homology domains and additional Gly-rich domains that add versatility to binding to RNA and/or that can extend their functions (Martinez-Contreras et al., 2007; Han et al., 2010; Wachter et al., 2012). Many of the hnRNPs are located in the nucleus and nuclear compartments, but some of them exhibit nucleus-cytoplasm shuttling that is regulated by posttranslational modifications, indicating a broader set of functions of these proteins besides nuclear mRNA processing (Wachter et al., 2012). The most extensively studied hnRNPs are the PTB proteins where their role in AS was first described in mammals (Mulligan et al., 1992; Wagner and Garcia-Blanco, 2001; Izquierdo et al., 2005). They act through their binding to pyrimidine-rich sequences in the pre-mRNA by competition and/or interaction with other hnRNPs and splicing factors exerting a positive or negative effect in the selection of specific splice sites that define the AS outcome in the transcript. In *Arabidopsis*, three genes have been identified (*PTB1*, *PTB2*, and *PTB3*; see Supplemental Table 1 online), all of them generating two alternatively spliced variants where in all the cases one of the two transcripts is targeted by NMD (Stauffer et al., 2010). The AS variants in PTB1 and PTB2 are derived from a mode of negative auto- and cross-regulation by these PTB proteins, a regulatory mechanism that is consistent with the human PTB orthologs (Grabowski, 2007; Stauffer et al., 2010). The analogies found in the mode of action of the PTBs in plants and human suggested that the plant proteins might also influence the AS of a considerable number of genes. A recent genome-wide study by Rühl et al. (2012) using overexpression and artificial miRNA lines of the different PTB proteins showed that mainly PTB1 and PTB2, but not PTB3, influence the AS outcome of at least 300 genes that comprise important factors involved in diverse developmental processes, including flowering and seed germination. Interestingly, in contrast with what has been found in mammals where principally exon skipping events are altered, diverse types of AS were affected by PTB1 and PTB2, including 5′ alternative splice site, IR, and exon inclusion events. Nevertheless, the mechanisms of action and the evidence for binding of PTB1 and PTB2 to generate the observed AS patterns remain to be elucidated.

Small Gly-rich hnRNPs-like GRP7 and GRP8 in *Arabidopsis* have also been shown to have a role in AS. These two proteins are controlled by the circadian clock, and their protein levels are regulated similarly to PTBs by AS and the production of NMD-sensitive transcripts (Carpenter et al., 1994; Heintzen et al., 1994; Staiger, 2001; Staiger et al., 2003). Using a high-resolution RT-PCR panel, it was shown that GRP7 affects 59 events from a total of 288 AS events tested (Streitner et al., 2012). RIP experiments confirmed binding of GRP7 to seven targets that was dependent on the RRM domain. Interestingly, some of the splicing events affected by GRP7 are also influenced by other RBPs, like SR proteins and the cap binding complex proteins (CBC), suggesting that the splicing profiles observed in *Arabidopsis* are the result of a complex interplay between RBPs and/or splicing factors (Streitner et al., 2012).

Despite the evident expansion in plants of plant-specific members of hnRNPs and hnRNP-like proteins, little is known about their actual role in AS. Few studies have identified functional features of plant specific hnRNPs-like proteins. The oligouridylate binding protein (UBP1) in *Nicotiana plumbaginifolia* was found to have binding affinity to uridine-rich sequences and increase the efficiency of poorly spliced introns (Lambermon et al., 2000). Nevertheless, the UBP1-related proteins RBP45 and RBP47 did not affect splicing efficiency (Lorković et al., 2000). Similar results were found for *Arabidopsis* UBA1a and UBA2a where no stimulation of splicing was detected, suggesting a role of UBP1 and their associated proteins in mRNA stabilization (Lambermon et al., 2002). The prevalence of AS in plants (Syed et al., 2012) implies that the presence of multiple factors involved in splicing play important roles in the production of alternative spliced transcripts and that crosstalk and tight regulation between these factors exist to produce, in a very orchestrated manner, transcripts relevant in important biological aspects like development and response to the environmental cues.

In addition to SR and hnRNP proteins, the involvement of nuclear CBC in AS regulation has been shown by analyzing the AS events in single/double *Arabidopsis* mutants of the CBC components, *cbp20* and *cbp80*/*abh1*. Changes in AS isoforms were shown for 101 genes, with 41% of them common to all mutants and 15% arise only in double mutants. The authors propose a direct involvement of CBC in plant AS of the first intron, particularly at the 5′SS (Raczynska et al., 2010).

Subcellular Localization and Dynamics of *Trans*-Acting Factors

Speckles (also called interchromatin granule clusters) are a type of nonmembranous structure enriched in snRNPs, SR proteins, and other splicing factors (Spector and Lamond, 2011). Using SR proteins tagged to fluorescent reporters, it was shown that plant SR proteins localize to the nucleoplasm and nuclear speckles (reviewed in Reddy et al., 2012a). The number, size, and shape of speckles vary depending on the cell type and in response to external signals. Using fluorescence recovery after photobleaching and fluorescence loss in photobleaching, it was shown that plant SR proteins show high mobility within the nucleus. SR proteins shuttle between the nucleoplasm and speckles, and this movement is regulated by protein phosphorylation and dephosphorylation. Localization and dynamics of plant SR proteins has been recently reviewed and hence is not discussed in detail here (Reddy et al., 2012a). Until recently, speckles were thought to be storage sites of splicing factors (Spector and Lamond, 2011). However, a recent demonstration of the presence of a protein that is part of catalytically active spliceosomes in speckles and the presence of spliceosomes near or within the speckles indicate that posttranscriptional splicing

may occur in speckles (Girard et al., 2012). Little is known about the biogenesis of nonmembranous bodies containing RNA and proteins. Recent studies show that lowcomplex-disordered regions present in many of the splicing and RNA processing proteins are necessary and sufficient for formation of these bodies. Furthermore, phosphorylation of these regions contributes to the dynamic nature of these bodies (Han et al., 2012; Kato et al., 2012).

AS REGULATION BY EPIGENETIC MARKS ON DNA AND HISTONES

Splicing occurs mostly cotranscriptionally, but removal of some introns, especially alternatively spliced introns, takes place posttranscriptionally (Khodor et al., 2011; Braunschweig et al., 2013). Evidence is accumulating that chromatin modifications, transcription, and splicing are highly integrated, each influencing the other (Braunschweig et al., 2013; Dujardin et al., 2013; Gómez Acuña et al., 2013). Multiple aspects associated with transcription, such as promoter strength, transcription factor occupancy, rate of transcription, speed of RNA polymerase, and chromatin modifications, are known to change AS (Braunschweig et al., 2013). Although the means by which chromatin modifications affect splicing remain to be elucidated, there are indications that these modifications may directly or indirectly recruit spliceosomal proteins (Luco et al., 2010; Pradeepa et al., 2012; Schor et al., 2012; Tolstorukov et al., 2012; Braunschweig et al., 2013). Phosphorylation of the catalytic subunit of polymerase II at its C-terminal domain recruits many splicing factors, such as SRs and hnRNPs, that function in AS (Hsin and Manley, 2012). Some studies show that spliceosomal proteins bound to nascent transcripts may impact chromatin architecture and polymerase II elongation (Braunschweig et al., 2013). An emerging theme is that polymerase II elongation rate affects AS through kinetic coupling (Luco et al., 2011). Fast elongation rates favor exon skipping, whereas slow elongation favors spliceosome assembly and exon inclusion.

Recent studies in animals indicate that chromatin remodeling, which in turn affects the rate of transcription as well as nucleosome positioning, can alter the AS pattern (Alló et al., 2009; Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009). Global studies in plants and animals indicate that DNA modifications, such as cytosine methylation, and chromatin changes, such as histone modifications and nucleosome occupancy, mark exon-intron boundaries (Braunschweig et al., 2013). Increased methylation is found in exons compared with introns. Further, exon–intron boundaries showed sharp transitions in methylation (Chodavarapu et al., 2010; Laurent et al., 2010). In *Arabidopsis* and human, using MNase-seq, it was found that nucleosomes are enriched in exons, especially at the boundaries (Schwartz et al., 2009; Chodavarapu et al., 2010). Furthermore, specific posttranslational modifications of histones are also enriched in exons (Schwartz et al., 2009). Differential nucleosome occupancy on exons that differ in splice site strength and a correlation between nucleosome occupancy and exon inclusion in metazoans suggest a role for chromatin architecture and regulated splicing (Tilgner et al., 2009; Braunschweig et al., 2013).

