Genome-wide analysis of splicing related genes and alternative splicing in plants

by

Bing-Bing Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Volker Brendel, Major Professor
Thomas Peterson
Philip Becraft
Xun Gu
Shashi Gadia

Iowa State University

Ames, Iowa

2005

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Bing-Bing Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

For my parents.

TABLE OF CONTENTS

Chapter 1: General Introduction 1
Dissertation Organization 3
Literature Cited 3
Chapter 2: The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing 5
Abstract 5
Rationale 6
ASRG: Database of Arabidopsis Splicing Related Genes 8
Arabidopsis small nuclear RNA (snRNA) genes 8
Arabidopsis splicing related protein-coding genes 13
Distribution and duplication of Arabidopsis splicing related genes 24
Alternative splicing of Arabidopsis splicing related genes 26
Discussion 27
Conclusion 31
Materials and methods 32
Acknowledgements 36
References 37
Figure Legends 49
Tables 50
Chapter 3: Genome-wide comparative analysis of alternative splicing in plants 67
Abstract 67
Introduction 68
Materials and Methods 71
Results and Discussion 76
Conclusions 88
Acknowledgments 89
References 89
Figure Legends and Tables 94
Supporting Data 100
Chapter 4: Molecular characterization and phylogeny of U2AF1 homologs in plants 111
Abstract 111
Introduction 112
Results 116
Discussion 132
Materials and Methods 135
Acknowledgments 139
Literature Cited 140
Figure Legends 145
Supplementary Data 159
Chapter 5: General Conclusion 164
Acknowledgements 166

Chapter 1: General Introduction

The accurate recognition and splicing of introns are fundamental to gene expression. The general mechanism of splicing is conserved in eukaryotes. Most introns start with the dinucleotide GU and end with AG. A consensus branch site exists near the 3' end and introns are excised through a lariat formation utilizing the branchpoint sequence. In yeast, the branchpoint sequence is completely conserved in all introns. In addition, mammalian introns have a polypyrimidine tract between the branch sequence and AG border. Plant introns are generally shorter than introns in vertebrates. They lack the polypyrimidine tract and have a loosely conserved branch region. The most important difference is that plant introns are U-rich in base composition. This composition bias is critical to the splicing of plant introns (Lorkovic et al., 2000).

The properties of plant introns make it reasonable to think that plant have special mechanism for intron recognition. In fact, plant cells can not splice heterologous pre-mRNA from mammalian sources and the mammalian *in vivo* and *in vitro* splicing systems fail to splice plant pre-mRNA faithfully (Reviewed in (Reddy, 2001)). These facts further prove that plant and animal have different splicing site recognition mechanisms. Differences are also reported to exist between monocot and dicot plants (Martin et al., 1997). Due to the lack of *in vitro* splicing system in plants, detailed splicing mechanism is poorly understood in plants. Very little is known about both the splicing machinery in plant splicing. Although hundreds of proteins were identified in metazoan splicing machinery (Krämer, 1996; Will and Lührmann, 1997; Mount and Salz, 2000), only a few were characterized in plants (Reddy, 2001). Little is

known about alternative splicing in plants, either. As an important post-transcription regulation method, alternative splicing plays important roles in the increase of protein diversity and regulating gene expression level. Over 50% of human genes can be alternative spliced. In contrast, only 5-10% genes were estimated to be alternative spliced in plants (Brett et al., 2002; Haas et al., 2003; Zhu et al., 2003).

U2AF (U2 snRNP Auxiliary Factor) is an essential splicing factor with critical roles in recognition of the 3'-splice site in animals. The U2AF protein is comprised of a large subunit ($U2AF^{65}$) and a small subunit ($U2AF^{35}$) (Zamore and Green, 1989), with $U2AF^{65}$ binding directly to the Py-tract (Zamore et al., 1992) and $U2AF^{35}$ binding to the 3' AG boundary (Merendino et al., 1999; Wu et al., 1999; Zorio and Blumenthal, 1999). Two $U2AF^{65}$ homologs were isolated from *Nicotiana plumbaginifolia* (Domon et al., 1998), while no $U2AF^{35}$ homologs have been characterized so far.

The main objective of this research is to learn more about splicing mechanism in plants. Specific goals include the following: (1) Computationally identify Arabidopsis homologs for all known genes involved in splicing. By comparing the splicing machinery of Arabidopsis with their counterparts in animals, we can get the general picture of plant splicing. This study will also serve as a base for the community to study splicing factors. (2) Identify and compare the alternative splicing events in both Arabidopsis and rice. As no such a study has been performed in monocot plants, our comparison will reveal the similarity as well as differences between the characteristics of dicot and monocot plants. (3) Experimentally characterize Arabidopsis $U2AF^{35}$ homologs.

Dissertation Organization

The dissertation is organized in the format consisting of three journal articles preceded by a General Introduction and followed by a General Conclusion. The journal articles are formatted according to the requirements of each journal. The first article "The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing" was published in *Genome Biology* (2004, 5(12):R102). The second article "Genome-wide comparative analysis of alternative splicing in plants" will be submitted for publication in *PNAS* and the third article "Molecular characterization and phylogeny of U2AF1 homologs in plants" will be submitted for publication in *Plant Physiology*. Bing-Bing Wang was the primary investigator for this work under the supervision of Dr. Volker Brendel and is the first author of all the three articles.

Literature Cited

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat Genet 30, 29-30.

Domon, C., Lorkovic, Z.J., Valcarcel, J., and Filipowicz, W. (1998). Multiple forms of the U2 small nuclear ribonucleoprotein auxiliary factor U2AF subunits expressed in higher plants. J Biol Chem 273, 34603-34610.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., and White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-5666.

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

Lorkovic, Z.J., Wieczorek Kirk, D.A., Lambermon, M.H., and Filipowicz, W. (2000). Pre-mRNA splicing in higher plants. Trends Plant Sci 5, 160-167.

Martin, D.J., Firek, S., Moreau, E., and Draper, J. (1997). Alternative processing of the maize Ac transcript in Arabidopsis. Plant J 11, 933-943.

Merendino, L., Guth, S., Bilbao, D., Martinez, C., and Valcarcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. Nature 402, 838-841.

Mount, S.M., and Salz, H.K. (2000). Pre-messenger RNA processing factors in the Drosophila genome. J Cell Biol 150, F37-F43.

Reddy, A.S.N. (2001). Nuclear pre-mRNA splicing in plants. Critical Rev Plant Sci 20, 523-571.

Will, C.L., and Lührmann, R. (1997). Protein functions in pre-mRNA splicing. Curr Opin Cell Biol 9, 320-328.

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

Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc Natl Acad Sci U S A 86, 9243-9247.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. Nature 355, 609-614.

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.

Zorio, D.A., and Blumenthal, T. (1999). U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. RNA 5, 487-494.

Chapter 2: The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing

A paper published in Genome Biology

Bing-Bing Wang[1] and Volker Brendel[1,2]

[1]Department of Genetics, Development and Cell Biology and [2]Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

Abstract

A total of 74 snRNA genes and 395 genes encoding splicing related proteins were identified in the Arabidopsis genome by sequence comparison and motif searches, including the previously elusive U4atac snRNA gene. Most of the genes have not been experimentally studied. Classification of these genes and detailed information on gene structure, alternative splicing, gene duplications, and phylogenetic relationships are made accessible as a comprehensive database of Arabidopsis Splicing Related Genes (ASRG) at http://www.plantgdb.org/SRGD/ASRG/.

Rationale

Most eukaryotic genes contain introns that are spliced from the precursor mRNA (pre-mRNA). The correct interpretation of splicing signals is essential to generate authentic mature mRNAs that yield correct translation products. As an important post-transcriptional mechanism, gene function can be controlled at the level of splicing through the production of different mRNAs from a single pre-mRNA (reviewed in [1]). The general mechanism of splicing has been well studied in human and yeast systems and is largely conserved between these organisms. Comparatively, plant splicing mechanisms remain poorly understood due in part to the lack of an *in vitro* plant splicing system. Although the splicing mechanisms in plants and animals appear to be similar overall, incorrect splicing of plant pre-mRNAs in mammalian systems (and vice versa) suggests that there are plant-specific characteristics, resulting from co-evolution of splicing factors with the signals they recognize or from the requirement for additional splicing factors (reviewed in [2, 3]).

Genome projects are accelerating research on splicing. For example, with the majority of splicing related genes already known in human and budding yeast, these gene sequences were used to query the Drosophila and fission yeast genomes in an effort to identify potential homologs [4, 5]. Most of the known genes were found to have homologs in both Drosophila and fission yeast. The availability of the near-complete genome of *Arabidopsis thaliana* [6] provides the foundation for the simultaneous study of all the genes involved in particular plant structures or physiological processes. For example, Barakat *et al.* [7] identified and mapped 249 genes encoding ribosomal proteins and analyzed gene number, chromosomal location, evolutionary history (including large-scale chromosomal duplications), and

expression of those genes. Beisson *et al.* [8] catalogued all genes involved in acyl lipid metabolism. Wang *et al.* [9] surveyed more than 1,000 Arabidopsis protein kinases and computationally compared derived protein clusters with established gene families in budding yeast. Previous surveys of Arabidopsis gene families that contain some splicing related genes include the DEAD box RNA helicase family [10] and RNA recognition motif (RRM) containing proteins [11]. Presently, the Arabidopsis Information Resource (TAIR) links to more than 850 such expert-maintained collections of gene families.

Here we present the results of computational identification of potentially all or nearly all Arabidopsis genes involved in pre-mRNA splicing. Recent mass spectrometry analyses revealed more than 200 proteins associated with human spliceosome ([12-16], reviewed in [17]). By extensive sequence comparisons using known plant and animal splicing related proteins as queries, we have identified 74 snRNA genes and 395 protein-coding genes in the Arabidopsis genome that are likely to be homologs of animal splicing related genes. About half of the genes occur in multiple copies in the genome and appear to have been derived from both chromosomal duplication events and from duplication of individual genes. All genes were classified into gene families, named, and annotated with respect to their inferred gene structure, predicted protein domain structure, and presumed function. The classification and analysis results are available as an integrated web resource, which should facilitate genome-wide studies of pre-mRNA splicing in plants.

ASRG: Database of Arabidopsis Splicing Related Genes

Our up to date web-accessible database comprising the Arabidopsis splicing related genes and associated information is available at . The web pages display gene structure, alternative splicing patterns, protein domain structure, and potential gene duplication origins in tabular format. Chromosomal locations and spliced alignment of cognate cDNAs and ESTs are viewable via links to AtGDB, which also provides other associated information for these genes and links to other databases. Text search functions are accessible from all the web pages. Sequence analysis tools including BLAST [18] and CLUSTALW [19] are integrated and facilitate comparison of splicing related genes and proteins across various species.

Arabidopsis small nuclear RNA (snRNA) genes

A total of 15 major snRNA and two minor snRNA genes were previously identified experimentally in Arabidopsis [20-25]. These genes were used as queries to search the Arabidopsis genome for other snRNA genes. A total of 70 major snRNAs and three minor snRNAs were identified by this method. In addition, a single U4atac snRNA gene was identified by sequence motif search. We assigned tentative gene names and gene models as shown in Table 1, together with chromosome locations and similarity scores relative to a representative query sequence. The original names for known snRNAs were preserved, following the convention atUx.y, where x indicates the U snRNA type and y the gene number. Computationally identified snRNAs were named similarly, but with a hyphen instead of a period separating type from gene number (atUx-y). Putative pseudogenes were indicated with a "p" following the gene name. Pseudogene status was assigned to gene

models for which sequence similarity to known genes was low, otherwise conserved transcription signals are missing, and the gene cannot fold into typical secondary structure. A recent experimental study of small non-messenger RNAs identified 14 tentative snRNAs in Arabidopsis by cDNA cloning ([26], GenBank accessions 22293580 to 22293592 and 22293600, Table 1). All these newly identified snRNAs were found in the set of our computationally predicted genes.

Conservation of major snRNA genes

As shown in Table 1, each of five major snRNA genes (U1, U2, U4, U5 and U6) exists in more than 10 copies in the Arabidopsis genome. U2 snRNA has the largest copy number, with a total of 18 putative homologs identified. Both U1 and U5 snRNAs have 14 copies, U6 snRNA has 13 copies, and U4 snRNA has only 11 copies. Sequence comparisons within Arabidopsis snRNA gene families showed that the U6 snRNA genes are the most similar, and the U1 snRNA genes are the most divergent. Eight active U6 snRNA copies are more than 93% identical to each other in the genic region, whereas active U1 snRNAs are on average only 87% identical. The U2 and U4 snRNAs are also highly conserved within each type, with more than 92% identity among the active genes. Details about the individual snRNAs and the respective sequence alignments are displayed at [27].

Previous studies identified two conserved transcription signals in most major snRNA gene promoters: USE (Upstream Sequence Element, RTCCACATCG) and TATA box [21-24]. All 14 U5 snRNAs have the USE and TATA box. Furthermore, their predicted secondary

structures are similar to the known structure of their counterparts in human, indicating that all these genes are active and functional (structure data not shown; for a review of the structures of human snRNAs, see [28]). Similarly, we identified 17 U2, 10 U1, nine U4, and nine U6 snRNA genes as likely active genes, with a few additional genes more likely to be pseudogenes because of various deletions. U4-10 and U6-7 do not have the conserved USE in the promoter region, but their U4-U6 interaction regions (stem I and stem II) are fairly well conserved. U2-16 is also missing the USE but has a secondary structure similar to other U2 snRNAs. These genes may be active, but differences in promoter motifs suggest that their expression may be under different control compared with other snRNAs homologs. The U2-17 snRNA has all conserved transcription signals, but 20 nucleotides are missing from its 3'-end. The predicted secondary structure of U2-17 is similar to that of other U2 snRNAs, with a significantly shorter stem-loop in the 3'-end due to the deletion. We are not sure if the U2-17 snRNA is functional, but the conserved transcription signals imply that it may be active.

Other conserved transcription signals were also identified in most active snRNAs, including the sequence element CAANTC in U2 snRNAs (located at -6 to -1) [20], and the termination signal $CAN_{3-10}AGTNNAA$ in Pol II transcribed U snRNAs (U1, U2, U4 and U5) [20, 21, 29]. The previously identified Monocot-Specific Promoter element (MSP, RGCCCR, located upstream of USE) in U6.1 and U6.26 [30] is also found in five other U6 snRNA genes (U6.29, U6-2, U6-3, U6-4, U6-5). In all seven U6 snRNAs the consensus MSP sequence extends by two thymine nucleotides to RGCCCRTT. Although the MSP does not contribute significantly to U6 snRNA transcription initiation in *N. plumbaginifolia* protoplasts [30], the extended consensus may implicate a role in gene expression regulation in Arabidopsis.

Low copy number of minor snRNA genes

The minor snRNAs are functional in the splicing of U12-type (AT-AC) introns. Four types of minor snRNAs, which correspond to four types of major snRNAs, exist in mammals. U11 is the analog to U1, U12 is the analog to U2, U4atac is the analog to U4, and U6atac is the analog to U6. The U5 snRNA seems to function in both the major and minor spliceosome [31]. Two minor snRNAs (atU12 and atU6atac) were experimentally identified in Arabidopsis [25]. Both have the conserved USE and TATA box in the promoter region. We identified another U6atac gene (atU6atac-2) by sequence mapping. This gene has a USE and a TATA box in the promoter region. The atU6atac-2 gene is more than 90% similar to atU6atac in both its 5'- and 3'-ends, with a 10-nucletotide deletion in the central region. The putative U4atac-U6atac interaction region in atU6atac-2 is 100% conserved with the interaction region previously identified in atU6atac [25, 32].