In plants, environmental signals are known to alter chromatin modifications in a gene-specific manner (Kim et al., 2010; Dowen et al., 2012; Gutzat and Mittelsten Scheid, 2012). Since stresses have a dramatic affect on pre-mRNA splicing, including those that are stress induced (Ali and Reddy, 2008b; Duque, 2011), it is likely that some of these changes in AS are regulated through stress-induced modification of chromatin. Nucleotide composition is different between exons and introns, with high GC content in exons and high AU in plant introns (Goodall and Filipowicz, 1989). Experimental evidence supports a role for differential nucleotide composition in pre-mRNA splicing. RNA polymerase II elongation rates are impacted by nucleotide sequence composition with a slow rate in adenine-thymine–rich regions.

A report in *S. pombe* implicates a role for splicing factors in RNA-induced chromatin silencing that is uncoupled from splicing (Bayne et al., 2008). In plants, recent reports indicate a role of splicing machinery and splicing factors in regulating transcription. Introduction of an intron into an intronless gene was shown to suppress RNA silencing that is dependent on efficient splicing of the intron, suggesting a role for splicing and transcriptional regulation (Christie et al., 2011). Some studies suggest a role for splicing machinery in RNA-directed methylation (RdDM). A mutant screen aimed at identifying proteins involved in transcriptional silencing through RdDM identified a splicing factor (ZINC FINGER AND OCRE DOMAIN-CONTAINING PROTEIN1) as one of the key regulators of RdDM (Zhang et al., 2013b). SR45, an SR-like splicing regulator, was isolated in a screen for mutants that affect the establishment of RNA-directed DNA methylation (Ausin et al., 2012). The loss-of-function mutant of *SR45* displayed reduced methylation, suggesting a role of this splicing factor in RdDM, but an indirect role through splicing regulation of pre-mRNAs of genes involved in RdDM is not excluded (Ausin et al., 2012).

ELUCIDATING PATHWAYS THAT CONNECT SIGNALS TO AS REGULATION

Rapid alteration of transcription complexity in response to stresses is a simple mechanism for rapid reprogramming of the transcriptome and alteration of proteome complexity. Several studies in plants indicate that splicing of pre-mRNAs is regulated by external signals (Ali and Reddy, 2008b; Duque, 2011; Mastrangelo et al., 2012; Leviatan et al., 2013). The mechanisms by which signals regulate pre-mRNA splicing are not known in plants. Figure 7 shows potential mechanisms through which signals regulate AS. Posttranslational modification of splicing factors and spliceosomal components, especially phosphorylation, are known to regulate splicing in animals. Several protein kinases that phosphorylate splicing regulators have been identified in plants (reviewed in Ali and Reddy, 2008a), and phosphoproteomic analysis revealed that several plant SR proteins and other splicing factors are phosphoproteins (de la Fuente van Bentem et al., 2006). It is possible that signals through second messengers, such as calcium, and reactive oxygen species modulate the activity of kinases and phosphatases that regulate the phosphorylation status of splicing regulators (de la Fuente van Bentem et al., 2006). Although second messengers are known to regulate transcription of many genes by modulating the activity of transcription factors (Reddy et al., 2011; Mittler et al., 2012; Smékalová et al., 2013), little is known about their role in posttranscriptional gene regulation.

Complex regulatory feedback loops linking AS, NMD, and kinase signaling may exist as exemplified by AFC2, a highly conserved LAMMER kinase, which phosphorylates splicing factors (Marquez et al., 2012). Other posttranslational modifications of splicing regulators, such as sumoylation, may also impact AS, as several RBPs including splicing regulators are found to be sumoylated in global proteomic studies (Miller et al., 2013). It is also likely that stresses may change the cellular environment (change in ion concentration and other molecules) that might have an impact on RNA structure leading to changes in splicing patterns. It has been documented that RNA structures change in response to concentration and types of ions and osmolytes (Draper, 2004; Lambert and Draper, 2007; Mullen et al., 2012) and temperature (Meyer et al., 2011). Stress signals in plants are known to alter cellular concentration of ions and osmolytes (Seki et al., 2007; Mullen et al., 2012; Chan et al., 2013). Hence, RNA structures may serve as sensors to detect changes in cellular environment or temperatures. The expression and AS of splicing regulators may also be regulated by these signals thereby impacting splicing. AS of master regulators of splicing, such as SRs and hnRNPs, is changed in response to stresses (Lazar and Goodman, 2000; Palusa et al., 2007; Reddy and Ali, 2011). Such changes will have global impacts on AS of other pre-mRNAs leading to reprogramming of gene expression.

GENOME DUPLICATION AND AS

Genome duplication and AS are two major mechanisms that lead to expansion and diversification of the proteome. Three models have been proposed to explain the relationship between genome duplication and AS. The independent model predicts no relationship between gene family size and the amount of AS, and the function-sharing model predicts a reverse correlation between gene family size and the extent of AS (Su et al., 2006; Jin et al., 2008). A study of genome duplication and AS events in mouse and humans by Jin et al. (2008) led to a third model called accelerated AS. They found that as predicted with the function-sharing model, AS decreased in large families but that AS increased in small families compared with singletons. The accelerated AS model suggests that singletons are likely to be duplicated, relaxing functional constraints allowing for more AS. Jin et al. (2008) argue that large families preclude the possibility for new useful functions to be created by AS, as different members in a large gene family could acquire new functions through sub- or neofunctionalization. In this respect, it is interesting that the family of 18 *Arabidopsis* SR genes, which partially originated from several genome duplication events (Kalyna and Barta, 2004), shows a remarkably high level of AS where 14 SR genes create at least 95 splice isoforms (Palusa et al., 2007).

Some plant studies have looked at duplication and AS in individual genes. Zhang et al. (2010c) used RT-PCR to look at 104 whole-genome gene duplicates in six organ types and in plants grown under abiotic stress. In a large majority of the duplicated pairs, they found differences in splicing patterns between the genes in one or more organs or under stress conditions, indicating AS divergence after genome duplication. Some plant genomes are polyploids as a result of hybridization rather than genome duplication. AS patterns in the wheat (*Triticum aestivum*) DREB2 homolog, WDREB2, were compared for polyploid wheat and for their diploid progenitors (Terashima and Takumi, 2009). Three variants of WDREB2 (α, β, and γ) result from AS, and the β-form is nonfunctional. In response to stresses, the WDREB2α and γ increase, but WDREB2γ remains constant in hexaploid wheat, while in the diploid progenitors the WDREB2γ decreased rapidly. The functional forms were not produced as rapidly in the hexaploids as in the diploid progenitors, suggesting that allopolyploidization inhibits the efficient splicing of the WDREB2 gene.

POTENTIAL BIOTECHNOLOGICAL APPLICATIONS OF AS

Transcriptional activation of transgenes has been widely used in plants to manipulate gene expression and to introduce desired traits into plants. Understanding the mechanisms involved in regulation of AS and the functions of splice variants will lead to novel ways to regulate gene expression and to develop crop plants with novel traits. For instance, expression of individual splice variants may have a different outcome compared with expression of all isoforms from a gene. Insertions of the first intron from the *UBQ10* gene into intronless genes markedly increased mRNA accumulation over 150-fold (Emami et al., 2013). Recently, plant 5S RNA mimic (discussed above), a naturally occurring alternatively spliced suicide exon (Hammond et al., 2009; Barbazuk, 2010), has been used to regulate gene expression at the pre-mRNA splicing level (Hickey et al., 2012). In this case, skipping of a suicide exon produces a functional transcript, whereas inclusion of this exon generates a transcript that is a target for NMD. With the optimized version of this suicide exon, ~100-fold gene activation was achieved by regulating AS. This work opens new avenues for the conditional expression of multiple genes at the splicing level by incorporating the same suicide exon into the transgenes. In another case, addition of the TPP riboswitch in *THIC* to a reporter construct resulted in metabolite-dependent regulation of reporter gene expression, with decreased expression under high TPP levels (Bocobza et al., 2007), suggesting a potential use of riboswitches for conditional expression of transgenes. Expression of a bacterial RNA chaperone that modulates RNA structure has also been shown to confer abiotic stress tolerance in maize under field conditions and increase yield (Castiglioni et al., 2008). A thorough understanding of AS regulatory elements and proteins that bind to these elements will offer novel ways to manipulate splicing in vivo and manipulate gene expression.

KEY FUTURE CHALLENGES

From limited transcriptome-wide studies in plants, it is clear that AS is much more prevalent in plants than previously thought. The true extent of AS from multiexon genes in plants is expected to be greater than now known when more information is gained from in-depth AS studies with different tissues/cell types and in response to developmental cues and diverse environmental and stress signals. Indeed, from recent studies, it is now becoming clear that AS is regulated developmentally and in response to diverse stress signals. These studies underscore the importance of understanding the functions of splice variants and mechanisms that regulate AS. Some important unanswered questions that need to be addressed are discussed below.

It is hypothesized that splice variants generated by AS have the potential to increase the number of proteins produced. However, currently, it is not known, both in plants and animals, what fraction of splice variants contributes to proteome diversity with functional consequences. In addition to contributing to proteome diversity, AS can be an important factor in fine-tuning gene expression at the transcript level by regulating the amount of stable and unstable transcripts. The percentage of mRNA isoforms that contributes to gene regulation at the mRNA stability through NMD, small noncoding RNAs, and other mechanisms (e.g., translation efficiency) is not known. Although a few recent studies point to an important role for AS in plant development and stress responses, functional analysis of individual mRNA isoforms is necessary to understand the importance of AS in plant growth, development, and, especially, stress responses. As there are thousands of mRNA isoforms, it is a daunting task to study the functions of each and every isoform. Initial focus should be on AS events that are conserved among all land plants, as such events are likely to have functional consequences as shown for the plant-specific SR protein families (Iida and Go, 2006; Kalyna et al., 2006). There is evidence that AS remodels protein–protein interactions in regulating cellular processes (Buljan et al., 2012; Ellis et al., 2012). Pairwise interaction of proteins coded by all AS forms using yeast two-hybrid could shed light on the functions of mRNA isoforms. It has been shown that tissue-specific regulated AS of exons modifies surface loops and unstructured binding regions of encoded proteins and rewires the protein–protein interaction networks (Buljan et al., 2012; Ellis et al., 2012).

The regulation of AS and the roles of the diverse set of RBPs (SRs and hnRNPs) are almost completely unknown. The availability of an in vitro spicing system in animals has contributed to enormous progress in understanding steps in spliceosome assembly, spliceosome composition, and *cis*-elements in RNA that regulate constitutive and regulated splicing. The Holy Grail in plant splicing research is a plant-derived in vitro splicing system. With new technological advances to isolate pure nuclei using tagged nuclei from plant cells (Deal and Henikoff, 2010, 2011) it is worth revisiting this. The accumulating data on AS in different plants across the phylogenetic spectrum will allow the development of computational tools to identify conserved, lineage-specific AS and predict regulatory elements (enhancers and suppressors) that determine the outcomes of splicing. The ultimate lofty goal in using computational tools to analyze AS is to develop algorithms that can predict if a pre-mRNA is alternatively spliced in a cell-specific manner and/or under certain cellular environments based on nucleotide sequence. To develop such algorithms requires a vast amount of empirical data on sequence elements that bind to splicing regulators and on the complexity of AS and splicing factor expression in diverse plants.

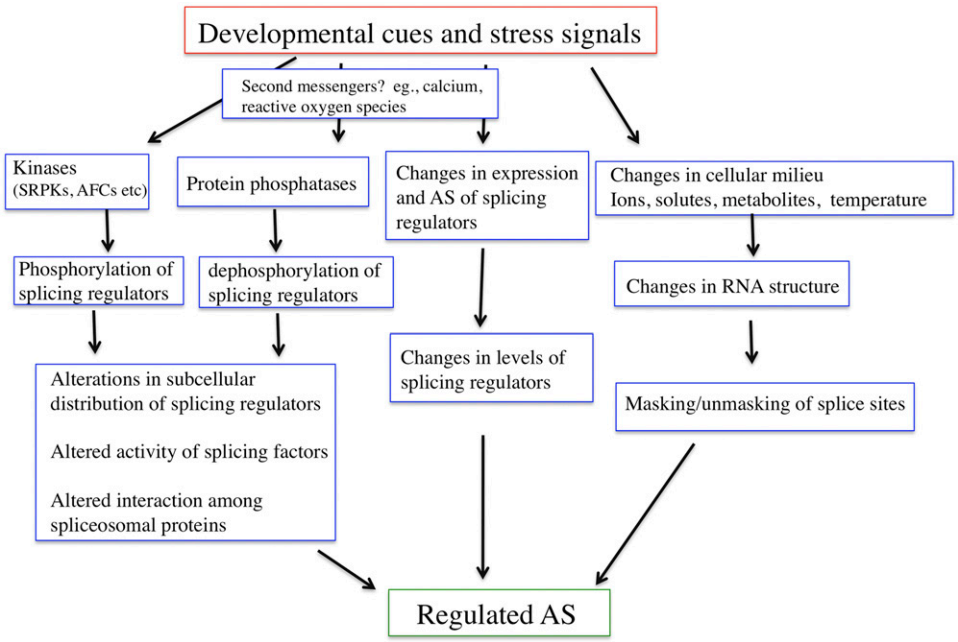


Figure 7. Potential Mechanisms through Which Developmental Cues and Environmental Signals Regulate AS.

Signal-induced changes in messengers such as calcium and reactive oxygen species, which are known regulate transcription of numerous genes (Reddy et al., 2011; Mittler et al., 2012; Smékalová et al., 2013), may also regulate AS through protein kinases and protein phosphatases as the activity of many splicing regulators is modulated by protein phosphorylation and dephosphorylation. In addition, the impact of signal-induced changes in cellular milieu, which are known to occur in plants (Seki et al., 2007; Mullen et al., 2012; Chan et al., 2013), on RNA structure and AS is unexplored in plants. See text under Elucidating Pathways That Connect Signals to AS Regulation for details pertinent to this figure. SRPK, Serine/arginine-rich protein-specific kinase; AFC2, *Arabidopsis fus3*-complementing cDNA 1 (AFC1) homolog (a LAMMER kinase).
[See online article for color version of this figure.]

Understanding how developmental and environmental signals are transduced into AS regulation and elucidating the role of AS in reprogramming gene expression in response to biotic and abiotic stresses will have implications in crop improvement. This will open new avenues to enhance and/or fine-tune gene regulation for biotechnological applications.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Table 1. Spliceosomal Proteins in Human and *Arabidopsis thaliana.*

ACKNOWLEDGMENTS

Pre-mRNA splicing research in the Reddy laboratory is supported by grants from the National Science Foundation, the Department of Energy, and the USDA. A.B. was supported by the Austrian Science Fund (DK W1207, SFB RNA-REG F43-P10, ERA-NET Plant Genomics [PASAS] I254) and the Austria Genomic Program (GENAU III [ncRNAs], the European Union FP6 NoE on AS [EURASNET]; LSHG-CT-2005-518238). We thank all current and past members of the Barta and Reddy laboratories for their contributions to splicing research. We also thank Irene Day, Denghui Xing, and Federico Martin for their comments on the article.

AUTHOR CONTRIBUTIONS

All authors contributed to writing the article.

Received August 17, 2013; revised September 27, 2013; accepted October 8, 2013; published October 31, 2013.

REFERENCES

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. **8:** 135–141.

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.

Alioto, T.S. (2007). U12DB: A database of orthologous U12-type spliceosomal introns. Nucleic Acids Res. **35** (Database issue)**:** D110–D115.

Alló, M., et al. (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat. Struct. Mol. Biol. **16:** 717–724.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T., and Ast, G. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. **1:** 543–556.

Arciga-Reyes, L., Wootton, L., Kieffer, M., and Davies, B. (2006). UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in *Arabidopsis*. Plant J. **47:** 480–489.

Au, K.F., Jiang, H., Lin, L., Xing, Y., and Wong, W.H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. **38:** 4570–4578.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. Plant Cell **15:** 2730–2741.

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 **7:** 29–33.

Awan, A.R., Manfredo, A., and Pleiss, J.A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc. Natl. Acad. Sci. USA **110:** 12762–12767.

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 Mol. Biol. **67:** 499–510.

Baltz, A.G., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell **46:** 674–690.

Bannikova, O., Zywicki, M., Marquez, Y., Skrahina, T., Kalyna, M., and Barta, A. (2013). Identification of RNA targets for the nuclear multidomain cyclophilin atCyp59 and their effect on PPIase activity. Nucleic Acids Res. **41:** 1783–1796.

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.

Barbazuk, W.B. (2010). A conserved alternative splicing event in plants reveals an ancient exonization of 5S rRNA that regulates TFIIIA. RNA Biol. **7:** 397–402.

Barbazuk, W.B., Fu, Y., and McGinnis, K.M. (2008). Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Res. **18:** 1381–1392.