U11 and U4atac have not been experimentally identified in Arabidopsis. BLAST searches using the human U11 and U4atac homologs as queries against the Arabidopsis genome failed to find any significant hits, indicating divergence of the minor snRNAs in plants and mammals. Using the strategy described below, we successfully identified a putative Arabidopsis U4atac gene. It is a single copy gene containing all conserved functional domains. We also found a single candidate U11 snRNA gene (chromosome 5, from 17,492,101 to 17,492,600) that has the USE and TATA box in the promoter region. This

gene also contains a putative Sm protein binding site and a region that could pair with the 5' splice site of the U12 type intron.

Identification of an Arabidopsis U4atac snRNA gene

Like U4 snRNA and U6 snRNA, human U4atac and U6atac snRNAs interact with each other through base pairing [33]. The same interaction is expected to exist between the Arabidopsis homologs. Therefore, we deduced the tentative AtU4atac stem II sequence (CCCGTCTCTGTCAGAGGAG) from AtU6atac snRNA and searched for matching sequences in the Arabidopsis genome. Hit regions together with 500 bp upstream and 500 bp downstream flanks were retrieved and screened for transcription signals (USE and TATA box). One sequence was identified that contains both the USE and TATA box in appropriate position, as shown in Figure 1.

The tentative U4atac snRNA gene contains not only the stem II sequence, but also the stem I sequence that presumably basepairs with U6atac snRNA stem I. Furthermore, a highly conserved Sm protein binding region exists at the 3'-end. The predicted secondary structure is nearly identical to hsU4atac, with a relative longer single stranded region (not shown). With the highly conserved transcriptional signals, functional domains and secondary structure, this candidate gene is likely to be a real U4atac snRNA homolog. We named it AtU4atac and assigned At4g16065 as its tentative gene model because it is located between gene models At4g16060 and At4g16070 on chromosome 4.

Tandem arrays of snRNAs genes

Some snRNAs genes exist as small groups on the Arabidopsis chromosomes [6]. We identified 10 snRNA gene clusters: seven U1-U4 snRNA clusters, one U2-U5 snRNA cluster, and tandem duplication for both U2 snRNA (U2-10) and U5 snRNA (U5.1) (Figure 2). All seven Arabidopsis U1-U4 clusters have the U1 snRNA gene located upstream of the U4 snRNA gene, with a 180-300 nucleotide intergenic region. Five of the U1-U4 arrays are located on chromosome five (U1a/U4.1, U1-4/U4-5, U1-8/U4-7, U1-9/U4-8, and U1-13p/U4.3p), and the remaining two are on chromosome one (U1-10/U4-6 and U1-14p/U4-10). The U2-17 and U5-10 occur in tandem array on chromosome five, separated by less than 200 nucleotides.

Arabidopsis splicing related protein-coding genes

Most of the proteins involved in splicing in mammals and Drosophila are known [4, 34, 35]. In addition, recent proteomics studies revealed many novel proteins associated with human spliceosomes (reviewed in [17]). Using all these animal proteins as query sequences, we identified a total of 395 tentative homologs in Arabidopsis. Sequence similarity scores and comparison of gene structure and protein domain structure were used to assign the genes to families. Each gene was assigned a tentative name based on the name of its respective animal homolog. Different homologs within a gene family were labelled by appending an Arabic number (1, 2, etc.) to the name. Close family members with similar gene structure were indicated by adding -a, -b, and -c to the name. The 395 genes were classified into five

different categories according to the presumed function of their products. 91 encode snRNP (small nuclear ribonucleoprotein particle) proteins, 109 encode splicing factors, and 60 encode potential splicing regulators. Details of EST evidence, alternative splicing patterns, duplication sources, and domain structure of these genes are listed in Table 2. We also identified 84 Arabidopsis proteins corresponding to 54 human spliceosome associated proteins. The remaining 51 genes encode proteins with domains or sequences similar to known splicing factors, but without enough similarity to allow unambiguous classification. These two categories are not discussed in detail here, but information about these genes is available at our ASRG site [36].

The majority of snRNP proteins are conserved in Arabidopsis

There are five snRNPs (U1, U2, U4, U5, and U6) involved in the formation of the major spliceosome, corresponding to five snRNAs. Five snRNPs (U1 snRNP, U2 snRNP, U5 snRNP, U4/U6 snRNP and U4.U6/U5 tri-snRNP) have been isolated experimentally in yeast or human [37-42]. Each snRNP contains the snRNA, a group of core proteins, and some snRNP-specific proteins. Most of these proteins are conserved in Arabidopsis. All U snRNPs except U6 snRNP contain seven common core proteins bound to snRNAs. These core proteins all have a Sm domain and have been called Sm proteins. The U6 snRNP contains seven LSM proteins ('like Sm' proteins). Another LSM protein (LSM1) is not involved in binding snRNA (reviewed in [43]).

As shown in Table 2, all Sm and LSM proteins have homologs in Arabidopsis, and eight of them are duplicated. It is likely that these genes existed as a single copy in the ancestor of animals and plants, but duplicated within the plant lineage. Only one of the 24 genes (LSM5, At5g48870) has been characterized experimentally in Arabidopsis. The LSM5 gene was cloned from a mutant supersensitive to ABA (abscisic acid) and drought (SAD1, [44]). LSM5 is expressed at low levels in all tissues and its transcription is not altered by drought stress [44]. CDNA and EST evidence exist for all other core protein genes, indicating that all 24 genes are active.

There are 63 Arabidopsis proteins corresponding to the 35 snRNP-specific proteins used as queries in our genome mapping. Very few of them were characterized experimentally, including U1-70K, U1A, and a tandem duplication pair of SAP130 [45-47]. U1-70K was reported as a single-copy essential gene. Expression of U1-70K antisense transcript under the APETALA3 promoter suppressed the development of sepals and petals [48]. We identified an additional homolog of U1-70K (At2g43370) and named it U1-70K2. The U1-70K2 proteins showed 48% similarity to the U1-70K protein according to Blast2 results. Both genes retain the sixth intron in some transcripts, a situation which would produce truncated proteins [45]. Interestingly, we found that five of the 10 Arabidopsis U1 snRNP proteins, including the U1-70K coding genes, may undergo alternative splicing.

Several genes in U2, U5, U4/U6 and U4.U6/U5 snRNPs but none in U1 snRNP occur in more than three copies in the Arabidopsis genome. The atSAP114 family has five members,

including two that occur in tandem (atSAP114-1a and atSAP114-1b). Three members have EST/cDNA evidence (Table 2). Interestingly, the predicted atSAP114p (At4g15580) protein contains a RNase H domain at the N-terminal, which shares similarity to At5g06805, a gene annotated as encoding a non-LTR retroelement reverse transcriptase-like protein. It is likely that the atSAP114p gene is a pseudogene that originated by retroelement insertion. The tri-snRNP 65 KD subunit has three gene copies clustered on chromosome four. Both the U4/U6 90kD protein and the U4/U6 15.5 KD protein also have three gene copies, and the 116kD and 200kD subunits in U5 snRNP have four copies apiece.

The yeast U1 snRNP contains several specific proteins that are not present in mammalian U1 snRNPs [49]. As is true in mammals, Arabidopsis also lacks Prp42 homologs, a known component of U1 snRNP in yeast [50]. However, Arabidopsis has two copies of Prp39, which are similar to Prp42. Furthermore, atPrp39a can produce a shorter protein isoform with a novel N-terminal sequence by exon skipping. It is possible that the duplicates and alternative isoforms of plant U1 snRNP proteins are functional homologs of the yeast specific proteins.

Several proteins specific to the minor spliceosome are also conserved in Arabidopsis. The human 18S U11/U12 snRNP contains several proteins found in U2 snRNP as well as seven novel proteins [13]. Four of the seven U11/U12 specific proteins (U11/U12-35K, 25K, 65K and 31K) are conserved in Arabidopsis, while the remaining three (59K, 48K and 20K) have

no clear homologs. Interestingly, all four Arabidopsis genes are single copy in the genome, and three of them are apparently alternatively spliced (Table 2).

Splicing factors are slightly different in Arabidopsis than in other organisms

We divided the splicing factors into eight subgroups according to recent human spliceosome studies [12, 13, 15, 17]: (1) splice site selection proteins; (2) SR proteins; (3) 17S U2 associated proteins; (4) 35S U5 associated proteins; (5) proteins specific to the BΔU1 complex; (6) Exon Junction Complex (EJC) proteins; (7) Second step splicing factors and (8) Other known splicing factors. We focused our analysis on the first two subgroups because their functions in splicing are well established. A total of 109 proteins in Arabidopsis were identified corresponding to 67 human queries from all eight subgroups. Most of the proteins are conserved among eukaryotes, but some human proteins have no obvious homologs in the Arabidopsis genome, and some novel splicing factors appear to exist within the Arabidopsis genome. About 43% of genes encoding splicing factors are duplicated in the genome, whereas some proteins such as SF1/BBP (branchpoint binding protein, which facilitates U2 snRNP binding in fission yeast [51]) and cap-binding proteins (CBP20 and CBP80, possibly involved in cap proximal intron splicing [52]) derive from single copy genes [53]. These single-copy gene products may work with all pre-mRNAs, including the ones with U12-type introns. Surprisingly, mutation of CBP80 (ABH1) gene is not lethal and nonpleiotropic. The *abh1* plants show ABA-hypersensitive stomata closing and reduced wilting during drought [54].

Many splicing factors were identified previously in Arabidopsis, including two U2AF65, two U2AF35, and 18 SR proteins [55-64]. The U2AF35 related protein (atUrp), which could interact with U2AF65 and position RS-domain containing splicing factors [65], is also present in the Arabidopsis genome. Although the Urp gene is expressed ubiquitously in human tissues, no ESTs from this gene were found in Arabidopsis. Three copies of PTB/hnRNP-I genes were identified in Arabidopsis. The PTB competes for the poly-pyrimidine tract with the U2AF large subunit, thus negatively regulating splicing [66].

We also identified a homolog related to atU2AF^{65} (At2g33440) and an additional SR protein (At2g46610). The U2AF^{65} related protein (atULrp, At2g33440) has 247 amino acids and shares over 40% similarity with the C-terminal of the two atU2AF^{65} homologs. Compared with three RRMs and one N-terminal RS domain in atU2AF^{65} proteins, only one RRM can be identified, and there is no apparent RS domain in atULrp. No animal homolog of atULrp could be identified. The function of this one-RRM U2AF^{65} related protein is not clear. As it lacks other functional motifs, this protein might act as a competitor of U2AF65. A two-RRM U2AF^{65} protein can be produced through alternative splicing. The 11th intron of atU2AF^{65}a can be retained (see RAFL full-length cDNA gi: 19310596) to produce a truncated protein with only the first two RRMs. Interestingly, the last RRM in atU2AF^{65}a contains several amino acid variations from the consensus pattern such that it could not be detected by InterPro and NCBI-CDD searches using default values, also suggesting that perhaps only the first two RRMs are essential.

The additional SR protein belongs to the atRSp31 family and was named atRSp32 (At2g46610). It shares 70% identity and 78% similarity with atRSp31. The protein is 250 amino acids in length and contains two RRMs and some RS dipeptides in the C-terminus. The gene structure of atRSp32 is similar to that of atRSp31. Two other genes (atRSp40 and atRSp41) are in the same family and also have similar exon and intron sizes (see gene structure information at [67]). Similar to the previous classification of 18 SR proteins [58], the 19 SR proteins (including SR45) can be grouped into four large families of 4-5 members according to sequence similarity, gene structure, and protein domain structure. The atRSp31 family (atRSp31, atRSp32, atRSp40, atRSp41) belongs to a novel plant SR family and has no clear animal ortholog. Other families include the SC35 (or SRrp/TASR2) family, SF2/ASF family, and 9G8 family. Arabidopsis has a single copy of the SC35 ortholog and four SC35-like proteins (atSR33, atSCL30a, atSCL30, atSCL28), which appear to have diverged significantly from SC35. It seems that this divergence predates the split of plants and animals because a similar SC35-like gene family exists in the human genome (SRrp35 and SRrp40). The SRrp35 and SRrp40 were found to antagonize other SR proteins *in vitro* and function in 5' splice site selection [68]. SF2/ASF has four copies (atSR1/SRp34, atSRp30, atSRp34a and atSRp34b) with similar gene structures and domains. Human 9G8 protein has five homologs in Arabidopsis, with three (atRSZp21, atRSZp22 and atRSZp22a) containing one CCHC-type zinc finger and two (atRSZ33, atRSZ32) containing two CCHC-type zinc fingers in addition to an RRM and an RS domain. Interestingly, several SR proteins (atRSZp21, atRSZp22, SR45, and SCL33) were found to interact with atU1-70K, and some SR proteins can interact with each other, thus forming a complicated interaction network to facilitate splice site selection and spliceosome assembly [3, 58-60]. atSR45 was initially regarded as a

novel plant SR protein [60], but by virtue of sequence similarity scores it actually may be the ortholog of the human RNPS1 gene, which encodes an Exon Junction Complex (EJC) protein. Other human SR proteins (SRp20, SRp30c, SRp40, SRp54, SRp55, and SRp75) lack clear orthologs in Arabidopsis. We conclude that SR protein families evolved differently in animals and plants from three to four common ancestors including SC35, SF2/ASF and 9G8/RSZ. The SRrp (SC35-like in plants) family may not be classical SR proteins but they play important roles in splice site selection.

Proteins in other subgroups such as 17S U2 snRNP associated proteins, 35S U5 snRNP associated proteins, and protein specific to the BΔU1 complex are also conserved in Arabidopsis. The BΔU1 complex is the spliceosome complex captured right before catalytic activation. Most proteins in the 35S U5 snRNP are absent in the BΔU1 complex but present in active B complex, indicating the important roles of 35S U5 snRNP associated proteins in spliceosome activation [12]. Conservation of these proteins in Arabidopsis revealed the same pathway of spliceosome activation in plants. A subcomplex named Prp19 complex in 35S U5 snRNP plays critical roles during spliceosome activation [12, 69]. All proteins in the human Prp19 complex have homologs in Arabidopsis, including a chromosomal duplication pair of Prp19 genes and a single copy of the CDC5 gene. For the BΔU1 complex, six human genes have homologs, and five of them are single copy in Arabidopsis. Two genes (NPW38BP/SNP70 and p220 (NPAT)) in the human BΔU1 complex have no apparent Arabidopsis homologs.

Arabidopsis also lacks an SMN protein complex. In human, the SMN protein (survival of motor neurons) can interact with a series of proteins including Gemin2, Gemin3 (helicase), Gemin4, Gemin5, and Gemin6 to form an SMN complex, which plays important roles in the biogenesis of snRNPs and the assembly of the spliceosome through direct interactions with Sm proteins and snRNA [70]. Although the SMN protein exists in the fission yeast genome (GenBank Accession CAA91173), none of SMN complex members can be identified in the Arabidopsis genome.

Splicing regulators are expanded in Arabidopsis

The term splicing regulators refers to proteins that can either modify splicing factors or compete with splicing factors for their binding site. Important splicing regulators are hnRNP proteins and SR protein kinases. The exact role of phosphorylation of SR proteins in splicing is not yet clear, but SR protein kinases are well conserved and exist as multiple copies in Arabidopsis. A total of eight SR protein kinases were identified in Arabidopsis, including three Lammer/CLK kinases (AFC1, AFC2 and AFC3), two SRPK1 homologs, and three SPRK2 homologs. The three Lammer/CLK kinases were identified previously, and AFC2 was shown to phosphorylate SR protein *in vitro* [60, 71]. Over-expression of tobacco AFC2 homolog PK12 in Arabidopsis changed alternative splice patterns of several genes, including atSRp30, atSR1/atSRp34 and U1-70K [72], indicating that these SR protein may function to modulate splicing in plants.

The hnRNPs (heterogeneous nuclear ribonucleoproteins) bind to splice sites and to binding sites of splicing factors on nascent pre-mRNAs, thus competing with splicing factors to negatively control splicing (reviewed in [73]). Humans have about 20 hnRNP proteins, many of which function in splicing. A total of 35 potential hnRNP proteins possibly related to splicing were found in Arabidopsis by sequence similarity searches, including a superfamily of glycine-rich RNA binding proteins. This family contains 21 members similar to human hnRNP A1 and hnRNP A2/B1. It can be further divided into two subfamilies. One includes eight proteins containing one RRM, and another has 13 members with two RRMs. 12 of these proteins were identified previously, including AtGRP7, AtGRP8, UBA2a, UBA2b, UBA2c and AtRNPA/B1-6 [11, 74, 75]. AtGRP7 was found to be able to influence alternative splicing of its own transcripts as well as AtGRP8 transcripts [76]. UBA2 proteins can interact with UBP1 and UBA1 proteins, which have three RRMs and one RRM respectively, to recognize U-rich sequences in the 3'-UTR and stabilize mRNA [75]. Although the over-expression of UBA2 did not stimulate splicing of a reporter gene in tobacco protoplasts [75], we can not rule out the possibility that it could be involved in splicing of other genes.

Other human hnRNPs related to splicing also have homologs in Arabidopsis. BLAST searches of the human CUG-BP against all Arabidopsis proteins revealed three putative homologs, including atFCA. AtFCA and CUG-BP share similarity within the RRMs and a region approximately 40 amino acids in length. An additional protein (At2g47310) related to FCA was identified and named FCA2, as it shares about 50% similarity with FCA. The FCA proteins have two RRMs and a WW domain, which interact with the FY protein, a homolog

of yeast polyadenylation factor Psf2p [77, 78]. The FCA-FY complex negatively regulates the FCA protein by favoring a polyadenylation site from the third intron of FCA pre-mRNA [77, 79]. FCA may be a multi-functional protein involved in mRNA processing, as human CUG-BP can function in both alternative splicing and deadenylation [80]. We also list 15 previously identified hnRNP-like proteins and two additional homologs as possible splicing regulators. The UBP1 proteins can strongly enhance splicing of some introns in protoplasts [81], while UBA1, RBP45 and RBP47 proteins have no similar function [75, 82].

Unclassified splicing protein candidates

In addition to the 260 proteins in the above three categories, there are also 84 Arabidopsis proteins corresponding to human spliceosome-associated proteins identified in recent proteomic studies [14-17]. Some of these proteins function in other processes, such as transcription, polyadenylation and even translation. Their association with spliceosomes provides evidence for the coupling of splicing and other processes. Other proteins have no known functions. Only 35.8% of the proteins in this category are duplicated in Arabidopsis. We also identified a total of 51 Arabidopsis protein-coding genes similar to known splicing proteins. They have conserved domains and some level of sequence similarity to known splicing factors. We did not include these two categories in Table 2, but detailed information about them is available at ASRG [36].

Distribution and duplication of Arabidopsis splicing related genes

The distribution of Arabidopsis snRNA and splicing related proteins across the genome is shown in Figure 2 and at the ASRG website. Overall, the genes appear evenly distributed on the chromosomes, with several small gene clusters. Only four snRNA genes are located on chromosome two, three of which are U2 snRNA genes. No U4 snRNA gene is located on chromosome four. For the protein-coding genes, most functional categories have members located on each chromosome. The only exception is the SR protein kinase family, which has no member on chromosome one. Interestingly, chromosome one contains the most snRNP proteins and splicing factors, but has the fewest splicing regulators. Several gene clusters encoding splicing related proteins were also identified. Some clusters, such as tandemly duplicated gene pairs, include genes from the same category. One cluster located on chromosome four includes four genes encoding tri-snRNP proteins (atTri65a, atTri65b, atTri65c, and atTri15.5c, homologs of tri-snRNP 65-KD protein and 15.5KD protein). Two other clusters, atU2A-atCdc5 and atCUG-BP1-atU1C, include genes from different functional categories. No clear snRNA-splicing related protein clusters were identified. Although about one third of snRNA genes are located near other protein coding genes, none of their neighbouring genes is related to splicing. As a caveat, we should point out, that our snRNA gene determination strongly suggests annotation errors in overlapping protein-coding gene models. Thus, atU2-1, atU2.3, atU4.2, atU4-11p, atU5-13, and atU6.26 overlap gene models At1g16820, At3g57770, At3g06895, At1g68390, At5g53740, and At3g13857, respectively, but none of these models are well supported by cDNA or EST evidence (see displays linked at ASRG [27]).

The 260 proteins in the first three categories could be grouped into 130 families, 66 of which consist of multiple members. The duplication rate is over 50%, which is higher than the 44% duplication rate of Arabidopsis transcription factors [83]. As shown in Table 3, about 50% of genes encoding snRNP proteins, 43% of splicing factors, and 78% of splicing regulators have duplications. The much higher duplication rate of splicing regulators may reflect diversification in splicing control.

At least 130 duplication events are required to yield the 260 proteins from 130 families given one single-copy ancestor per family. Thirty-three duplication events (about a quarter of the total) are likely the result of chromosome duplications. The chromosomal duplication ratio is 18.9%-27.5% among the three groups (see Table 3). Some snRNA genes pairs, such as U2-14/U2-10, U5-3/U5-5 and (U6.1 U6.26) / (U6-8p U6-9p), may also have been produced by chromosome duplication. The C.D.2-3 region (chromosome duplication region between chromosomes two and three, see [84]) has the most splicing related gene pairs. Six genes in this region on chromosome two were duplicated in the same order on chromosome three. EST evidence shows that all these genes are expressed. Three U5 snRNA genes (U5.1, U5.1b, and U5-4) and four U2 snRNA genes (U2.2, U2.3, U2.4, and U2.6) also are located in the same region on chromosome three. No U5 and U2 homologs exist in the corresponding region on chromosome two, suggesting that the snRNA duplication events in that region may have happened after the chromosome duplication event, or that the snRNA duplicates were lost subsequent to the chromosome duplication events.

Chromosomal duplication rather than individual gene duplication appears to be the predominant mode of amplification for some types of genes. As shown in Table 2, the 24 genes encoding core proteins have nine duplication pairs, five of which can be attributed to chromosomal duplications. The 19 SR protein genes include eight duplication pairs, six of which are probably the results of chromosomal duplications. At least five chromosomal duplication events contributed to the super family of 21 hnRNP glycine-rich RBD and A/B genes. It is not clear why these functional categories have high chromosomal duplication ratios. It is possible that chromosomal duplication could create positive selection to maintain similar copy numbers of other genes encoding proteins that interact with the products of already duplicated genes.

Alternative splicing of Arabidopsis splicing related genes

According to EST/cDNA alignments, 80 of the 260 protein coding genes show 66 alternative splicing events. This rate (30.8%) is much higher than the overall frequency of alternative splicing in Arabidopsis, which is about 13% using the same criteria (2,747 genes out of 20,446 genes with EST/cDNA evidence; Brendel *et al.*, in preparation). As shown in Table 4, the snRNP protein coding genes have the lowest alternative splicing ratio (24.2%), whereas the ratios for splicing factor and splicing regulator genes are both over 33%. More than half of the genes encoding Exon Junction Complex proteins, proteins specific for the BΔU1 complex, SR proteins, U11/U12 snRNP specific proteins, and U1 snRNP proteins undergo alternative splicing.

Among different types of alternative splicing, intron retention is the most abundant of the alternative transcripts identified for the 260 classified splicing related genes. As shown in Table 4, 44 of the total 80 alternative splicing genes (about 55%) involve intron retention, 28 (35%) involve alternative acceptor site selection, and 15 (18.7%) are due to exon skipping. Compared with the corresponding ratio in all Arabidopsis alternative splicing events, (55.3% intron retention, 23.4% alternative acceptor site selection, and 6.3% exon skipping; Brendel *et al.*, in preparation), the ratio of intron retention in splicing related genes is similar and the ratio of exon skipping is higher. Interestingly, only one of the 20 splicing regulator genes processed by alternative splicing (about 5%) shows exon skipping, indicating that exon skipping is an important post-transcriptional method for controlling the expression of splicing factor coding genes, but not for the splicing regulator genes.

Discussion

Previous studies had determined 30 snRNA genes and 46 protein-coding genes related to splicing in Arabidopsis (see Tables 1 and 2). In this study, we have computationally identified an additional 44 snRNA genes (Table 1) and 349 protein-coding genes (Table 2) that also may be involved in splicing. Among the five types of U snRNAs, U6 is the most conserved and U1 is the least conserved. We identified seven U1-U4 snRNA gene clusters. We were surprised to see so many U1-U4 clusters in Arabidopsis. In Drosophila, four snRNA clusters were reported [4], but none of them include U1-U4 gene pairs. It is likely that a U1-U4 snRNA cluster existed in a progenitor of the current Arabidopsis genome, which was duplicated several times to form the extant seven clusters. The non-clustered U1

and U4 snRNA genes may have arisen by individual gene duplication or gene loss in duplicated clusters.

Among the proteins involved in splicing, most animal homologs are conserved in plants, indicating an ancient, monophyletic origin for the splicing mechanism. A striking feature of plant splicing related genes is their duplication ratio. 50% of the splicing genes are duplicated in Arabidopsis. The duplication ratio of the splicing related genes increases from genes encoding snRNP proteins to genes encoding splicing regulators. These data strongly suggest that the general splicing mechanism is conserved, but that the control of splicing may be more diverse in plants.

The high duplication ratio of Arabidopsis splicing related genes could be the result of evolutionary selection. Unlike animals, which can move around to maintain more homogeneous physiological conditions, plants are exposed to a larger range of stress conditions such as heat and cold. The duplicates will more likely be maintained in the genome as their functions become diversified, and potentially plant-specific, to ensure the fidelity of splicing under such varied conditions. Chromosome duplication produced several Sm proteins, SR proteins, and hnRNP proteins in Arabidopsis, which in turn could create positive selective pressures influencing the rate of duplication for functionally related genes. Because chromosome duplication occurred differentially within each plant lineage, we would expect different duplication patterns of these genes in, for example, monocots and dicots.

To confirm the above hypothesis, we searched the recently sequenced rice genome using the five Arabidopsis SC35 and SC35-like proteins as probes. Eight distinct genome loci were found to encode SC35 and SC35-like proteins, including three homologs of atSC35, two homologs of atSR33/SCL33 and atSCL30a, two homologs of atSCL30, and one homolog of atSCL28. Five of the eight rice genes are currently annotated in GenBank with accession numbers BAC79909 (osSC35a), BAD09319 (osSC35b), AAP46199 (osSR33-1), BAC799901 (osSCL30a/osSR33-2), and BAD19168 (osSCL30-1). As shown in the phylogenetic tree displayed in Figure 3, the two rice SC35 genes and atSC35 are likely to be orthologs of the animal SC35 gene. The other sequences cluster in SC35-like (SRrp/TASR) clades, indicating that the SC35 and SRrp/TASR genes diverged before the divergence of monocot and dicot plants (the divergence presumably happened even before the divergence of animals and plants, as described earlier). In addition, there are species-specific duplications. Thus, the Arabidopsis chromosomal duplication pair atSR33 – atSCL30a forms a clade, while their rice copies (osSR33-1 and osSCL30a) form another clade. Also there are additional duplications for the rice SC35 and SCL30 genes. We are currently working to identify all rice splicing related genes. The complete sets of these genes in two plant species should provide a good foundation for assessing similarities and differences in splicing mechanisms used by monocot and dicot plants.

As introns evolve rapidly, the mechanism to recognize and splice them should either evolve correspondingly or be flexible enough to accommodate the changes. It seems that plants deploy the most economic and practical way by keeping a largely conserved splicing mechanism and a very flexible recognition and control mechanism. Direct evidence comes

from the presence of plant-specific splicing proteins, such as the novel SR protein family and the super family of hnRNP A/B. The absence of SMN complex and some yeast U1 snRNP proteins in Arabidopsis indicates that other organisms also have integrated new proteins or pathways into the splicing mechanism over the course of evolution relative to other eukaryotes. Other evidence supporting the conserved splicing but flexible regulating mechanism include differential conservation among U snRNAs (U1 snRNAs is less conserved than U6 snRNAs) and high alternative splicing frequency in U1 snRNP proteins, SR proteins and hnRNP proteins. The SR proteins and U1 snRNP proteins are involved in early steps of splicing and 5'- and 3'-splice site selection; multiple isoforms of these proteins may be functionally significant in the control of splicing.

It is interesting to note that the overall alternative splicing frequency in splicing related genes is much higher than the frequency averaged over all Arabidopsis genes. More than half of SR proteins and U1 snRNP proteins show alternative splicing. Alternative splicing might increase protein diversity derived from splicing related genes, which would further add flexibility to the splicing mechanism. The high frequency of alternative transcripts from splicing related genes raises another interesting question – how is splicing regulated in these splicing related genes? One possible answer is that some splicing related genes may be auto-regulated. Accumulation of one transcript would feed back to inhibit/promote other isoforms. Several splicing related genes were reported previously to be regulated in this way. For example, AtGRP7 (hnRNP A/B super family) is a circadian clock regulated protein which auto-regulates its expression negatively [76]. When the AtGRP7 protein accumulates over the circadian cycle, it promotes the alternative transcript using a cryptic 5' splice site. The

alternative transcripts contain pre-mature stop codons and do not accumulate to high levels due to message instability, thus decreasing the level of AtGRP7 protein [76]. atSRp30 has similar effects on its own transcripts [62]. Another possible answer is that some splicing related genes could be able to regulate the splicing of other splicing related genes. For example, over-expression of AtGRP7 and atSRp30 is known to affect the splicing of AtGRP8 and atSR1, respectively [62, 76]. A third possibility is that the environment could affect the alternative splicing pattern. A good example is the SR1 gene. The ratio of two transcripts from the SR1 gene (SR1B/SR1) increases in a temperature-dependent manner [64]. Generally, heat or cold stress could cause intron retention in some splicing regulators, which could further alter the splicing pattern of other genes. The fourth possible regulators are intronless genes. Combining all these possibilities, a pathway to regulate splicing could be inferred as follows: (1) Environmental changes → (2) Splicing pattern changes in some specific splicing related genes and/or intronless genes → (3) Expression pattern changes (including splicing pattern changes) in general splicing related genes → (4) Changes in splicing patterns for specific genes.

Conclusion

A large number of Arabidopsis splicing related genes were computationally identified in this study by means of sequence comparisons and motif searches, including a tentative U4atac snRNA gene containing all conserved motifs, a new SR protein-coding gene (atRSp32) belonging to the atRSp31 family, and several genes related to genes encoding known splicing related proteins (atULrp and atFCA2). A web accessible database containing all the

Arabidopsis splicing related genes has been constructed and will be expanded to other organisms in the near future. This compilation should provide a good foundation to study the splicing process in more detail and to determine to what extent these genes are conserved across the entire plant kingdom. Our data show that about 50% of the splicing related genes are duplicated in Arabidopsis. The duplication ratios for splicing regulators are even higher, indicating that the splicing mechanism is generally conserved among plants, but that the regulation of splicing may be more variable and flexible, thus enabling plants to respond to their specific environments.