Barbosa-Morais, N.L., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science **338:** 1587–1593.

Barkan, A. (2009). Genome-wide analysis of RNA-protein interactions in plants. Methods Mol. Biol. **553:** 13–37.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell **129:** 823–837.

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

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

Barta, A., Marquez, Y., and Brown, J.W.S. (2012). Challenges in plant alternative splicing. In Alternative Pre-mRNA Splicing: Theory and Protocols, C.W.J. Smith and R. Lührmann, eds (Weinheim, Germany: Wiley-VCH Verlag), pp. 79–89.

Barta, A., Sommergruber, K., Thompson, D., Hartmuth, K., Matzke, M.A., and Matzke, M.A. (1986). The expression of a nopaline synthase-human growth hormone chimeric gene in transformed tobacco and sunflower callus tissue. Plant Mol. Biol. **6:** 347–357.

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

Basu, M.K., Makalowski, W., Rogozin, I.B., and Koonin, E.V. (2008). U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. Biol. Direct **3:** 19.

Bayne, E.H., Portoso, M., Kagansky, A., Kos-Braun, I.C., Urano, T., Ekwall, K., Alves, F., Rappsilber, J., and Allshire, R.C. (2008). Splicing factors facilitate RNAi-directed silencing in fission yeast. Science **322:** 602–606.

Behm-Ansmant, I., Kashima, I., Rehwinkel, J., Saulière, J., Wittkopp, N., and Izaurralde, E. (2007). mRNA quality control: An ancient machinery recognizes and degrades mRNAs with nonsense codons. FEBS Lett. **581:** 2845–2853.

Bielewicz, D., Kalak, M., Kalyna, M., Windels, D., Barta, A., Vazquez, F., Szweykowska-Kulinska, Z., and Jarmolowski, A. (2013). Introns of plant pri-miRNAs enhance miRNA biogenesis. EMBO Rep. **14:** 622–628.

Blencowe, B.J. (2006). Alternative splicing: New insights from global analyses. Cell **126:** 37–47.

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 Dev. **21:** 2874–2879.

Bocobza, S.E., and Aharoni, A. (2008). Switching the light on plant riboswitches. Trends Plant Sci. **13:** 526–533.

Bocobza, S.E., Malitsky, S., Araújo, W.L., Nunes-Nesi, A., Meir, S., Shapira, M., Fernie, A.R., and Aharoni, A. (2013). Orchestration of thiamin biosynthesis and central metabolism by combined action of the thiamin pyrophosphate riboswitch and the circadian clock in *Arabidopsis*. Plant Cell **25:** 288–307.

Boothby, T.C., Zipper, R.S., van der Weele, C.M., and Wolniak, S.M. (2013). Removal of retained introns regulates translation in the rapidly developing gametophyte of *Marsilea vestita*. Dev. Cell **24:** 517–529.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell **132:** 311–322.

Bradley, R.K., Merkin, J., Lambert, N.J., and Burge, C.B. (2012). Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. **10:** e1001229.

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

Breaker, R.R. (2011). Prospects for riboswitch discovery and analysis. Mol. Cell **43:** 867–879.

Brown, J.W., Marshall, D.F., and Echeverria, M. (2008). Intronic noncoding RNAs and splicing. Trends Plant Sci. **13:** 335–342.

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

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell **46:** 871–883.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. Mol. Cell. Biol. **24:** 10505–10514.

Buratti, E., Muro, A.F., Giombi, M., Gherbassi, D., Iaconcig, A., and Baralle, F.E. (2004). RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. Mol. Cell. Biol. **24:** 1387–1400.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics **7:** 327.

Carpenter, C.D., Kreps, J.A., and Simon, A.E. (1994). Genes encoding glycine-rich *Arabidopsis thaliana* proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm. Plant Physiol. **104:** 1015–1025.

Castiglioni, P., et al. (2008). Bacterial RNA chaperones confer abiotic stress tolerance in plants and improved grain yield in maize under water-limited conditions. Plant Physiol. **147:** 446–455.

Chan, K.X., Wirtz, M., Phua, S.Y., Estavillo, G.M., and Pogson, B.J. (2013). Balancing metabolites in drought: The sulfur assimilation conundrum. Trends Plant Sci. **18:** 18–29.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. **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, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. **10:** 741–754.

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

Christie, M., Croft, L.J., and Carroll, B.J. (2011). Intron splicing suppresses RNA silencing in *Arabidopsis*. Plant J. **68:** 159–167.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature **452:** 215–219.

Coulter, L.R., Landree, M.A., and Cooper, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol. **17:** 2143–2150.

Cui, L., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. **16:** 738–749.

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 J. **71:** 936–947.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev. Cell **18:** 1030–1040.

Deal, R.B., and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. Nat. Protoc. **6:** 56–68.

de la Fuente van Bentem, S., Anrather, D., Roitinger, E., Djamei, A., Hufnagl, T., Barta, A., Csaszar, E., Dohnal, I., Lecourieux, D., and Hirt, H. (2006). Phosphoproteomics reveals extensive in vivo phosphorylation of *Arabidopsis* proteins involved in RNA metabolism. Nucleic Acids Res. **34:** 3267–3278.

Ding, Y., Kwok, C., Tang, Y., Bevilacqua, P.C., and Assmann, S.M. (2013). HIdden Codes in RNA: *In vivo* genome-wide profiling of RNA secondary structure. In Post-Transcriptional Gene Regulation in Plants (Providence, RI: American Soc. Pl. Biol), pp. 8.

Dixon, D.P., Hawkins, T., Hussey, P.J., and Edwards, R. (2009). Enzyme activities and subcellular localization of members of the *Arabidopsis* glutathione transferase superfamily. J. Exp. Bot. **60:** 1207–1218.

Donahue, C.P., Muratore, C., Wu, J.Y., Kosik, K.S., and Wolfe, M.S. (2006). Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. J. Biol. Chem. **281:** 23302–23306.

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E., and Ecker, J.R. (2012). Widespread dynamic DNA methylation in response to biotic stress. Proc. Natl. Acad. Sci. USA **109:** E2183–E2191.

Draper, D.E. (2004). A guide to ions and RNA structure. RNA **10:** 335–343.

Drechsel, G. , Kahles, A., Kesarwani, A.K., Stauffer, E., Behr, J., Drewe, P., Rätsch, G., and Wachter, A. (2013). Nonsense-mediated decay of alternative pre-mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. Plant Cell **25:** 3726–3742.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Muñoz, M.J., Alló, M., Schor, I.E., and Kornblihtt, A.R. (2013). Transcriptional elongation and alternative splicing. Biochim. Biophys. Acta **1829:** 134–140.

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

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., Wrana, J.L., and Blencowe, B.J. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol. Cell **46:** 884–892.

Emami, S., Arumainayagam, D., Korf, I., and Rose, A.B. (2013). The effects of a stimulating intron on the expression of heterologous genes in *Arabidopsis thaliana*. Plant Biotechnol. J. **11:** 555–563.

English, A.C., Patel, K.S., and Loraine, A.E. (2010). Prevalence of alternative splicing choices in *Arabidopsis thaliana*. BMC Plant Biol. **10:** 102.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., and Lührmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. Mol. Cell **36:** 593–608.

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.

Fouquet, R., Martin, F., Fajardo, D.S., Gault, C.M., Gómez, E., Tseung, C.W., Policht, T., Hueros, G., and Settles, A.M. (2011). Maize rough endosperm3 encodes an RNA splicing factor required for endosperm cell differentiation and has a nonautonomous effect on embryo development. Plant Cell **23:** 4280–4297.

Friend, K., Kolev, N.G., Shu, M.D., and Steitz, J.A. (2008). Minor-class splicing occurs in the nucleus of the Xenopus oocyte. RNA **14:** 1459–1462.

Fu, Y., Bannach, O., Chen, H., Teune, J.H., Schmitz, A., Steger, G., Xiong, L., and Barbazuk, W.B. (2009). Alternative splicing of anciently exonized 5S rRNA regulates plant transcription factor TFIIIA. Genome Res. **19:** 913–921.

Furey, T.S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. **13:** 840–852.

Gagete, A.P., Riera, M., Franco, L., and Rodrigo, M.I. (2009). Functional analysis of the isoforms of an ABI3-like factor of *Pisum sativum* generated by alternative splicing. J. Exp. Bot. **60:** 1703–1714.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods **8:** 469–477.

Gilbert, W. (1978). Why genes in pieces? Nature **271:** 501.