Materials and methods

Search for Arabidopsis snRNAs

Sequences of the 15 experimentally identified major snRNAs were downloaded from GenBank. The two minor snRNAs sequences were compiled from the literature[25]. These genes were used to search against the Arabidopsis genome at the AtGDB BLAST server [85] and at the SALK T-DNA Express web server [86]. Our initial analysis was based on Release 3.0 of the Arabidopsis genome (GenBank accession numbers NC_003070.4, NC_003071.3, NC_003074.4, NC_003075.3, and NC_003076.4). Local BLAST [18] was employed to derive the locations of the snRNA homologs from more recently sequenced regions of the genome. Criteria used for local BLAST were "e 1 -F F -W 7" (cut-off eval is 1, dust filter on, with a minimum word size of 7). Human and maize snRNAs were also included as query sequences, and all hits with e-values less than 10^{-5} were regarded as possible homologs. A total of 70 major snRNAs and three minor snRNAs were identified by this method. Each

major snRNA type has 10-18 copies in the genome. A tentative gene name and gene model was assigned to each snRNA gene after comparison with the snRNAs identified in MATDB [87]. Sequence similarity values were based on BLAST alignments.

Search for Arabidopsis splicing related proteins

A three-round BLAST search strategy was used to identify Arabidopsis splicing related protein-coding genes. (1) Splicing related proteins from human and Drosophila were downloaded from GenBank according to several recent proteomic studies [14-17] and the web site compilation of the S. Mount group available at [88]. Human hnRNP proteins identified in a recent review [73] were downloaded from GenBank. All these sequences were used as queries in a local BLAST search against Arabidopsis annotated proteins (obtained from TIGR at [89]). All hits with an e-value less than 10^{-10} were collected as candidates. Many of these candidates had highly significant e-values (usually 10^{-30} or below and much lower than other hits). These candidates were regarded as true homologs. (2) All identified true homologs were used to query the Arabidopsis protein set again. An e-value of 10^{-20} was used as a cut-off value to find possible paralogs of the true homologs. Sequences identified in both rounds of BLAST hits were regarded as main candidates for splicing related proteins. (3) Finally, the main candidates were queried against GenPept and all annotated human proteins (obtained from Ensembl). All candidates with significant similarity to proteins unrelated to splicing were removed from the main candidate list, and all candidates with significant similarity to proteins related to splicing were regarded as true splicing related genes and were promoted to the status of true homologs. The remaining candidates were

regarded as unclassified splicing related proteins. BLAST results were initially analyzed by MuSeqBox (Multiple Sequence Blast Output eXamination; [90]). Two custom scripts were written to read MuSeqBox output files, largely automating the search procedure.

Gene structure and chromosomal locations

The gene structure and chromosomal locations for the genes encoding splicing related proteins were retrieved from Arabidopsis genome database AtGDB [91]. The chromosomal locations of the snRNA genes were inferred from the BLAST results. The location maps (Fig. 1) were generated using the AtGDB advanced search function [92]. Spliced alignments of ESTs and cDNAs generated by GeneSeqer [93] were used to verify gene models. Gene structure information was used as important criteria to group homologs into gene families.

Protein domains

InterProScan 3.3 was downloaded from [94] and subsequently was used to search protein domain databases using default parameters [95]. A Perl script was written to process the text results from InterProScan. Protein domain information was used in comparisons of homologs from different species. NCBI-CDD (Conserved Domain Database) search [96] was conducted manually for certain genes to confirm the InterPro results.

Duplication source

The gene families with multiple copies were inspected to determine whether they were likely to have derived from chromosome duplication events. Gene models of the duplicated gene were searched against the gene list of each chromosome redundancy region at MATDB [97]. If the gene and its duplicate were both in the list, they were regarded as a chromosome duplication pair. Otherwise, they were assumed to be produced by random gene duplication.

Identification of alternative splicing

All Arabidopsis EST and cDNAs were aligned against the genome using the spliced alignment program GeneSeqer as made available through AtGDB [98]. We retrieved the intron and exon coordinates of the reliable cognate alignments from the database. Scripts were written to identify introns that overlap with other introns or exons. We defined the alternative splicing cases as follows: (1) Alternative Donor (AltD): an intron has the same 3'-end coordinate but different 5'-end coordinate as another overlapping intron; (2) Alternative Acceptor (AltA): an intron has the same 5'-end coordinate but different 3'-end coordinate as another intron; (3) Alternative Position (AltP): an intron has different 5'-end and 3'-end coordinates as another overlapping intron; (4) Exon Skipping (ExonS): an annotated intron completely contains an alternatively identified exon in the same transcription direction; (5) Intron Retention (IntronR): an annotated intron is completely contained by an alternatively identified exon.

Database and interface construction

Details about each splicing related gene were saved in a MySQL database. PHP scripts were written to interact with the database and generate the interface web pages. Text and BLAST searches were implemented by Perl-cgi scripts.

List of abbreviations

snRNA: small nuclear RNA

snRNP: small nuclear ribonucleoprotein particle

hnRNP: heterogeneous nuclear ribonucleoproteins

USE: Upstream Sequence Element

MSP: Monocot-Specific Promoter

RRM: RNA Recognition Motif

C.D.: Chromosomal Duplication

Acknowledgements

We would like to thank Shannon Schlueter for help with the web page and database design and implementation. We are also grateful to Shailesh Lal, Carolyn Lawrence, and Michael Sparks for discussions and critical reading of the manuscript and to the anonymous reviewers for excellent suggestions. This work was supported in part by a grant from the ISU Plant Sciences Institute and NSF grants DBI-0110189 and DBI-0110254 to V.B.

References

1. Kazan K: **Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged**. *Trends Plant Sci* 2003, **8:**468-471.
2. Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W: **Pre-mRNA splicing in higher plants**. *Trends Plant Sci* 2000, **5:**160-167.
3. Reddy ASN: **Nuclear pre-mRNA splicing in plants**. *Critical Rev Plant Sci* 2001, **20:**523-571.
4. Mount SM, Salz HK: **Pre-messenger RNA processing factors in the Drosophila genome**. *J Cell Biol* 2000, **150:**F37-F43.
5. Käufer NF, Potashkin J: **Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals**. *Nucleic Acids Res* 2000, **28:**3003-3010.
6. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408:**796-815.
7. Barakat A, Szick-Miranda K, Chang IF, Guyot R, Blanc G, Cooke R, Delseny M, Bailey-Serres J: **The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome**. *Plant Physiol* 2001, **127:**398-415.
8. Beisson F, Koo AJ, Ruuska S, Schwender J, Pollard M, Thelen JJ, Paddock T, Salas JJ, Savage L, Milcamps A, *et al.*: **Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database**. *Plant Physiol* 2003, **132:**681-697.

9. Wang D, Harper JF, Gribskov M: **Systematic trans-genomic comparison of protein kinases between Arabidopsis and Saccharomyces cerevisiae.** *Plant Physiol* 2003, **132:**2152-2165.

10. Aubourg S, Kreis M, Lecharny A: **The DEAD box RNA helicase family in Arabidopsis thaliana.** *Nucleic Acids Res* 1999, **27:**628-636.

11. Lorkovic ZJ, Barta A: **Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana.** *Nucleic Acids Res* 2002, **30:**623-635.

12. Makarova OV, Makarov EM, Urlaub H, Will CL, Gentzel M, Wilm M, Lührmann R: **A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing.** *Embo J* 2004, **23:**2381-2391.

13. Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Lührmann R: **The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome.** *RNA* 2004, **10:**929-941.

14. Zhou Z, Sim J, Griffith J, Reed R: **Purification and electron microscopic visualization of functional human spliceosomes.** *Proc Natl Acad Sci U S A* 2002, **99:**12203-12207.

15. Makarov EM, Makarova OV, Urlaub H, Gentzel M, Will CL, Wilm M, Lührmann R: **Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome.** *Science* 2002, **298:**2205-2208.

16. Rappsilber J, Ryder U, Lamond AI, Mann M: **Large-scale proteomic analysis of the human spliceosome.** *Genome Res* 2002, **12:**1231-1245.

17. Jurica MS, Moore MJ: **Pre-mRNA splicing: awash in a sea of proteins**. *Mol Cell* 2003, **12:**5-14.

18. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25:**3389-3402.

19. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22:**4673-4680.

20. Vankan P, Filipowicz W: **Structure of U2 snRNA genes of Arabidopsis thaliana and their expression in electroporated plant protoplasts.** *Embo J* 1988, **7:**791-799.

21. Vankan P, Edoh D, Filipowicz W: **Structure and expression of the U5 snRNA gene of Arabidopsis thaliana. Conserved upstream sequence elements in plant U-RNA genes**. *Nucleic Acids Res* 1988, **16:**10425-10440.

22. Vankan P, Filipowicz W: **A U-snRNA gene-specific upstream element and a -30 'TATA box' are required for transcription of the U2 snRNA gene of Arabidopsis thaliana**. *Embo J* 1989, **8:**3875-3882.

23. Waibel F, Filipowicz W: **U6 snRNA genes of Arabidopsis are transcribed by RNA polymerase III but contain the same two upstream promoter elements as RNA polymerase II-transcribed U-snRNA genes**. *Nucleic Acids Res* 1990, **18:**3451-3458.

24. Hofmann CJ, Marshallsay C, Waibel F, Filipowicz W: **Characterization of the genes encoding U4 small nuclear RNAs in Arabidopsis thaliana**. *Mol Biol Rep* 1992, **17:**21-28.

25. Shukla GC, Padgett RA: **Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants**. *RNA* 1999, **5:**525-538.

26. Marker C, Zemann A, Terhorst T, Kiefmann M, Kastenmayer JP, Green P, Bachellerie JP, Brosius J, Huttenhofer A: **Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant Arabidopsis thaliana**. *Curr Biol* 2002, **12:**2002-2013.

27. **ASRG snRNAs** [http://www.plantgdb.org/SRGD/ASRG/AtsnRNA.php]

28. Patel AA, Steitz JA: **Splicing double: insights from the second spliceosome**. *Nat Rev Mol Cell Biol* 2003, **4:**960-970.

29. Connelly S, Filipowicz W: **Activity of chimeric U small nuclear RNA (snRNA)/mRNA genes in transfected protoplasts of Nicotiana plumbaginifolia: U snRNA 3'-end formation and transcription initiation can occur independently in plants**. *Mol Cell Biol* 1993, **13:**6403-6415.

30. Connelly S, Marshallsay C, Leader D, Brown JW, Filipowicz W: **Small nuclear RNA genes transcribed by either RNA polymerase II or RNA polymerase III in monocot plants share three promoter elements and use a strategy to regulate gene expression different from that used by their dicot plant counterparts**. *Mol Cell Biol* 1994, **14:**5910-5919.

31. Tarn WY, Steitz JA: **Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge**. *Trends Biochem Sci* 1997, **22:**132-137.

32. Shukla GC, Padgett RA: **U4 small nuclear RNA can function in both the major and minor spliceosomes**. *Proc Natl Acad Sci U S A* 2004, **101**:93-98.

33. Shukla GC, Cole AJ, Dietrich RC, Padgett RA: **Domains of human U4atac snRNA required for U12-dependent splicing in vivo**. *Nucleic Acids Res* 2002, **30**:4650-4657.

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

35. Will CL, Lührmann R: **Protein functions in pre-mRNA splicing**. *Curr Opin Cell Biol* 1997, **9**:320-328.

36. **ASRG proteins** [http://www.plantgdb.org/SRGD/ASRG/ASRP-home.php]

37. Stevens SW, Abelson J: **Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins**. *Proc Natl Acad Sci U S A* 1999, **96**:7226-7231.

38. Stevens SW, Barta I, Ge HY, Moore RE, Young MK, Lee TD, Abelson J: **Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from Saccharomyces cerevisiae**. *RNA* 2001, **7**:1543-1553.

39. Gottschalk A, Neubauer G, Banroques J, Mann M, Lührmann R, Fabrizio P: **Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6.U5] tri-snRNP**. *Embo J* 1999, **18**:4535-4548.

40. Caspary F, Shevchenko A, Wilm M, Seraphin B: **Partial purification of the yeast U2 snRNP reveals a novel yeast pre-mRNA splicing factor required for pre-spliceosome assembly**. *Embo J* 1999, **18**:3463-3474.

41. Krämer A, Grüter P, Gröning K, Kastner B: **Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP**. *J Cell Biol* 1999, **145:**1355-1368.

42. Fabrizio P, Esser S, Kastner B, Lührmann R: **Isolation of S. cerevisiae snRNPs: comparison of U1 and U4/U6.U5 to their human counterparts**. *Science* 1994, **264:**261-265.

43. Will CL, Lührmann R: **Spliceosomal UsnRNP biogenesis, structure and function**. *Curr Opin Cell Biol* 2001, **13:**290-301.

44. Xiong L, Gong Z, Rock CD, Subramanian S, Guo Y, Xu W, Galbraith D, Zhu JK: **Modulation of abscisic acid signal transduction and biosynthesis by an Sm-like protein in Arabidopsis**. *Dev Cell* 2001, **1:**771-781.

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

46. Simpson GG, Clark GP, Rothnie HM, Boelens W, van Venrooij W, Brown JW: **Molecular characterization of the spliceosomal proteins U1A and U2B" from higher plants**. *Embo J* 1995, **14:**4540-4550.

47. Casacuberta E., Puigdomenech P., Monofort A.: **A genomic duplication in Arabidopsis thaliana contains a sequence similar to the human gene coding for SAP130.** *Plant Physiol Biochem* 2001, **39:**565-573.

48. Golovkin M, Reddy AS: **Expression of U1 small nuclear ribonucleoprotein 70K antisense transcript using APETALA3 promoter suppresses the development of sepals and petals**. *Plant Physiol* 2003, **132:**1884-1891.

49. Gottschalk A, Tang J, Puig O, Salgado J, Neubauer G, Colot HV, Mann M, Seraphin B, Rosbash M, Lührmann R, *et al.*: **A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins**. *RNA* 1998, **4:**374-393.

50. McLean MR, Rymond BC: **Yeast pre-mRNA splicing requires a pair of U1 snRNP-associated tetratricopeptide repeat proteins**. *Mol Cell Biol* 1998, **18:**353-360.

51. Huang T, Vilardell J, Query CC: **Pre-spliceosome formation in S.pombe requires a stable complex of SF1-U2AF(59)-U2AF(23)**. *Embo J* 2002, **21:**5516-5526.

52. Lewis JD, Gorlich D, Mattaj IW: **A yeast cap binding protein complex (yCBC) acts at an early step in pre-mRNA splicing**. *Nucleic Acids Res* 1996, **24:**3332-3336.

53. Kmieciak M, Simpson CG, Lewandowska D, Brown JW, Jarmolowski A: **Cloning and characterization of two subunits of Arabidopsis thaliana nuclear cap-binding complex**. *Gene* 2002, **283:**171-183.

54. Hugouvieux V, Kwak JM, Schroeder JI: **An mRNA cap binding protein, ABH1, modulates early abscisic acid signal transduction in Arabidopsis**. *Cell* 2001, **106:**477-487.

55. Domon C, Lorkovic ZJ, Valcarcel J, Filipowicz W: **Multiple forms of the U2 small nuclear ribonucleoprotein auxiliary factor U2AF subunits expressed in higher plants**. *J Biol Chem* 1998, **273:**34603-34610.

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

57. Lopato S, Mayeda A, Krainer AR, Barta A: **Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors.** *Proc Natl Acad Sci U S A* 1996, **93:**3074-3079.

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

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

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

61. Lazar G, Schaal T, Maniatis T, Goodman HM: **Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF.** *Proc Natl Acad Sci U S A* 1995, **92:**7672-7676.

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

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

64. Lazar G, Goodman HM: **The Arabidopsis splicing factor SR1 is regulated by alternative splicing**. *Plant Mol Biol* 2000, **42:**571-581.

65. Tronchere H, Wang J, Fu XD: **A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA**. *Nature* 1997, **388:**397-400.

66. Lin CH, Patton JG: **Regulation of alternative 3' splice site selection by constitutive splicing factors**. *RNA* 1995, **1:**234-245.

67. **ASRG SR protein gene structure** [http://www.plantgdb.org/SRGD/ASRG/Display.php?GID=2.2&Gst=1]

68. Cowper AE, Caceres JF, Mayeda A, Screaton GR: **Serine-arginine (SR) protein-like factors that antagonize authentic SR proteins and regulate alternative splicing**. *J Biol Chem* 2001, **276:**48908-48914.

69. Chan SP, Kao DI, Tsai WY, Cheng SC: **The Prp19p-associated complex in spliceosome activation**. *Science* 2003, **302:**279-282.

70. Yong J, Pellizzoni L, Dreyfuss G: **Sequence-specific interaction of U1 snRNA with the SMN complex**. *Embo J* 2002, **21:**1188-1196.

71. Bender J, Fink GR: **AFC1, a LAMMER kinase from Arabidopsis thaliana, activates STE12-dependent processes in yeast**. *Proc Natl Acad Sci U S A* 1994, **91:**12105-12109.

72. Savaldi-Goldstein S, Aviv D, Davydov O, Fluhr R: **Alternative splicing modulation by a LAMMER kinase impinges on developmental and transcriptome expression**. *Plant Cell* 2003, **15:**926-938.

73. Krecic AM, Swanson MS: **hnRNP complexes: composition, structure, and function**. *Curr Opin Cell Biol* 1999, **11**:363-371.

74. Heintzen C, Melzer S, Fischer R, Kappeler S, Apel K, Staiger D: **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* 1994, **5**:799-813.

75. Lambermon MH, Fu Y, Wieczorek Kirk DA, Dupasquier M, Filipowicz W, Lorkovic ZJ: **UBA1 and UBA2, two proteins that interact with UBP1, a multifunctional effector of pre-mRNA maturation in plants**. *Mol Cell Biol* 2002, **22**:4346-4357.

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

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

78. Macknight R, Bancroft I, Page T, Lister C, Schmidt R, Love K, Westphal L, Murphy G, Sherson S, Cobbett C, *et al.*: **FCA, a gene controlling flowering time in Arabidopsis, encodes a protein containing RNA-binding domains**. *Cell* 1997, **89**:737-745.

79. Quesada V, Macknight R, Dean C, Simpson GG: **Autoregulation of FCA pre-mRNA processing controls Arabidopsis flowering time**. *Embo J* 2003, **22**:3142-3152.

80. Paillard L, Legagneux V, Osborne HB: **A functional deadenylation assay identifies human CUG-BP as a deadenylation factor**. *Biol Cell* 2003, **95:**107-113.

81. Lambermon MH, Simpson GG, Wieczorek Kirk DA, Hemmings-Miészczak M, Klahre U, Filipowicz W: **UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation**. *Embo J* 2000, **19:**1638-1649.

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

83. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, *et al.*: **Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes**. *Science* 2000, **290:**2105-2110.

84. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in Arabidopsis**. *Science* 2000, **290:**2114-2117.

85. **AtGDB BLAST** [http://www.plantgdb.org/cgi-bin/PlantGDB/AtGDB/BRview.pl]

86. **T-DNA express** [http://signal.salk.edu/cgi-bin/tdnaexpress]

87. **MATDB snRNAs** [http://mips.gsf.de/cgi-bin/proj/thal/search_type?all/185]

88. **Drosophila mRNA processing factors** [http://www.life.umd.edu/labs/Mount/factors/]

89. **TIGR ftp site** [ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep]

90. Xing L, Brendel V: **Multi-query sequence BLAST output examination with MuSeqBox**. *Bioinformatics* 2001, **17:**744-745.

91. **AtGDB** [http://www.plantgdb.org/AtGDB/]

92. **AtGDB advanced search** [http://www.plantgdb.org/AtGDB-cgi/search.pl]

93. Brendel V, Xing L, Zhu W: **Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus**. *Bioinformatics* 2004, **20:**1157-1169.

94. **InterPro** [http://www.ebi.ac.uk/interpro/]

95. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17:**847-848.

96. **NCBI-CDD search** [http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml]

97. **MATDB Redundancy Viewer** [http://mips.gsf.de/proj/thal/db/gv/rv/rv_frame.html]

98. Zhu W, Schlueter SD, Brendel V: **Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping**. *Plant Physiol* 2003, **132:**469-484.

99. **PHYLIP** [http://evolution.genetics.washington.edu/phylip.html]

100. Hirayama T, Shinozaki K: **A cdc5+ homolog of a higher plant, Arabidopsis thaliana**. *Proc Natl Acad Sci U S A* 1996, **93:**13371-13376.

101. Landsberger M, Lorkovic ZJ, Oelmuller R: **Molecular characterization of nucleus-localized RNA-binding proteins from higher plants**. *Plant Mol Biol* 2002, **48:**413-421.

Figure Legends

Figure 1. Sequence alignment of U4atac and U6atac snRNAs.

The tentative Arabidopsis U4atac snRNA was aligned against the human U4atac snRNA (U62822) using CLUSTALW [19]. Possible sequence domains are indicated by different background colors, with cyan indicating transcription signals (USE: Upstream Sequence Element; TATA: TATA box), green indicating the region involved in the stem-loop-stem structure, and pink indicating the Sm protein binding domain. The corresponding interaction region in U6atac snRNA is also marked in green. Red background indicates G-T base pairs in the stem-loop structure.

Figure 2. Chromosomal locations of Arabidopsis snRNAs.

Chromosomes one to five are represented to scale by the long thick lines in dark green in ascending numeric order from top to bottom. The small bars above the chromosome lines indicate the presence of a snRNA gene in that region. Different colors represent different snRNA types. Red: U1 snRNA; Magenta: U2 snRNA; Blue: U4 snRNA; Green: U5 snRNA; Peru: U6; Black: minor snRNA. The seven U1-U4 snRNA gene clusters (red-blue) and the single U2-U5 snRNA gene cluster (magenta-green) are indicated by red circles.

Figure 3. Phylogenetic tree of the SC35 protein family.

The phylogenetic tree was constructed based on protein sequence alignments of the SC35 homologs in human, Drosophila, Arabidopsis and rice. The GenBank accession numbers for the sequences are as follows: hsSC35, Q01130; hsSRrp40, AAL57514; hsSRrp35: AAL57515; dmSC35, AAF53192; atSC35, NP_851261; atSR33/SCL33, NP_564685;

atSCL30a, NP_187966; atSCL30, NP_567021; atSCL28, NP_197382; osSC35a, BAC79909; osSC35b, BAD09319; osSR33-1, AAP46199; osSCL30a/SR33-2, BAC799901; osSCL30-2, BAD19168. The sequences were aligned using CLUSTALW [19] with default parameters, and the phylogenetic tree was produced according to the neighbour-joining method using PAM substitution model distances as implemented in the PHYLIP package [99].

Tables

Table 1. Arabidopsis snRNA genes

Chromosomal locations were determined by conducting BLAST searches against the Arabidopsis genome (Release 5.0). C in table head stands for Chromosome and S for strand. The gene used for query in the BLAST search is marked by an asterisk (*), the dollar sign ($) marks atU12 and atU6atac sequences, which were experimentally identified [25] but had no GenBank accession number. Their sequences were compiled manually from the cited paper. The GenBank gi numbers for the chromosome sequences used are as follows: chromosome 1, 42592260; chromosome 2, 30698031; chromosome 3, 30698537; chromosome 4, 30698542; chromosome 5, 30698605.

Gene	GeneID	C	S	From	To	Len	eVal	Similarity	GenBank
atU1a*	At5g49054	5	-	19903323	19903158	166	1E-89	1 - 166, 100%	gi17660
atU1-2	At4g23415	4	+	12225621	12225786	166	1E-58	1 - 166, 92%	gi22293582
atU1-3	At5g51675	5	+	21013986	21014149	164	4E-55	3 - 166, 91%	
atU1-4	At5g25774	5	-	8972971	8972807	165	2E-51	1 - 166, 90%	gi22293583
atU1-5	At1g08115	1	-	2538238	2538073	166	1E-46	1 - 166, 89%	gi22293581
atU1-6	At3g05695	3	+	1681815	1681977	163	4E-40	4 - 166, 87%	
atU1-7	At3g05672	3	+	1657766	1657928	163	4E-40	4 - 166, 87%	gi22293580
atU1-8	At5g27764	5	+	9832576	9832740	165	1E-39	1 - 166, 87%	
atU1-9	At5g26694	5	-	9494594	9494430	165	1E-27	1 - 166, 84%	
atU1-10	At1g11884	1	-	4007396	4007236	161	1E-18	4 - 61, 93%; 80 - 166, 88%	
atU1-11p	At4g16645	4	+	9370786	9370841	56	7E-17	4 - 59, 94%	

atU1-12p	At4g23565	4	-	12298871	12298802	70	1E-15	94 - 163, 90%	
atU1-13p	At5g49524	5	-	20112431	20112275	157	2E-14	4 - 50, 91%; 91 - 166, 88%	
atU1-14p	At1g35354	1	+	12986822	12986908	87	1E-06	10 - 60, 88%; 84 - 118, 88%	
atU2-1	At1g16825	1	+	5758381	5758575	195	2E-88	1 - 196, 96%	
atU2.2*	At3g57645	3	+	21357718	21357913	196	1E-107	1 - 196, 100%	gi17661
atU2.3	At3g57765	3	-	21408595	21408400	196	1E-95	1 - 196, 97%	gi17662
atU2.4	At3g56825	3	-	21052994	21052800	195	5E-86	1 - 196, 95%	gi17663
atU2.5	At5g09585	5	+	2975013	2975208	196	7E-79	1 - 196, 93%	gi17664
atU2.6	At3g56705	3	+	21015472	21015667	196	1E-83	1 - 196, 94%	gi17665
atU2.7	At5g61455	5	-	24730829	24730634	196	5E-86	1 - 196, 95%	gi17666
atU2-8	At5g67555	5	+	26966884	26967079	196	5E-86	1 - 196, 95%	
atU2.9	At4g01885	4	+	815273	815466	194	2E-82	1 - 194, 94%	gi17667
atU2-10	At2g02938	2	+	849777	849972	196	3E-93	1 - 196, 96%	gi22293586
atU2-10b/12	At2g02940	2	+	852859	853054	196	3E-93	1 - 196, 96%	
atU2-11	At1g09805/09895	1	-	3180736	3180547	190	8E-85	1 - 190, 95%	
atU2-13	At2g20405	2	+	8809169	8809364	196	3E-81	1 - 196, 94%	gi22293584
atU2-14	At1g14165	1	+	4842274	4842469	196	3E-81	1 - 196, 94%	gi22293585
atU2-15	At5g62415	5	+	25083790	25083985	196	4E-74	1 - 196, 92%	
atU2-16	At5g57835	5	-	23448717	23448522	196	2E-67	1 - 196, 92%	
atU2-17	At5g14545	5	-	4690105	4690008	98	3E-44	1 - 98, 97%	
atU2-18p	At3g26815	3	+	9881236	9881303	68	2E-14	1 - 68, 89%	
atU4.1*	At5g49056	5	-	19902970	19902817	154	4E-80	1 - 154, 99%	gi17673
atU4.2	At3g06900	3	-	2178343	2178190	154	2E-75	1 - 154, 98%	gi17674
atU4.3p	At5g49526	5	-	20112072	20112030	43	2E-11	15 - 57, 95%	gi17675
atU4-4	At1g49242/49235	1	-	18222354	18222201	154	2E-75	1 - 154, 98%	gi22293588
atU4-5	At5g25776	5	-	8972618	8972465	154	1E-70	1 - 154, 96%	
atU4-6	At1g11886	1	-	4007020	4006867	154	1E-70	1 - 154, 96%	gi22293587
atU4-7	At5g27766	5	+	9832934	9833083	150	7E-66	1 - 150, 96%	
atU4-8	At5g26996	5	-	9494230	9494081	150	7E-66	1 - 150, 96%	
atU4-9	At1g79965	1	+	30086031	30086168	138	9E-47	18 - 154, 92%	
atU4-10	At1g35356	1	+	12987189	12987313	125	3E-34	1 - 124, 90%	
atU4-11p	At1g68395	1	+	25647322	25647396	75	9E-07	18 - 37, 100%; 60 - 102, 90%	
atU5.1*	At3g55865	3	-	20740607	20740503	105	6E-35	1 - 105, 94%	gi17676
atU5.1b	At3g55855	3	-	20736881	20736780	102	7E-38	1 - 102, 96%	gi22293592
atU5-2	At1g65115	1	+	24194482	24194586	105	1E-39	1 - 105, 96%	
atU5-3	At1g70185	1	+	26433396	26433497	102	7E-38	1 - 102, 96%	gi22293590
atU5-4	At3g55645	3	+	20653843	20653947	105	3E-37	1 - 105, 95%	
atU5-5	At1g24105/24095	1	-	8525204	8525103	102	2E-35	1 - 102, 95%	gi22293591
atU5-6	At1g04475	1	-	1215831	1215730	102	2E-35	1 - 102, 95%	gi22293589
atU5-7	At4g02535	4	-	1114629	1114528	102	1E-30	2 - 103, 93%	
atU5-8	At3g25445	3	-	9227212	9227116	97	1E-20	5 - 101, 89%	
atU5-9	At1g79545	1	-	29928543	29928447	97	1E-20	5 - 101, 89%	
atU5-10	At5g14547	5	-	4690412	4690370	43	3E-12	24 - 67, 97%	
atU5-11	At5g54065	5	-	21957066	21957023	44	2E-10	20 - 64, 95%	
atU5-12	At1g71355	1	+	26895255	26895298	44	2E-10	20 - 64, 95%	

atU5-13	At5g53745	5	-	21829988	21829943	46	3E-09	24 - 70, 93%	
atU6.1*	At3g14735	3	+	4951596	4951697	102	1E-51	1 - 102, 100%	gi16516
atU6.26	At3g13855	3	+	4561111	4561212	102	2E-49	1 - 102, 99%	gi16517
atU6.29	At5g46315	5	+	18804616	18804717	102	2E-49	1 - 102, 99%	gi16518
atU6-2	At5g62995	5	+	25296825	25296926	102	1E-51	1 - 102, 100%	
atU6-3	At4g27595	4	+	13782215	13782316	102	1E-51	1 - 102, 100%	
atU6-4	At4g03375	4	-	1483121	1483020	102	1E-51	1 - 102, 100%	
atU6-5	At4g33085	4	-	15965258	15965158	101	8E-37	1 - 101, 94%	
atU6-6	At4g35225	4	+	16754836	16754931	96	1E-32	1 - 102, 93%	
atU6-7	At2g15532	2	+	6784793	6784869	77	7E-25	4 - 80, 93%	
atU6-8p	At1g52605	1	+	19596398	19596476	96	2E-19	4 - 99, 87%	
atU6-9p	At1g53465	1	-	19960538	19960485	54	9E-09	21 - 74, 88%	
atU6-10p	At3g45705	3	+	16792802	16792888	87	2E-06	1 - 46, 89%; 62 - 100, 89%	
atU6-11p	At5g11085	5	-	3522167	3522143	25	9E-06	1 - 25, 100%	
atU12*	At1g61275	1	+	22606785	22606960	176	1E-95	1 - 176, 100%	$gi22293600
atU6atac*	At5g40395	5	-	16183534	16183413	122	1E-63	1 - 122, 100%	$
atU6atac-2	At1g21395	1	-	7491489	7491378	112	5E-20	1 - 65, 95%; 81 - 110, 93%	
atU4atac	At4g16065	4	+	9096374	9096532	159	N/A	N/A	

Table 2. Arabidopsis splicing related proteins

Gene names were kept consistent with names used in previous publications or derived from the names of the respective homologs (yeast names are given in the S.c. column, where available). The Tnb column gives the numbers of cognate cDNAs and ESTs supporting the gene structure. The AltS column indicates evidence for alternative splicing, including Alternative Donor site (AltD), Alternative Acceptor site (AltA), Alternative Position (AltP, both acceptor and donor sites are different), Exon Skipping (ExonS), and Intron Retention (IntronR). C.D. indicates a known Chromosome Duplication region. Functional groups of proteins are separated by long lines spanning all columns. Different members in the group are separated by short lines starting at Arabidopsis gene name. Genes duplicated in Arabidopsis are clustered together with no line between them. Dash line separate the Prp19 complex from other 35S U5 associated proteins and * indicates proteins in that complex.