Girard, C., Will, C.L., Peng, J., Makarov, E.M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K., and Lührmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. Nat. Commun. **3:** 994.

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.

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

Gómez Acuña, L.I., Fiszbein, A., Alló, M., Schor, I.E., and Kornblihtt, A.R. (2013). Connections between chromatin signatures and splicing. Wiley Interdiscip. Rev. RNA **4:** 77–91.

Goodall, G.J., and Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell **58:** 473–483.

Goodall, G.J., and Filipowicz, W. (1991). Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J. **10:** 2635–2644.

Grabowski, P.J. (2007). RNA-binding proteins switch gears to drive alternative splicing in neurons. Nat. Struct. Mol. Biol. **14:** 577–579.

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., and Pierce, E.A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics **27:** 2518–2528.

Grojean, J., and Downes, B. (2010). Riboswitches as hormone receptors: Hypothetical cytokinin-binding riboswitches in *Arabidopsis thaliana*. Biol. Direct **5:** 60.

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.

Guo, S., Zheng, Y., Joung, J.G., Liu, S., Zhang, Z., Crasta, O.R., Sobral, B.W., Xu, Y., Huang, S., and Fei, Z. (2010). Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics **11:** 384.

Guttman, M., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. **28:** 503–510.

Gutzat, R., and Mittelsten Scheid, O. (2012). Epigenetic responses to stress: Triple defense? Curr. Opin. Plant Biol. **15:** 568–573.

Haas, B.J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. **8:** 1494–1512.

Hafner, M., et al. (2010a). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell **141:** 129–141.

Hafner, M., et al. (2010b). PAR-CliP—A method to identify transcriptome-wide the binding sites of RNA binding proteins. J. Vis. Exp. **41:** 2034.

Hallegger, M., Llorian, M., and Smith, C.W. (2010). Alternative splicing: Global insights. FEBS J. **277:** 856–866.

Hammond, M.C., Wachter, A., and Breaker, R.R. (2009). A plant 5S ribosomal RNA mimic regulates alternative splicing of transcription factor IIIA pre-mRNAs. Nat. Struct. Mol. Biol. **16:** 541–549.

Han, S.P., Tang, Y.H., and Smith, R. (2010). Functional diversity of the hnRNPs: Past, present and perspectives. Biochem. J. **430:** 379–392.

Han, T.W., Kato, M., Xie, S., Wu, L.C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G., and McKnight, S.L. (2012). Cell-free formation of RNA granules: Bound RNAs identify features and components of cellular assemblies. Cell **149:** 768–779.

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

Heintzen, C., Melzer, S., Fischer, R., Kappeler, S., Apel, K., and Staiger, D. (1994). A light- and temperature-entrained circadian clock controls expression of transcripts encoding nuclear proteins with homology to RNA-binding proteins in meristematic tissue. Plant J. **5:** 799–813.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., Fields, S., and Stamatoyannopoulos, J.A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods **6:** 283–289.

Hickey, S.F., Sridhar, M., Westermann, A.J., Qin, Q., Vijayendra, P., Liou, G., and Hammond, M.C. (2012). Transgene regulation in plants by alternative splicing of a suicide exon. Nucleic Acids Res. **40:** 4701–4710.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004). Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat. Genet. **36:** 1255–1257.

Hirsch, J., Lefort, V., Vankersschaver, M., Boualem, A., Lucas, A., Thermes, C., d'Aubenton-Carafa, Y., and Crespi, M. (2006). Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts. Plant Physiol. **140:** 1192–1204.

Hori, K., and Watanabe, Y. (2005). UPF3 suppresses aberrant spliced mRNA in *Arabidopsis*. Plant J. **43:** 530–540.

Hori, K., and Watanabe, Y. (2007). Context analysis of termination codons in mRNA that are recognized by plant NMD. Plant Cell Physiol. **48:** 1072–1078.

Hoskins, A.A., and Moore, M.J. (2012). The spliceosome: A flexible, reversible macromolecular machine. Trends Biochem. Sci. **37:** 179–188.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. **26:** 2119–2137.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., Ares, M., Jr., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. **1:** 167–178.

Hunt, A.G. (2011). RNA regulatory elements and polyadenylation in plants. Front. Plant Sci. **2:** 109.

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.

Iida, K., Shionyu, M., and Suso, Y. (2008). Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals. Mol. Biol. Evol. **25:** 709–718.

Isken, O., and Maquat, L.E. (2007). Quality control of eukaryotic mRNA: Safeguarding cells from abnormal mRNA function. Genes Dev. **21:** 1833–1856.

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.

Iwata, H., and Gotoh, O. (2011). Comparative analysis of information contents relevant to recognition of introns in many species. BMC Genomics **12:** 45.

Izquierdo, J.M., Majós, N., Bonnal, S., Martínez, C., Castelo, R., Guigó, R., Bilbao, D., and Valcárcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol. Cell **19:** 475–484.

James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., Herzyk, P., Brown, J.W., and Nimmo, H.G. (2012). Alternative splicing mediates responses of the *Arabidopsis* circadian clock to temperature changes. Plant Cell **24:** 961–981.

Jean, G., Kahles, A., Sreedharan, V.T., De Bona, F., and Ratsch, G. (2010). RNA-Seq read alignments with PALMapper. Curr. Protoc. Bioinformatics. Unit 11.6.

Jiao, Y., and Meyerowitz, E.M. (2010). Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. Mol. Syst. Biol. **6:** 419.

Jin, L., Kryukov, K., Clemente, J.C., Komiyama, T., Suzuki, Y., Imanishi, T., Ikeo, K., and Gojobori, T. (2008). The evolutionary relationship between gene duplication and alternative splicing. Gene **427:** 19–31.

Jin, Y., Yang, Y., and Zhang, P. (2011). New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. RNA Biol. **8:** 450–457.

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.

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

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., Lopato, S., Voronin, V., and Barta, A. (2006). Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. Nucleic Acids Res. **34:** 4395–4405.

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

Kato, M., et al. (2012). Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. Cell **149:** 753–767.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. Nat. Rev. Genet. **11:** 345–355.

Kerényi, Z., Mérai, Z., Hiripi, L., Benkovics, A., Gyula, P., Lacomme, C., Barta, E., Nagy, F., and Silhavy, D. (2008). Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. EMBO J. **27:** 1585–1595.

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.

Kertész, S., Kerényi, Z., Mérai, Z., Bartos, I., Pálfy, T., Barta, E., and Silhavy, D. (2006). Both introns and long 3′-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. Nucleic Acids Res. **34:** 6147–6157.

Kervestin, S., and Jacobson, A. (2012). NMD: A multifaceted response to premature translational termination. Nat. Rev. Mol. Cell Biol. **13:** 700–712.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H., Marr, M.T., II., and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. **25:** 2502–2512.

Kim, J.M., To, T.K., Nishioka, T., and Seki, M. (2010). Chromatin regulation functions in plant abiotic stress responses. Plant Cell Environ. **33:** 604–611.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. **41:** 376–381.

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.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science **329:** 336–339.

König, 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. Nat. Struct. Mol. Biol. **17:** 909–915.

Kortmann, J., and Narberhaus, F. (2012). Bacterial RNA thermometers: Molecular zippers and switches. Nat. Rev. Microbiol. **10:** 255–265.

Kurihara, Y., et al. (2009). Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **106:** 2453–2458.

Kurihara, Y., and Watanabe, Y. (2004). *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. Proc. Natl. Acad. Sci. USA **101:** 12753–12758.

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.

Lambermon, M.H., Fu, Y., Wieczorek Kirk, D.A., Dupasquier, M., Filipowicz, W., and Lorković, 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.

Lambermon, M.H., Simpson, G.G., Wieczorek Kirk, D.A., Hemmings-Mieszczak, M., Klahre, U., and Filipowicz, W. (2000). UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation. EMBO J. **19:** 1638–1649.

Lambert, D., and Draper, D.E. (2007). Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA-Mg2+ interactions. J. Mol. Biol. **370:** 993–1005.

Lamberto, I., Percudani, R., Gatti, R., Folli, C., and Petrucco, S. (2010). Conserved alternative splicing of *Arabidopsis* transthyretin-like determines protein localization and S-allantoin synthesis in peroxisomes. Plant Cell **22:** 1564–1574.

Lander, E.S., et al; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature **409:** 860–921. Erratum. Nature **411:** 720.

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

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., and Wei, C.L. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. **20:** 320–331.