Human homologs	S.c.	Gene_Name	GeneID	C	Tnb	AltS	C.D.	Protein Domain	Ref
1.1 Sm Core proteins									
SmB	SmB1	atSmB-a	At5g44500	5	7		>4-5a	Sm, 1	
		atSmB-b	At4g20440	4	21	IntronR (1);	>4-5a	Sm, 1	
SmD1	SmD1	atSmD1-a	At3g07590	3	7	IntronR (1);		Sm, 1	
		atSmD1-b	At4g02840	4	13			Sm, 1	
SmD2	SmD2	atSmD2-a	At2g47640	2	7	AltA (1); AltD (1);		Sm, 1	
		atSmD2-b	At3g62840	3	25	AltA (1);		Sm, 1	
SmD3	SmD3	atSmD3-a	At1g76300	1	9		>1-1c	Sm, 1	
		atSmD3-b	At1g20580	1	7		>1-1c	Sm, 1	
SmE	SmE	atSmE-a	At4g30330	4	2		>2-4b	Sm, 1	
		atSmE-b	At2g18740	2	10	AltA (1);	>2-4b	Sm, 1	
SmF	SmF	atSmF	At4g30220	4	6			Sm, 1	
SmG	SmG	atSmG-a	At2g23930	2	13			Sm, 1	
		atSmG-b	At3g11500	3	9			Sm, 1	
LSM2	LSm2	atLSM2	At1g03330	1	7			Sm, 1	
LSM3	LSm3	atLSM3a	At1g21190	1	6		>1-1c	Sm, 1	
		atLSM3b	At1g76860	1	16		>1-1c	Sm, 1	
LSM4	LSm4	atLSM4	At5g27720	5	13			Sm, 1	
LSM5	LSm5	atLSM5 / SAD1	At5g48870	5	7	AltA (1);		Sm, 1	[44]
LSM6	LSm6	atLSM6a	At3g59810	3	7		>2-3	Sm, 1	

		atLSM6b	At2g43810	2	5		>2-3	Sm, 1	
LSM7	LSm7	atLSM7	At2g03870	2	6			Sm, 1	
LSM8	LSm8	atLSM8	At1g65700	1	9			Sm, 1	
LSM1	LSm1	atLSM1a	At1g19120	1	8			Sm, 1	
		atLSM1b	At3g14080	3	9	IntronR (1);		Sm, 1	
1.2 U1 snRNP specific proteins									
U1A Subunit	Mud1	atU1A	At2g47580	2	14	ExonS (1);		RRM, 2	[46]
U1C Subunit	Yhc1	atU1C	At4g03120	4	5			C2H2, 1; mrCtermi, 3	
U1-70K	Snp1	atU1-70K	At3g50670	3	32	IntronR (1);		RRM, 1	[45]
-	Prp39	atPrp39a	At1g04080	1	12	ExonS (6);		HAT, 7; TPR-like, 1	
		atPrp39b	At5g46400	5	1			HAT, 4;	
FBP11	Prp40	atPrp40a	At1g44910	1	10	IntronR (1);		WW, 2; FF, 5	
FBP11	Prp40	atPrp40b	At3g19670	3	5			WW, 2; FF, 5	
Luc7-like protein	Luc7	atLuc7a	At3g03340	3	6			DUF259, 1	
		atLuc7b	At5g17440	5	8			DUF259, 1	
Related to Luc7-like protein	Luc7	atLuc7-rl	At5g51410	5	7	IntronR (1);		DUF259, 1	
1.3 17S U2 snRNP specific proteins									
U2A' Subunit	Lea1p	atU2A	At1g09760	1	21			LRR 4;	
U2B" Subunit	Msl1p	atU2B"a	At1g06960	1	6	AltD (1);	>1-2a	RRM, 2	
		atU2B"b	At2g30260	2	13	AltA (1); IntronR (1);	>1-2a	RRM, 2;	
SF3a120/SAP114 Subunit	Prp21p	atSAP114-1a	At1g14650	1	17	AltB (1);		SWAP/Surp, 2; Ubiquitin, 1	
		atSAP114-1b	At1g14640	1				SWAP/Surp, 2	
		atSAP114-2	At5g06520	5				SWAP/Surp, 4	
		atSAP114-3	At4g16200	4	1			SWAP/Surp, 3	
		atSAP114p	At4g15580	4				SWAP/Surp, 3; Ubiquitin, 1	
SF3a60/SAP61 Subunit	Prp9p	atSAP61	At5g06160	5	10	AltD (1);		C2H2, 1	
SF3a66/SAP62 Subunit	Prp11p	atSAP62	At2g32600	2	13			C2H2, 1;	
SF3b120/SAP130 Subunit	Rse1p	atSAP130a	At3g55200	3	6			CPSF_A, 1; WD40-like, 1	[47]
		atSAP130b	At3g55220	3	7			CPSF_A, 1; WD40-like, 1	[47]
SF3b150/SAP145 Subunit	Cus1p	atSF3b150	At4g21660	4	16			PSP, 1; DUF382, 1	
		atSF3b150p	At1g11520	1					
SF3b160/SAP155 Subunit	Hsh155	atSAP155	At5g64270	5	11			HEAT, 1; ARM, 2; SAP_155, 1	
SF3b53/SAP49 Subunit	Hsh49p	atSAP49a	At2g18510	2	20			RRM, 2	
		atSAP49b	At2g14550	2				RRM, 2	
p14	Snu17p	atP14-1	At5g12190	5	7			RRM, 1;	
		atP14-2	At2g14870	2				RRM, 1;	
SF3b 14b /PHP5A	Rds3p	atSF3b_14b-a	At1g07170	1	10		>1-2a	UPF0123, 1;	
		atSF3b_14b-b	At2g30000	2	8		>1-2a	UPF0123, 1;	
SF3b 10		SF3b10a	At4g14342	4	11			SF3b10, 1;	
		SF3b10b	At3g23325	3	6			SF3b10, 1;	
1.4 U5 snRNP specific proteins									
15kD Subunit	Dib1p	atU5-15	At5g08290	5	28			DIM1, 1; Thioredoxin_2; 1	
40kD Subunit		atU5-40	At2g43770	2	21			WD-40, 7;	

100kD Subunit	Prp28p	atU5-100KD	At2g33730	2	13		DEAD, 1; Helicase_C, 1
102KD/Prp6-like	Prp6p	atU5-102KD	At4g03430	4	18		Ubiquitin, 1; TPR, 3; HAT, 15; TPR-like, 2; Prp1_N, 1
116 kD Subunit /elongation	Snu114p	atU5-116-1a	At1g06220	1	19	ExonS (1);	EFG_C, 1; GTP_EFTU, 1; GTP_EFTU_D2; 1; Small_GTP, 1; EFG_IV, 1;
		atU5-116-1b	At5g25230	5			EFG_C, 1; GTP_EFTU, 1; GTP_EFTU_D2; 1; EFG_IV, 1;
		atU5-116-2	At1g56070	1	214		EFG_C, 1; GTP_EFTU, 1; GTP_EFTU_D2; 1; EFG_IV, 1;
		atU5-116-3	At3g22980	3	3		EFG_C, 1; GTP_EFTU, 1; Small_GTP, 1;
200kD Subunit/Helicase	Brr2p	atU5-200-1	At5g61140	5	11	IntronR (1);	DEAD, 2; Helicase_C, 2; Sec63, 2; ARM, 1
		atU5-200-2a	At1g20960	1	23		DEAD, 2; Helicase_C, 2; Sec63, 2
		atU5-200-2b	At2g42270	2	5		DEAD, 2; Helicase_C, 2; Sec63, 2
		atU5-200-3	At3g27730	3			DEAD, 1; Sec63, 1; RuvA domain 2-like, 1
220kD Subunit	Prp8p	atU5-220/ Prp8a	At1g80070	1	33		Mov34, 1
		atU5-220/ Prp8b	At4g38780	4	2		Mov34, 1
1.5 U4/U6 snRNP specific proteins							
U4/U6-90K / SAP90	Prp3p	atSAP90-1	At1g28060	1	10		
		atSAP90-2	At3g55930	3			
		atSAP90-3	At3g56790	3			
U4/U6-60K / SAP60	Prp4p	atSAP60	At2g41500	2	8		WD-40, 7; SFM, 1; WD40-like, 1
U4/U6-20K / CYP20		atTri-20	At2g38730	2	11		Pro_isomerase, 1
U4/U6-61KD	Prp31	atU5-61/ Prp31a	At1g60170	1	26		Nop, 1
		atU5-61/ Prp31b	At3g60610	3			Nop, 1
U4/U6-15.5K	Snu13p	atU4/ U6-15.5a	At5g20160	5	18	IntronR (2);	Ribosomal_L7Ae, 1
		atU4/ U6-15.5b	At4g12600	4	14		Ribosomal_L7Ae, 1
		atU4/ U6-15.5c	At4g22380	4	9		Ribosomal_L7Ae, 1
1.6 Tri-snRNP specific proteins							
Tri-65 KD	Snu66p	atTri65a	At4g22350	4	7		UCH; 1; ZnF_UBP, 1
		atTri65b	At4g22290	4	20		UCH; 1; ZnF_UBP, 1; Pentaxin, 1
		atTri65a	At4g22410	4			UCH; 1; ZnF_UBP, 1
Tri-110 KD	SAD1	atTri110	At5g16780	5	7		SART-1, 1
Tri-27kD/RY1		atTri-27kD/ RY1	At5g57370	5	14		
hSnu23/FLJ31121	Snu23p	atSnu23	At3g05760	3	7		ZnF_U1, 1;
1.7 18s U11/U2 snRNP specific proteins							
U11/U12-35K		atU11/ U12-35kD	At2g43370	2	7	IntronR (1);	RRM, 1
U11/U12-25K (-99 protein)		atU11/U12-25K	At3g07860	3	6	IntronR (2);	C2H2, 1;
U11/U12-65K		atU11/ U12-65K	At1g09230	1	15	AltA (1);	RRM, 2;PHOSPHOPANTETHEINE, 2;
U11/U12-31K (MADP1)		atU11/ U12-31K	At3g10400	3	5		RRM, 1;CCHC, 1;
2.1 Splice site selection							
U2AF35		atU2AF35a/ AUSa	At1g27650	1	26		RRM, 1; CCCH, 2;
		atU2AF35/	At5g42820	5	8		RRM, 1; CCCH, 2; [55]

		AUSb						
U2AF65	Mud2	atU2AF65b/ AULa	At1g60900	1	10		RRM, 3;	[55]
		atU2AF65a/ AULb	At4g36690	4	29	AltA (1); IntronR (2);	RRM, 2;	[55]
		atULrp	At2g33440	2	2		RRM, 1	
		AUL3p	At1g60830	1				
U2AF35 related protein		atUrp	At1g10320	1			RRM, 1; CCCH, 2;	
SF1/BBP		atSF1/ BBP	At5g51300	5	23	IntronR (1);	RRM, 1; CCHC, 2; KH, 1;	
CBP20	Cbc1	atCBP20	At5g44200	5	8		RRM, 1	[53]
CBP80	Cbc2p	atCBP80	At2g13540	2	21		MIF4G, 1; ARM, 3	[53]
PTB/hnRNP I		atPTB1	At1g43190	1	26		RRM, 4;	
		atPTB2a	At3g01150	3	21	AltD (1); ExonS (1);	RRM, 2	
		atPTB2b	At5g53180	5	17	ExonS (1);	RRM, 2	
2.2 SR proteins								
SC35		atSC35	At5g64200	5	32	AltD (1);	RRM, 1;	[58]
SRrp40/TASR-2		atSR33/ atSCL33	At1g55310	1	12	IntronR (1); >1-3b	RRM, 1	[60]
		atSCL30a	At3g13570	3	32	ExonS (2); IntronR (4); >1-3b	RRM, 1	[58]
		atSCL30	At3g55460	3	14	ExonS (1);	RRM, 1	[58]
		atSCL28	At5g18810	5	5		RRM, 1	[58]
SF2/ASF		atSR1/ atSRp34	At1g02840	1	37	AltA (1); IntronR (1); >1-4	RRM, 2	[61, 64]
		atSRp34a	At4g02430	4	13	AltA (1); ExonS (1); IntronR (4); >1-4	RRM, 2	
		atSRp34b	At3g49430	3	3	ExonS (1); IntronR (1);	RRM, 2	
		atSRp30	At1g09140	1	15	AltA (1);	RRM, 2	[62]
9G8		atRSZp22/ atSRZ22	At4g31580	4	26	>2-4e	RRM, 1; CCHC, 1	[60, 63]
		atRSZp22a	At2g24590	2	7	>2-4e	RRM, 1; CCHC, 1	[60, 63]
		atRSzp21/ atSRZ21	At1g23860	1	18		RRM, 1; CCHC, 1	[60, 63]
		atRSZ33	At2g37340	2	30	IntronR (1); >2-3	RRM, 1; CCHC, 2	[58]
		atRSZ34	At3g53500	3	36	AltA (1); IntronR (3); >2-3	RRM, 1; CCHC, 2	[58]
-		atRSp32	At2g46610	2	23	AltD (1); IntronR (1); >2-3	RRM, 2	
		atRSp31	At3g61860	3	17	AltA (1); >2-3	RRM, 2	[56]
		atRSp41	At5g52040	5	34	AltA (1); >4-5b	RRM, 2	[56]
		atRSp40/ atRSP35	At4g25500	4	15	ExonS (1); IntronR (1); >4-5b	RRM, 2	[56]
2.3 17S U2 associated proteins								
hPrp43	Prp43p	atPrp43-1	At5g14900	5			HA2, 1	
		atPrp43-2a	At3g62310	3	17	AltA (1); >2-3	DEAD, 1; Helicase_C, 1; HA2, 1	
		atPrp43-2b	At2g47250	2	14	>2-3	DEAD, 1; Helicase_C, 1; HA2, 1	
SR140		atSR140-1	At5g25060	5	11		Surp, 1;RRM, 1;, 1;RPR, 1;	
		atSR140-2	At5g10800	5	2		Surp, 1;RRM, 1;RPR, 1;	
SPF45		atSPF45	At1g30480	1	9		D111/G-patch domain, 1; RRM, 1;	
SPF30		atSPF30	At2g02570	2	9	AltA (1);	Tudor, 1;	
2.4 35S U5 associated proteins								
hPrp19 *	Prp19p	atPrp19a	At1g04510	1	18	>1-2a	WD-40, 7; Ubox, 1;	

		atPrp19b	At2g33340	2	27	IntronR (1);	>1-2a	WD-40, 7; Ubox, 1;	
CDC5 *	Cef1	atCDC5	At1g09770	1	12			SANT, 2;	[100]
PRL1 *	Prp46p	atPRL1	At4g15900	4	14			WD-40, 2;WD40like, 1;	
		atPRL2	At3g16650	3	6			WD-40, 2;WD40like, 1;	
AD-002 *	Cwc15p	atAD-002	At3g13200	3	22			Cwf_Cwc_15, 1;	
HSP73/HSPA8 *		HSP73-1	At3g12580	3	35			Hsp70, 1;	
		HSP73-2	At5g42020	5	51	IntronR (1);		Hsp70, 1;	
		HSP73-3	At5g02500	5	553	IntronR (1);		Hsp70, 1;	
SPF27/BCAS2 *		atSPF27	At3g18165	3	15			BCAS2, 1;	
beta catenin-like 1 *		atCTNNNBL1	At3g02710	3	12			Armadillo, 1;ARM, 1;	
hSyf1	Syf1p	atSyf1	At5g28740	5	7			TPR, 1;HAT, 10;TPRlike, 3;	
hSyf3/CRN	Syf3	atCRN1a	At5g45990	5				TPR, 1; HAT, 14; TPR-like, 2	
		atCRN1b	At3g13210	3				TPR, 1; HAT, 12; TPR-like, 2	
		atCRN1c	At5g41770	5	13			TPR, 1; HAT, 14; TPR-like, 2	
		atCRN2	At3g51110	3	8			TPR, 1; HAT, 9; TPR-like, 1	
hIsy1	Isy1p	atIsy1	At3g18790	3	10			Isy1, 1;	
GCIP p29	Syf2	atGCIPp29	At2g16860	2	12				
SKIP	Prp45p	atSKIP	At1g77180	1	28			SKIP/SNW, 1;	
hECM2	Ecm2p	atECM2-1a	At1g07360	1	21		>1-2a	RRM, 1;CCCH, 1;	
		atECM2-1b	At2g29580	2	10		>1-2a	RRM, 1;CCCH, 1;	
		atECM2-2	At5g07060	5				CCCH, 1;	
KIAA0560		atAquarius	At2g38770	2	11				
MGC23918		atMGC23918	At3g05070	3	7				
G10	Cwc14p	atG10	At4g21110	4	12			G10, 1;	
Cyp E		atCypE1a/ CYP2	At2g21130	2	4		>2-4c	Pro_isomerase, 1	
		atCypE1b	At4g38740	4	59		>2-4c	Pro_isomerase, 1;	
		atCypE2a/ ROC3	At2g16600	2	39		>2-4a	Pro_isomerase, 1	
		atCypE2b	At4g34870	4	80		>2-4a	Pro_isomerase, 1;	
PPIase-like 1		atPPIase-like1	At2g36130	2	10			Pro_isomerase, 1;	
2.5 Proteins specific for BAU1									
NPW38		atNPW38	At2g41020	2	16	AltD (1); IntronR (1);		WW, 2;	
N-CoR1		atN-CoR1	At3g52250	3	3			SANT, 2;Homeodomain_like, 2;	
hPrp4 kinase		atPRP4K-1	At3g25840	3	13	ExonS (1);		Pkinase, 1;TyrKc, 1;S_Tkc, 1;, 1;Kinase_like, 1;	
		atPRP4K-2	At1g13350	1	5	IntronR (1);		Pkinase, 1;TyrKc, 1;S_Tkc, 1;, 1;Kinase_like, 1;	
		atPRP4K-3	At3g53640	3				Pkinase, 1;TyrKc, 1;S_Tkc, 1;, 1;Kinase_like, 1;	
FBP-21		atFBP21	At1g49590	1	12	ExonS (1); IntronR (3);		C2H2, 1;	
TBL1-rp 1		atTBL1-rp1	At5g67320	5	14			WD-40, 5;Peptidase_S9A_N, 1;LisH, 1;WD40like, 1;	
Smc-1		atSmc1	At3g54670	3	12			ATP_GTP_A_BS, 1;SMC_N, 1;SMC_C, 1;ABC_transporter, 1;SMC_hinge, 1;	
2.6 Exon junction complex (EJC) proteins									
ALY	Yra1p	atALY-1a	At5g02530	5	19	IntronR (1);		RRM, 1;	
		atALY-1b	At5g59950	5	16	IntronR (1);		RRM, 1;	
		atALY-2a	At5g37720	5	17		>1-5b	RRM, 1;	
		atALY-2b	At1g66260	1	38	ExonS (1);	>1-5b	RRM, 1;	

Y14		atY14	At1g51510	1	10	IntronR (1);	RRM, 1;RBM8, 4;	
Srm160-like		atSRM102	At2g29210	2	18	AltA (1);	PWI, 1	
Magoh		atMagoh	At1g02140	1	19		Mago_nashi, 1;	
Nuk-34/eIF4A3 /DDX48		atDDX48/ eIF4A3-1	At3g19760	3	50	>1-3a	DEAD, 1;Helicase_C, 1;	
		atDDX48/ eIF4A3-2	At1g51380	1	5	>1-3a	DEAD, 1;Helicase_C, 1;	
RNPS1		atSR45/ atRNPS1	At1g16610	1	27	AltA (1);	RRM, 1	[60]
UAP56		atUAP56a	At5g11200	5	21	AltA (1);	DEAD, 1; Helicase_C, 1	
		atUAP56b	At5g11170	5	25		DEAD, 1; Helicase_C, 1	
pinin		atPinin	At1g15200	1	9	AltA (1);	Pinin/SDK/memA, 1;	
2.7 Second step splicing factors								
Prp22	Prp22	atPrp22-1	At3g26560	3	11		DEAD, 1; Helicase_C, 1; S1, 1; HA2, 1;	
		atPrp22-2	At1g26370	1	5		DEAD, 1; Helicase_C, 1; HA2, 1	
		atPrp22-3	At1g27900	1	15		DEAD, 1; Helicase_C, 1; HA2, 1	
Prp17	Prp17p	atPrp17-1	At1g10580	1	10		WD-40, 7;	
		atPrp17-2	At5g54520	5	5	AltA (1);	WD-40, 6;	
Prp18	Prp18	atPrp18-1	At1g03140	1	16		Prp18, 1; SFM 1;	
		atPrp18-2	At1g54590	1			Prp18, 1	
Slu7	Slu7p	atSLU7-1a	At1g65660	1	6			
		atSLU7-1b	At4g37120	4	11			
		atSLU7-2	At3g45950	3				
Prp16	Prp16p	atPrp16	At5g13010	5	22		DEAD, 1; Helicase_C, 1; HA2, 1	
2.8 Other known splicing factors								
SRm300		atSRM300like	At3g23900	3	5	AltD (1);	RRM, 1; Filamin/ABP280 repeat, 1	
hTra-2/SFRS10		atTra/ SFRS1	At1g07350	1	25	ExonS (1); IntronR (3);	RRM, 1	
Prp2		atPrp2-1a	At1g32490	1	9	>1-2c	DEAD, 1; Helicase_C, 1; HA2, 1	
		atPrp2-1b	At2g35340	2		>1-2c	DEAD, 1; Helicase_C, 1; HA2, 1	
		atPrp2-2	At4g16680	4			DEAD, 1; Helicase_C, 1; HA2, 1	
Prp5		atPrp5-1a	At3g09620	3			DEAD, 1; Helicase_C, 1	
		atPrp5-1b	At1g20920	1	11		DEAD, 1; Helicase_C, 1	
		atPrp5-2	At2g47330	2	9		DEAD, 1; Helicase_C, 1	
hDbr1	dbr1	atDbr1	At4g31770	4	12		Metallophos, 1; DBR1, 1	
3.1 SR protein Kinase								
Lammer/CLK kinase		AFC1	At3g53570	3	11	AltA (1); IntronR (3);	PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	[71]
		AFC2	At4g24740	4	9	ExonS (1);	PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	[71]
		AFC3	At4g32660	4	9	AltD (1); IntronR (1);	PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	[71]
SRPK1		atSRPK1a	At2g17530	2	7	>2-4a	PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	
		atSRPK1b	At4g35500	4	10	>2-4a	PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	
SRPK2		atSRPK2a	At5g22840	5	2		PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	
		atSRPK2b	At3g53030	3	7		PKinase, 1; TyrKc, 1; S_Tkc, 1; PKinase-like, 1	
		atSRPK2c	At3g44850	3	1		PKinase, 1; TyrKc, 1; PKinase-like, 1	
3.2 Glycine-Rich RNA binding protein								
hnRNP A/B		atGRBP1a	At1g18630	1	5	>1-1c	RRM, 1	

	atGRBP1b	At1g74230	1	14		>1-1c	RRM, 1; Eggshell, 4	
	atGRBP1c	At4g13850	4	17		>3-4	RRM, 1	
	atGRBP1d	At3g23830	3	12	AltA (1);	>3-4	RRM, 1	
	atGRBP1e	At5g61030	5	8			RRM, 1; PfkB_Kinase, 1	
	atGRBP2	At2g16260	2				RRM, 1	
	AtGRP7/ atGRBP3a	At2g21660	2	182	AltD (1); IntronR (3);	>2-4c	RRM, 1	[74]
	AtGRP8/ atGRBP3b	At4g39260	4	67	AltB (1); AltD (1); IntronR (5);	>2-4c	RRM, 1	[74]
3.3 hnRNP A/B family								
hnRNP A/B	AtRNPA/ B_1	At4g14300	4	4			RRM, 2	[11]
	AtRNPA/ B_2	At2g33410	2	13			RRM, 2	[11]
	AtRNPA/ B_3	At5g55550	5	13	IntronR (3);	>4-5c	RRM, 2	[11]
	AtRNPA/ B_4	At4g26650	4	21		>4-5c	RRM, 2	[11]
	AtRNPA/ B_5	At5g47620	5	12	AltD (2);		RRM, 2	[11]
	AtRNPA/ B_6	At3g07810	3	18	AltA (1);		RRM, 2; FKBP_PPIASE_2, 2	[11]
	AtRNPA/ B_7	At1g58470	1	6			RRM, 2	
	AtRNPA/ B_8a	At5g40490	5	3			RRM, 2; Eggshell, 4	
	AtRNPA/ B_8b	At1g17640	1				RRM, 2	
	AtRNP_N1	At3g13224	3	16	IntronR (1);		RRM, 2;HUDSXLRNA, 2;	
	UBA2a	At3g56860	3	23	IntronR (1);	>2-3	RRM, 2	[75]
	UBA2b	At2g41060	2	9		>2-3	RRM, 2	[75]
	UBA2c	At3g15010	3	10	IntronR (1);		RRM, 2	[75]
3.4 Other hnRNPs (with animal homologs)								
hnRNP E1/E2	at-hnRNP-E	At3g04610	3	10			KH, 3;	
hnRNP F/ hnRNP H	at-hnRNP-F/ AtRNPH/ F_1	At5g66010	5	9	AltA (1);		RRM, 2	[11]
	at-hnRNP-H/ AtRNPH/ F_2	At3g20890	3				RRM, 2	[11]
hnRNP G	at-hnRNP-G1	At5g04280	5	6			RRM, 1; CCHC, 1	
	at-hnRNP-G2	At3g26420	3	35	AltA (1);		RRM, 1; CCHC, 1	
	at-hnRNP-G3	At1g60650	1	7			RRM, 1; CCHC, 1	
hnRNP P2	at-hnRNP-P	At1g50300	1	9			RRM, 1; ZnF_RBZ, 2	
hnRNP R/Q	hnRNP-R1	At4g00830	4	19			RRM, 3;	
	hnRNP-R2	At3g52660	3	1		>2-3	RRM, 3;	
	hnRNP-R3 / AtRNPA/ B_9	At2g44710	2	13		>2-3	RRM, 3	
CUG-BP	AtCUG-BP1	At4g03110	4	4	AltA (1); IntronR (1);		RRM, 3; HUDSXLRNA, 4	[11]
	AtCUG-BP2	At1g03457	1	9	AltA (1);		RRM, 3; HUDSXLRNA, 4	[11]
(CUG-BP)	atFCA1	At4g16280	4	13	AltB (1); IntronR (1);		RRM, 2; WW, 1	[78]
	atFCA2	At2g47310	2	6			RRM, 2; WW, 1	
3.5 Other plant hnRNPs								
	AtUBP1a	At1g54080	1	48	AltA (1);	>1-3b	RRM, 3	[81]
	AtUBP1c	At3g14100	3	13		>1-3b	RRM, 3	[81]
	AtUBP1b	At1g17370	1	17			RRM, 3	[81]
	UBA1a	At2g22090	2	15		>2-4c	RRM, 1	[75]
	UBA1b	At2g22100	2	2		>2-4c	RRM, 1	[75]
	UBA1c	At2g19380	2	1			RRM, 1; C2H2, 3	[75]
	atRBP45a	At5g54900	5	42		>4-5c	RRM, 3	[82]

atRBP45c	At4g27000	4	52		>4-5c	RRM, 3	[82]
AtRBP45b	At1g11650	1	53			RRM, 3	[82]
atRBP45d	At5g19350	5	10			RRM, 3	
AtRBP47a	At1g49600	1	10		>1-3a	RRM, 3	[82]
AtRBP47b	At3g19130	3	21		>1-3a	RRM, 3	[82]
AtRBP47c	At1g47490	1	23	IntronR (1);		RRM, 3	[82]
AtRBP47c'	At1g47500	1	12			RRM, 3	[82]
Ath1	At4g16830	4	34			HANP4_PAI-RBP1, 1	[101]
Ath2	At4g17520	4	29		>4-5a	HANP4_PAI-RBP1, 1	[101]
Ath3	At5g47210	5	67	IntronR (1);	>4-5a	HANP4_PAI-RBP1, 1	

Abbreviations for domains are as follows:

ABC_transporter: ABC transporter; Armadillo: Armadillo; ARM: ARM repeat fold; ATP_GTP_A_BS: ATP/GTP-binding site motif A (P-loop); BCAS2: Breast carcinoma amplified sequence 2; C2H2: Zn-finger, C2H2 matrin type; C2H2: Zn-finger, C2H2 type; CCCH: Zn-finger, C-x8-C-x5-C-x3-H type; CCHC: Zn-finger, CCHC type; CPSF_A: CPSF A subunit, C-terminal; Cwf_Cwc_15: Cwf15/Cwc15 cell cycle control protein; DBR1: Lariat debranching enzyme, C-terminal; DEAD: ATP-dependent helicase, DEAD-box; DEAD: DEAD/DEAH box helicase; DIM1: Pre-mRNA splicing protein; DUF259: Protein of unknown function DUF259; DUF382: Protein of unknown function DUF382; EFG_C: Elongation factor G, C-terminal; EFG_IV: Elongation factor G, domain IV; Eggshell: Eggshell protein; FF: FF domain; FKBP_PPIASE_2: Peptidylprolyl isomerase, FKBP-type; G10: G10 protein; GTP_EFTU_D2: Elongation factor Tu, domain 2; GTP_EFTU: Protein synthesis factor, GTP-binding; HA2: Helicase-associated region; HANP4_PAI-RBP1, 1: Hyaluronan/mRNA binding protein; HAT: RNA-processing protein, HAT helix; Helicase_C: Helicase, C-terminal; Homeodomain_like: Homeodomain-like; Hsp70: Heat shock protein Hsp70; HUDSXLRNA: Paraneoplastic encephalomyelitis antigen; Isy1: Isy1-like splicing; Kinase_like: Protein kinase-like; LisH: Lissencephaly type-1-like homology motif; LRR: Leucine-rich repeat,; Mago_nashi: Mago nashi protein; Metalloph: Metallophosphoesterase; MIF4G: Initiation factor eIF-4 gamma, middle; Mov34: Mov34/MPN/PAD-1; mrCtermi: Molluscan rhodopsin C-terminal tail; Nop: Pre-mRNA processing ribonucleoprotein, binding region; Peptidase_S9A_N: Peptidase S9A, prolyl oligopeptidase, N-terminal beta-propeller domain; PfkB_Kinase: Carbohydrate kinase, PfkB; PHOSPHOPANTETHEINE: Phosphopantetheine attachment site; Pinin/SDK/memA: Pinin/SDK/memA protein; Pkinase: Protein kinase; Pro_isomerase:

Peptidyl-prolyl cis-trans isomerase, cyclophilin type; Prp18: Prp18 domain; Prp1_N: PRP1 splicing factor, N-terminal; PSP: PSP, proline-rich; PWI: Splicing factor PWI; RBM8: RNA binding motif protein 8; Ribosomal_L7Ae: Ribosomal protein L7Ae/L30e/S12e/Gadd45; RPR: Regulation of nuclear pre-mRNA protein; RRM: RNA-binding region RNP-1 (RNA recognition motif); S1: RNA binding S1; SANT: Myb DNA-binding domain; SAP_155: Splicing factor 3B subunit_1; SART-1: SART-1 protein; Sec63: Sec63 domain; SF3b10: Splicing factor 3B subunit 10; SFM: Splicing factor motif; SKIP/SNW: SKIP/SNW domain; Small_GTP: Small GTP-binding protein domain; SMC_C: Structural maintenance of chromosome protein SMC, C-terminal; SMC_hinge: SMCs flexible hinge; SMC_N: SMC protein, N-terminal; Sm_like_riboprot: Small nuclear-like ribonucleoprotein; Sm: Small nuclear ribonucleoprotein (Sm protein); S_Tkc: Serine/threonine protein kinase; Surp: SWAP/Surp; Thioredoxin_2: Thioredoxin domain 2; TPRlike: TPR-like; TPR: TPR repeat; Tudor: Tudor domain; TyrKc: Tyrosine protein kinase; Ubox: Zn-finger, modified RING; UCH: Peptidase C19, ubiquitin carboxyl-terminal hydrolase family 2; UPF0123: Protein of unknown function UPF0123; UPF0123: Protein of unknown function UPF0123; WD-40: G-protein beta WD-40 repeat; WD40like: WD40-like; WW: WW/Rsp5/WWP domain; ZnF_RBZ: Zn-finger, Ran-binding; ZnF_U1: Zn-finger, U1-like; ZnF_UBP: Zn-finger in ubiquitin thiolesterase;

Table 3. Duplication source involving Arabidopsis splicing related proteins

*Family indicates both single copy gene and multiple-copy gene families. The C.D. Ratio column gives the fraction of all duplication events caused by chromosomal duplications.