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

Lee, B.H., Kapoor, A., Zhu, J., and Zhu, J.K. (2006). STABILIZED1, a stress-upregulated nuclear protein, is required for pre-mRNA splicing, mRNA turnover, and stress tolerance in *Arabidopsis*. Plant Cell **18:** 1736–1749.

Leviatan, N., Alkan, N., Leshkowitz, D., and Fluhr, R. (2013). Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. PLoS ONE **8:** e66511.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012). Regulatory impact of RNA secondary structure across the *Arabidopsis* transcriptome. Plant Cell **24:** 4346–4359.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: Global insights into biological networks. Nat. Rev. Genet. **11:** 75–87.

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.

Liu, H.X., Goodall, G.J., Kole, R., and Filipowicz, W. (1995). Effects of secondary structure on pre-mRNA splicing: Hairpins sequestering the 5′ but not the 3′ splice site inhibit intron processing in *Nicotiana plumbaginifolia*. EMBO J. **14:** 377–388.

Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., and Qi, X. (2013). An autoregulatory loop controlling Arabidopsis HsfA2 expression: Role of heat shock-induced alternative splicing. Plant Physiol. **162:** 512–521.

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.

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 Mol. Biol. **62:** 637–653. Erratum. Plant Mol. Biol. **63:** 589.

Lopato, S., Forstner, C., Kalyna, M., Hilscher, J., Langhammer, U., Indrapichate, K., Lorković, 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.

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 Mol. Biol. **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., Waigmann, E., and Barta, A. (1996). Characterization of a novel arginine/serine-rich splicing factor in *Arabidopsis*. Plant Cell **8:** 2255–2264.

Loraine, A.E., McCormick, S., Estrada, A., Patel, K., and Qin, P. (2013). RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. Plant Physiol. **162:** 1092–1109.

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

Lorkovic, Z.J., Lehner, R., Forstner, C., and Barta, A. (2005). Evolutionary conservation of minor U12-type spliceosome between plants and humans. RNA **11:** 1095–1107.

Lorković, Z.J., Wieczorek Kirk, D.A., Klahre, U., Hemmings-Mieszczak, M., and Filipowicz, W. (2000). RBP45 and RBP47, two oligouridylate-specific hnRNP-like proteins interacting with poly(A)+ RNA in nuclei of plant cells. RNA **6:** 1610–1624.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. USA **108:** 11063–11068.

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.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science **327:** 996–1000.

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 Res. **22:** 1184–1195.

Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics **11:** 663.

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

Mastrangelo, A.M., Marone, D., Laidò, G., De Leonardis, A.M., and De Vita, P. (2012). Alternative splicing: Enhancing ability to cope with stress via transcriptome plasticity. Plant Sci. **185-186:** 40–49.

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.

McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. Curr. Opin. Genet. Dev. **21:** 373–379.

Melamud, E., and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. **37:** 4873–4886.

Meng, Y., Shao, C., Ma, X., and Wang, H. (2013). Introns targeted by plant microRNAs: A possible novel mechanism of gene regulation. Rice **6:** 1–10.

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

Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E., and Vilardell, J. (2011). Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. Mol. Cell **43:** 1033–1039.

Miller, M.J., Scalf, M., Rytz, T.C., Hubler, S.L., Smith, L.M., and Vierstra, R.D. (2013). Quantitative proteomics reveals factors regulating RNA biology as dynamic targets of stress-induced SUMOylation in *Arabidopsis*. Mol. Cell. Proteomics **12:** 449–463.

Mittler, R., Finka, A., and Goloubinoff, P. (2012). How do plants feel the heat? Trends Biochem. Sci. **37:** 118–125.

Monaghan, J., Xu, F., Gao, M., Zhao, Q., Palma, K., Long, C., Chen, S., Zhang, Y., and Li, X. (2009). Two Prp19-like U-box proteins in the MOS4-associated complex play redundant roles in plant innate immunity. PLoS Pathog. **5:** e1000526.

Morrison, M., Harris, K.S., and Roth, M.B. (1997). smg mutants affect the expression of alternatively spliced SR protein mRNAs in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA **94:** 9782–9785.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **5:** 621–628.

Mullen, M.A., Assmann, S.M., and Bevilacqua, P.C. (2012). Toward a digital gene response: RNA G-quadruplexes with fewer quartets fold with higher cooperativity. J. Am. Chem. Soc. **134:** 812–815.

Mulligan, G.J., Guo, W., Wormsley, S., and Helfman, D.M. (1992). Polypyrimidine tract binding protein interacts with sequences involved in alternative splicing of beta-tropomyosin pre-mRNA. J. Biol. Chem. **267:** 25480–25487.

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.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M., Jr.,. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. **21:** 708–718.

Nicholson, P., Yepiskoposyan, H., Metze, S., Zamudio Orozco, R., Kleinschmidt, N., and Mühlemann, O. (2010). Nonsense-mediated mRNA decay in human cells: Mechanistic insights, functions beyond quality control and the double-life of NMD factors. Cell. Mol. Life Sci. **67:** 677–700.

Nyikó, T., Sonkoly, B., Mérai, Z., Benkovics, A.H., and Silhavy, D. (2009). Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. Plant Mol. Biol. **71:** 367–378.

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

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.

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.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R., and Blencowe, B.J. (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet. **21:** 73–77.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. **20:** 153–158.

Pelechano, V., Wei, W., and Steinmetz, L.M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. Nature **497:** 127–131.

Penfield, S., Josse, E.M., and Halliday, K.J. (2010). A role for an alternative splice variant of PIF6 in the control of *Arabidopsis* primary seed dormancy. Plant Mol. Biol. **73:** 89–95.

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.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. **6:** e1001236.

Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R., and Bickmore, W.A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. PLoS Genet. **8:** e1002717.

Raczynska, K.D., Simpson, C.G., Ciesiolka, A., Szewc, L., Lewandowska, D., McNicol, J., Szweykowska-Kulinska, Z., Brown, J.W., and Jarmolowski, A. (2010). Involvement of the nuclear cap-binding protein complex in alternative splicing in *Arabidopsis thaliana*. Nucleic Acids Res. **38:** 265–278.

Ramani, A.K., Nelson, A.C., Kapranov, P., Bell, I., Gingeras, T.R., and Fraser, A.G. (2009). High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective *Caenorhabditis elegans*. Genome Biol. **10:** R101.

Reddy, A.S., Ali, G.S., Celesnik, H., and Day, I.S. (2011). Coping with stresses: Roles of calcium- and calcium/calmodulin-regulated gene expression. Plant Cell **23:** 2010–2032.

Reddy, A.S., Day, I.S., Göhring, 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. Front Plant Sci **3:** 18.

Reddy, A.S.N. (2001). 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 serine/arginine-rich proteins: Roles in precursor messenger RNA splicing, plant development and stress responses. Wiley Interdiscip. Rev. RNA **2:** 875–889.

Rehwinkel, J., Letunic, I., Raes, J., Bork, P., and Izaurralde, E. (2005). Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. RNA **11:** 1530–1544.

Remy, E., Cabrito, T.R., Baster, P., Batista, R.A., Teixeira, M.C., Friml, J., Sá-Correia, I., and Duque, P. (2013). A major facilitator superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in *Arabidopsis*. Plant Cell **25:** 901–926.

Riehs, N., Akimcheva, S., Puizina, J., Bulankova, P., Idol, R.A., Siroky, J., Schleiffer, A., Schweizer, D., Shippen, D.E., and Riha, K. (2008). *Arabidopsis* SMG7 protein is required for exit from meiosis. J. Cell Sci. **121:** 2208–2216.

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.

Robertson, G., et al. (2010). De novo assembly and analysis of RNA-seq data. Nat. Methods **7:** 909–912.

Roca, X., Krainer, A.R., and Eperon, I.C. (2013). Pick one, but be quick: 5′ Splice sites and the problems of too many choices. Genes Dev. **27:** 129–144.

Rogers, K., and Chen, X. (2013a). Biogenesis, turnover, and mode of action of plant microRNAs. Plant Cell **25:** 2383–2399.

Rogers, K., and Chen, X. (2013b). MicroRNA biogenesis and turnover in plants. Cold Spring Harb. Symp. Quant. Biol. **77:** 183–194.

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.

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

Rühl, C., Stauffer, E., Kahles, A., Wagner, G., Drechsel, G., Rätsch, G., and Wachter, A. (2012). Polypyrimidine tract binding protein homologs from *Arabidopsis* are key regulators of alternative splicing with implications in fundamental developmental processes. Plant Cell **24:** 4360–4375.

Sakharkar, M.K., Chow, V.T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol. (Gedrukt) **4:** 387–393.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. **28:** 4320–4330.

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.

Schoenberg, D.R., and Maquat, L.E. (2012). Regulation of cytoplasmic mRNA decay. Nat. Rev. Genet. **13:** 246–259.

Schöning, 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.

Schöning, 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., Llères, D., Risso, G.J., Pawellek, A., Ule, J., Lamond, A.I., and Kornblihtt, A.R. (2012). Perturbation of chromatin structure globally affects localization and recruitment of splicing factors. PLoS ONE **7:** e48084.

Schwab, R., Speth, C., Laubinger, S., and Voinnet, O. (2013). Enhanced microRNA accumulation through stemloop-adjacent introns. EMBO Rep. **14:** 615–621.

Schwartz, A.M., Komarova, T.V., Skulachev, M.V., Zvereva, A.S., Dorokhov, IuL., and Atabekov, J.G. (2006). Stability of plant mRNAs depends on the length of the 3′-untranslated region. Biochemistry Mosc. **71:** 1377–1384.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. **16:** 990–995.

Seki, M., Umezawa, T., Urano, K., and Shinozaki, K. (2007). Regulatory metabolic networks in drought stress responses. Curr. Opin. Plant Biol. **10:** 296–302.

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.

Seo, P.J., Park, M.J., and Park, C.M. (2013). Alternative splicing of transcription factors in plant responses to low temperature stress: Mechanisms and functions. Planta **237:** 1415–1424.

Serganov, A., and Nudler, E. (2013). A decade of riboswitches. Cell **152:** 17–24.

Severing, E.I., van Dijk, A.D., Morabito, G., Busscher-Lange, J., Immink, R.G., and van Ham, R.C. (2012). Predicting the impact of alternative splicing on plant MADS domain protein function. PLoS ONE **7:** e30524.

Severing, E.I., van Dijk, A.D., Stiekema, W.J., and van Ham, R.C. (2009). Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC Genomics **10:** 154.

Shen, H., Kan, J.L., and Green, M.R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol. Cell **13:** 367–376.

Shepard, P.J., and Hertel, K.J. (2008). Conserved RNA secondary structures promote alternative splicing. RNA **14:** 1463–1469.

Shepard, P.J., and Hertel, K.J. (2009). The SR protein family. Genome Biol. **10:** 242.

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.

Silverman, I.M., Li, F., and Gregory, B.D. (2013). Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. Plant Sci. **205-206:** 55–62.

Simpson, C.G., and Brown, J.W. (2008). U12-dependent intron splicing in plants. Curr. Top. Microbiol. Immunol. **326:** 61–82.

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

Simpson, G.G. (2004). The autonomous pathway: Epigenetic and post-transcriptional gene regulation in the control of *Arabidopsis* flowering time. Curr. Opin. Plant Biol. **7:** 570–574.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell **151:** 750–764.

Sinha, R., Zimmer, A.D., Bolte, K., Lang, D., Reski, R., Platzer, M., Rensing, S.A., and Backofen, R. (2010). Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens*. BMC Plant Biol. **10:** 76.

Smékalová, V., Doskočilová, A., Komis, G., and Samaj, J. (2013). Crosstalk between secondary messengers, hormones and MAPK modules during abiotic stress signalling in plants. Biotechnol. Adv. (in press), doi/10.1016/j.biotechadv.2013.07.009.

Spector, D.L., and Lamond, A.I. (February 1, 2011). Nuclear speckles. Cold Spring Harb. Perspect. Biol. http://dx.doi.org/10.1101/cshperspect.a000646.

Staiger, D. (2001). RNA-binding proteins and circadian rhythms in *Arabidopsis thaliana*. Philos. Trans. R. Soc. Lond. B Biol. Sci. **356:** 1755–1759.

Staiger, D., and Brown, J.W.S. (2013). Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell **25:** 3640–3656.

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.

Stalder, L., and Mühlemann, O. (2008). The meaning of nonsense. Trends Cell Biol. **18:** 315–321.

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

Stauffer, E., Westermann, A., Wagner, G., and Wachter, A. (2010). Polypyrimidine tract-binding protein homologues from *Arabidopsis* underlie regulatory circuits based on alternative splicing and downstream control. Plant J. **64:** 243–255.

Steitz, J.A., Dreyfuss, G., Krainer, A.R., Lamond, A.I., Matera, A.G., and Padgett, R.A. (2008). Where in the cell is the minor spliceosome? Proc. Natl. Acad. Sci. USA **105:** 8485–8486.

Streitner, C., Köster, T., Simpson, C.G., Shaw, P., Danisman, S., Brown, J.W., and Staiger, D. (2012). An hnRNP-like RNA-binding protein affects alternative splicing by in vivo interaction with transcripts in *Arabidopsis thaliana*. Nucleic Acids Res. **40:** 11240–11255.

Su, Z., Wang, J., Yu, J., Huang, X., and Gu, X. (2006). Evolution of alternative splicing after gene duplication. Genome Res. **16:** 182–189.

Sugliani, M., Brambilla, V., Clerkx, E.J., Koornneef, M., and Soppe, W.J. (2010). The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in *Arabidopsis*. Plant Cell **22:** 1936–1946.

Sultan, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321:** 956–960.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W. (2012). Alternative splicing in plants—Coming of age. Trends Plant Sci. **17:** 616–623.

Szarzynska, B., Sobkowiak, L., Pant, B.D., Balazadeh, S., Scheible, W.R., Mueller-Roeber, B., Jarmolowski, A., and Szweykowska-Kulinska, Z. (2009). Gene structures and processing of *Arabidopsis thaliana* HYL1-dependent pri-miRNAs. Nucleic Acids Res. **37:** 3083–3093.

Terashima, A., and Takumi, S. (2009). Allopolyploidization reduces alternative splicing efficiency for transcripts of the wheat DREB2 homolog, WDREB2. Genome **52:** 100–105.

Thomas, J., Palusa, S.G., Prasad, K.V., 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 auto-regulation of alternative splicing of SCL33 pre-mRNA. Plant J. **72:** 935–946.

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

Tolstorukov, M.Y., Goldman, J.A., Gilbert, C., Ogryzko, V., Kingston, R.E., and Park, P.J. (2012). Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. Mol. Cell **47:** 596–607.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25:** 1105–1111.

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.

Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: Splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA **4:** 61–76.

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. Nat. Methods **7:** 995–1001.

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

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

Wachter, A., Rühl, C., and Stauffer, E. (2012). The role of polypyrimidine tract-binding proteins and other hnRNP proteins in plant splicing regulation. Front Plant Sci **3:** 81.

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.

Wagner, E.J., and Garcia-Blanco, M.A. (2001). Polypyrimidine tract binding protein antagonizes exon definition. Mol. Cell. Biol. **21:** 3281–3288.

Walters, B., Lum, G., Sablok, G., and Min, X.J. (2013). Genome-wide landscape of alternative splicing events in *Brachypodium distachyon*. DNA Res. **20:** 163–171.

Wan, Y., Kertesz, M., Spitale, R.C., Segal, E., and Chang, H.Y. (2011). Understanding the transcriptome through RNA structure. Nat. Rev. Genet. **12:** 641–655.

Wan, Y., Qu, K., Ouyang, Z., and Chang, H.Y. (2013). Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. Nat. Protoc. **8:** 849–869.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D.L., Nutter, R.C., Segal, E., and Chang, H.Y. (2012). Genome-wide measurement of RNA folding energies. Mol. Cell **48:** 169–181.

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. (2006). Genomewide comparative analysis of alternative splicing in plants. Proc. Natl. Acad. Sci. USA **103:** 7175–7180.

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, K., et al. (2010a). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. **38:** e178.

Wang, X., et al. (2012). SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in *Arabidopsis*. Plant Cell **24:** 3278–3295.

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

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010b). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol. **8:** e1000530.

Wang, Z., Tollervey, J., Briese, M., Turner, D., and Ule, J. (2009b). CLIP: Construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. Methods **48:** 287–293.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. **35:** 169–178.

Weatheritt, R.J., Davey, N.E., and Gibson, T.J. (2012). Linear motifs confer functional diversity onto splice variants. Nucleic Acids Res. **40:** 7123–7131.

Weatheritt, R.J., and Gibson, T.J. (2012). Linear motifs: Lost in (pre) translation. Trends Biochem. Sci. **37:** 333–341.

Weber, A.P., Weber, K.L., Carr, K., Wilkerson, C., and Ohlrogge, J.B. (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol. **144:** 32–42.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J.S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B.T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol. **13:** R35.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. **3:** 3.

Wu, G., and Poethig, R.S. (2006). Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. Development **133:** 3539–3547.

Wu, J., Kang, J.H., Hettenhausen, C., and Baldwin, I.T. (2007). Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. Plant J. **51:** 693–706.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26:** 873–881.

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. Proc. Natl. Acad. Sci. USA **108:** 12533–12538.

Yan, K., Liu, P., Wu, C.A., Yang, G.D., Xu, R., Guo, Q.H., Huang, J.G., and Zheng, C.C. (2012). Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in *Arabidopsis thaliana*. Mol. Cell **48:** 521–531.

Yang, G.D., Yan, K., Wu, B.J., Wang, Y.H., Gao, Y.X., and Zheng, C.C. (2012a). Genomewide analysis of intronic microRNAs in rice and *Arabidopsis*. J. Genet. **91:** 313–324.

Yang, X., Zhang, H., and Li, L. (2012b). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in *Arabidopsis*. Plant J. **70:** 421–431.

Yap, K., Lim, Z.Q., Khandelia, P., Friedman, B., and Makeyev, E.V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev. **26:** 1209–1223.

Yoine, M., Ohto, M.A., Onai, K., Mita, S., and Nakamura, K. (2006). The Iba1 mutation of UPF1 RNA helicase involved in nonsense-mediated mRNA decay causes pleiotropic phenotypic changes and altered sugar signalling in *Arabidopsis*. Plant J. **47:** 49–62.

Yoon, S.P., Kim, H.H., Kim, J., Park, R.Y., and Ohn, T. (2013). Regulation of cellular RNA nano-particle assembly by splicing factor SRp20. J. Nanosci. Nanotechnol. **13:** 184–187.

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.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M., and Delledonne, M. (2010). Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. Plant Physiol. **152:** 1787–1795.

Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J., and Darnell, R.B. (2010a). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science **329:** 439–443.

Zhang, C., Lee, K.Y., Swanson, M.S., and Darnell, R.B. (2013a). Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res. **41:** 6793–6807.

Zhang, C.J., Zhou, J.X., Liu, J., Ma, Z.Y., Zhang, S.W., Dou, K., Huang, H.W., Cai, T., Liu, R., Zhu, J.K., and He, X.J. (2013b). The splicing machinery promotes RNA-directed DNA methylation and transcriptional silencing in *Arabidopsis*. EMBO J. **32:** 1128–1140.

Zhang, G., et al. (2010b). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. **20:** 646–654.

Zhang, P.G., Huang, S.Z., Pin, A.L., and Adams, K.L. (2010c). Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of *Arabidopsis*. Mol. Biol. Evol. **27:** 1686–1697.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell **40:** 939–953.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.S., and Gregory, B.D. (2010). Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. PLoS Genet. **6:** e1001141.

Zhu, W., and Brendel, V. (2003). Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. Nucleic Acids Res. **31:** 4561–4572.

Zhu, W., Schlueter, S.D., and Brendel, V. (2003). Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. Plant Physiol. **132:** 469–484.

Curriculum Vitae

General Information

First name:	Yamile
Last name:	Márquez Ortiz
Date and place of birth:	October17th, 1986, Mexico City, Mexico
Current address:	Max F. Perutz Laboratories, Dept. of Medical Biochemistry. Dr. Bohr Gasse 9/3, 1030, Vienna, Austria.
Phone number:	(+43) 1-4277-61641
E-mail:	yamile.marquez@univie.ac.at
Current activity:	PhD student in Dr. Andrea Barta Laboratory
Last update:	March, 2014

Academic Preparation

University:	Bachelor in Genomic Sciences (Honours), Center for Genomic Sciences and Institute of Biotechnology National Autonomous University of Mexico (Aug 2004 - Dec 2008)
Spoken languages:	Spanish English

Computational and research skills

Programming in C and Perl. Design and creation of databases and web pages using as tools MySQL and PHP. Bioinformatic searching with web tools and databases. Basic training in molecular biology techniques

Professional experience

Member of Dr. Andrea Barta laboratory. Dept. of Medical Biochemistry, Medical University of Vienna, Max F. Perutz Laboratories (Jun 2009 - to date).

Stay in Dr. Gautier Koscielny group. European Bioinformatic Insitute (EBI), Hinxton, UK (Oct - Nov 2009).

Member of Dr. Federico Sánchez Rodriguez laboratory. Department of Molecular Plant Biology, Institute of Biotechnology, UNAM (Jan 2006 - Apr 2009).

Stay in Dr. Hilda María Lomelí Buyoli laboratory. Department of Genetics of the Development and Molecular Physiology, Institute of Biotechnology, UNAM (Sep - Dec 2005).

Scholarships and Awards

Honorific mention Bachelor in Genomics Sciences, UNAM (Dec 2009)

UNAM Scholarship for High Performance Students (Sep 2003 - Jun 2004).

Courses and workshops

Workshop, Modulators of RNA Fate and Function III, Vienna, Austria (Feb 2011).

Winter School in Genomics at the Center for Genomic Sciences of the UNAM (Feb 2006).

Conferences

Oral presentation “Genome-wide investigation of the Arabidopsis transcriptome reveals a new type of alternative splicing event”, ICGEB 2nd Post-EURASNET “Symposium RNA Alternative Splicing”, Trieste, Italy, Nov 2013.

Oral presentation “Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis” Plant RNA workshop, Vienna, July 2012.

Poster presentation “Arabidopsis ecotype alternative splicing variants”, Plant RNA workshop, Vienna, July 2012.

Poster presentation “Unraveling the alternative splicing landscape in Arabidopsis using RNA-seq”, 1st Post-EURASNET Symposium, Trieste, Italy, Mar 2012.

Oral presentation “Analysis of alternative splicing in Arabidopsis using RNA-seq” Minisiymposium Next generation sequencing, Vienna, Austria, Nov, 2011.

Oral presentation “Analysis of Alternative Splicing in Arabidopsis using RNA-seq”. ISMB, 8th Special Interest Group meeting on Alternative Splicing, Vienna, Austria, July 2011.

Poster presentation “RNA-seq reveals an unexpected high level of Alternative splicing in Arabidopsis” RNA Conference, Kyoto, Japan, June, 2011.

Poster presentation “Massive sequencing reveals an unexpected high level of Alternative splicing in Arabidopsis” Second International EURASNET Conference on Alternative Splicing, Granada, Spain, Feb, 2011.

Poster presentation “Massive sequencing reveals an unexpected high level of Alternative splicing in Arabidopsis” IFM, Berlin, Germany, Sep, 2011.

Oral presentation “Massive sequencing reveals an unexpected high level of alternative splicing in Arabidopsis”. SEB Annual Meeting, Prague, Czech Republic, July 2010.

Oral presentation “A new phylogenetic analysis of *Phaseolus* species: Patterns of diversification and biogeography”. Phaseomics V, Varenna, Italy, May 2007.

Publications

Reddy, A. S., Marquez, Y., Kalyna, M., & Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell, 25*(10), 3657-3683.

Bannikova, O., Zywicki, M., Marquez, Y., Skrahina, T., Kalyna, M., and Barta, A. (2013). Identification of RNA targets for the nuclear multidomain cyclophilin atCyp59 and their effect on PPIase activity. *Nucleic Acids Res*. 41: 1783–1796.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W. (2012). Alternative splicing in plants—Coming of age. *Trends Plant Sci.* 17: 616–623.

Marquez, Y., Brown, J. W., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res, 22*(6), 1184-1195.

Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., . . . Brown, J. W. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res, 40*(6), 2454-2469.

Barta, A., Marquez, Y., and Brown, J.W.S. (2012). Challenges in plant alternative splicing. In Alternative Pre-mRNA Splicing: Theory and Protocols, C.W.J. Smith and R. Lührmann, eds (Weinheim, Germany: Wiley-VCH Verlag), pp. 79–89.

Olivares JE, Díaz-Camino C, Estrada-Navarrete G, Alvarado-Affantranger X, Rodríguez-Kessler M, Zamudio FZ, Olamendi-Portugal T, Márquez Y, Servín LE and Sánchez F. Nodulin 41, a novel late nodulin of common bean with peptidase activity (2011). *BMC Plant Biol*. 11:134

Zepeda-Mendoza CJ, Lemus T, Yáñez O, García D, Valle-García D, Meza-Sosa KF, Gutiérrez-Arcelus M, Márquez-Ortiz Y, Domínguez-Vidaña R, Gonzaga-Jauregui C, Flores M, Palacios R. Identical repeated backbone of the human genome (2010). *BMC Genomics*. 11:60