[$] C.D. indicates chromosomal duplications

	Genes	Family*	Single / Multi.	Duplication Ratio	Duplication events	C.D.[$]	C.D. Ratio
snRNP proteins	91	54	27 / 27	50.0%	37	7	18.9%
Splicing factors	109	58	33 / 25	43.1%	51	14	27.5%
Splicing regulator	60	18	4 / 14	77.8%	42	11	26.2%
Total	260	130	64 / 66	50.8%	130	32	24.6%

Table 4. Alternative splicing in splicing related genes.

The column entries are the numbers of genes which the respective alternative splicing events. AltA: Alternative Acceptor site; AltD: Alternative Donor site; AltP: Alternative intron Position (both acceptor and donor sites are different); ExonS: Exon Skipping; IntronR: Intron Retention. The Overall and Ratio columns give the number and fraction of genes with any type of alternative splicing, respectively.

	Genes	AltA	AtlD	AltP	ExonS	IntronR	Overall	Ratio
snRNP proteins	91	6	3	1	3	11	22	23.2%
Splicing factors	109	14	5	0	11	21	38	34.9%
Splicing regulator	60	8	4	2	1	12	20	33.3%
Total	260	28	12	3	15	44	80	30.8%

Figure 1

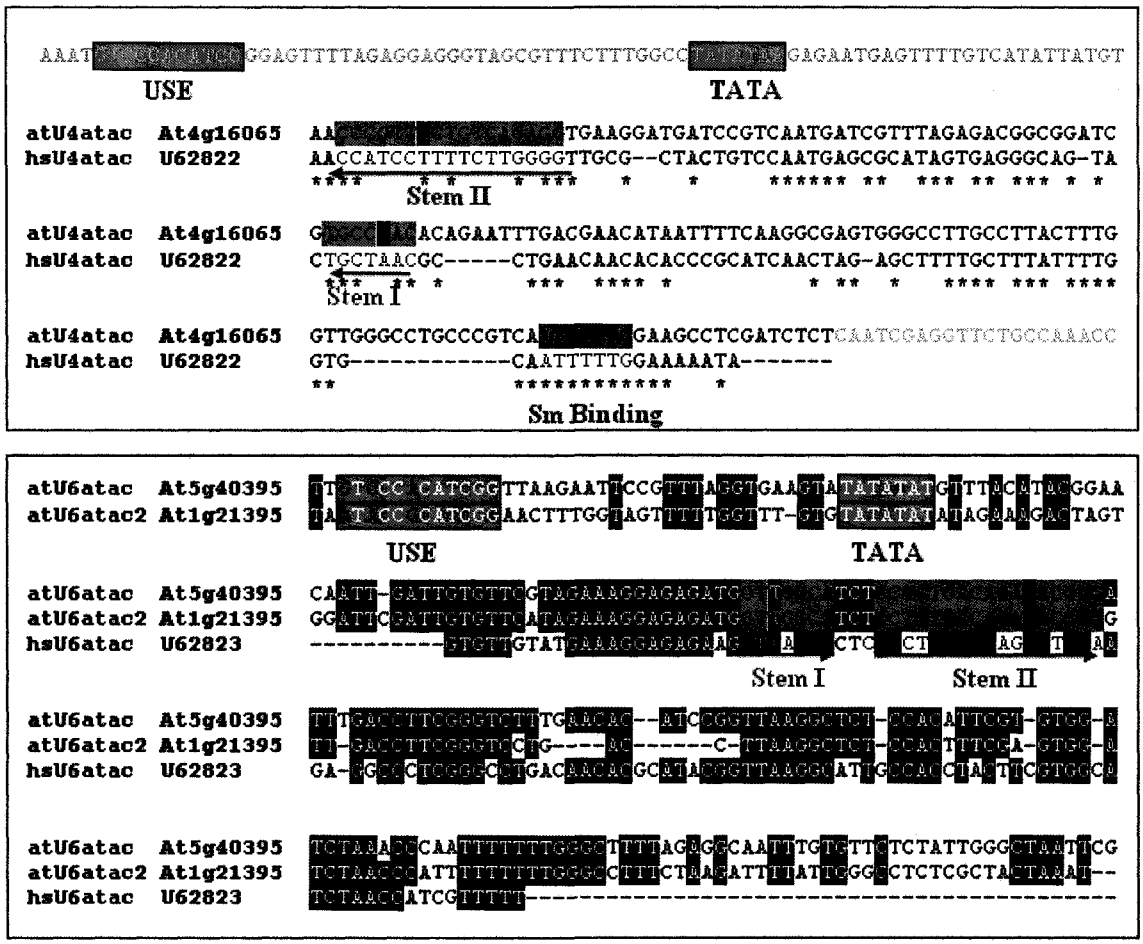

USE
TATA
atU4atac At4g16065
hsU4atac U62822
Stem II
Stem I
Sm Binding
atU6atac At5g40395
atU6atac2 At1g21395
hsU6atac U62823
USE
TATA
Stem I
Stem II

Figure 2

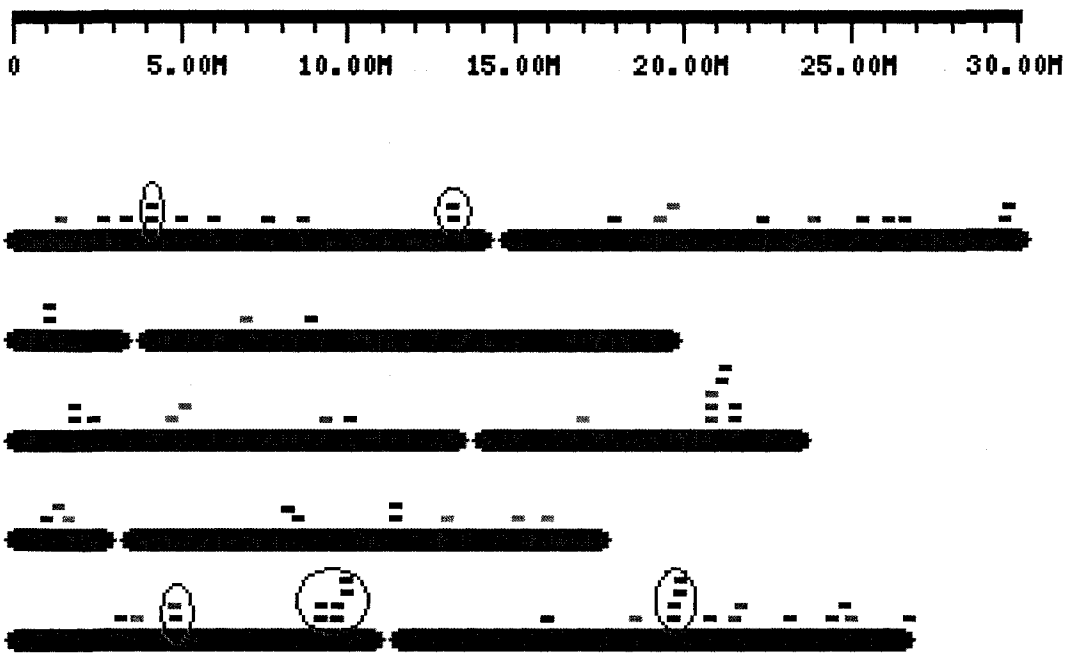
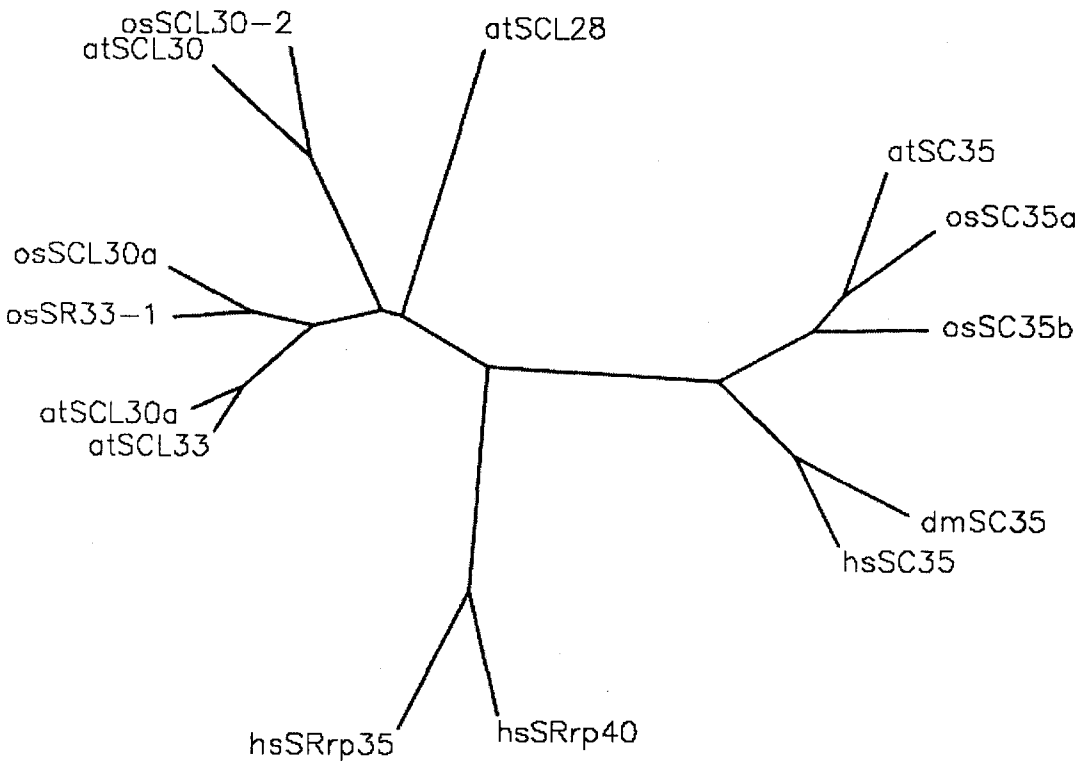


Figure 3


osSCL30-2
atSCL30
atSCL28
atSC35
osSC35a
osSCL30a
osSR33-1
osSC35b
atSCL30a
atSCL33
dmSC35
hsSC35
hsSRrp35
hsSRrp40

Chapter 3: Genome-wide comparative analysis of alternative splicing in plants

A paper to be submitted to *Proceedings of the National Academy of Sciences USA (PNAS)*

Bing-Bing Wang[1] and Volker Brendel[1,2]

[1]Department of Genetics, Development and Cell Biology and [2]Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

Abstract

Alternative splicing (AS) has been extensively studied in mammalian systems, but much less in plants. Here we report AS events deduced from EST/cDNA analysis in two model plants: Arabidopsis and rice. In Arabidopsis, 4,768 (22%) of the genes with EST/cDNA evidence show 8,423 AS events. About 56% of these events are intron retention (IntronR), and only 8% are exon skipping (ExonS). In rice, 6,638 (21.5%) of the expressed genes display 14,825 AS events, of which 54% are IntronR and 13.5% are ExonS. The consistent high frequency of IntronR suggests prevalence of splice site recognition by intron definition in plants. 22%-30% of the AS events occur in untranslated regions (UTRs), and 14-16% of the AS events are read-through with respect to the constitutive open reading frame (ORF). The remaining AS events change the start and/or stop codon position. In total, 36-43% of the AS events produce transcripts that would be targets of the nonsense-mediated decay (NMD) pathway, if

that pathway were to operate in plants as in humans. 40% of Arabidopsis AS genes are also alternatively spliced in rice, with some examples strongly suggesting a role of the AS event as an evolutionary conserved mechanism of post-transcriptional regulation. A sample of previously un-annotated AS events in Arabidopsis were experimentally confirmed by RT-PCR. We created a comprehensive web-interfaced database to compile and visualize the evidence for alternative splicing in plants (ASIP, available at: http://www.plantgdb.org/ASIP/).

Introduction

Alternative splicing (AS) is an important post-transcriptional regulatory mechanism that can increase protein diversity and affect mRNA stability (1, 2). Relative to the predominant transcript isoform, different types of AS have been observed, including exon skipping (ExonS), Alternative donor or acceptor (AltD/AltA) site, and intron retention (IntronR) (reviewed in (3), shown in Figure 1). AS has been extensively studied by EST/cDNA-based analysis in mammalian systems and 35-60% human genes were suggested to be alternatively spliced (4-9). Different types of AS have been observed with markedly different frequencies. In human, ExonS is the most common type (58% of the total AS is ExonS, plus 11% multiple skipped-exon) and IntronR is the least (5%) (4). About 70-88% AS occur in protein coding region (reviewed in (10)) and about one third of AS in human produce pre-mature termination codons (PTCs) (11). These PTC-containing transcripts are apparent targets for Nonsense-mediated mRNA decay (NMD) (11). Not all the predicted AS are real and functional, as many possible sources of false positives exist (10). The AS is defined as

functional "if it is required during the life-cycle of the organism and activated in a regulated manner" (12). To identify functional AS, conserved AS between human and mouse were studied (12-17), with the assumption that conservation indicates function. 25% of a subset of 980 ExonS events in human were found to be conserved in mouse (12). About 10% human gene loci with mouse orthologs showed conserved AS events (15). 38.4% of these events are ExonS and only 2.8% are IntronR (15).

The splicing mechanism in plants is generally conserved compared to mammalians (18, 19). However, introns in plant are usually short in length and U rich (18, 20), with much less apparent polypyrimidine tract near the 3' splice site than in vertebrate (21). Our recent genome-wide survey on Arabidopsis splicing related genes revealed variations in SR proteins and hnRNP proteins between plants and mammalians, suggesting differences in splicing regulation mechanism in plants (22). A few of AS events were identified experimentally in plants, including genes involved in splicing (23, 24), transcription (25), flowering regulation (26), disease resistance (27), enzyme activities (28, 29) and many other physiological processes and functions (19). A database (PASDB) collecting known alternatively spliced genes in plants is available at http://pasdb.genomics.org.cn (30). Although AS is thought to be less abundant in plants compared to animals, it begins to be recognized as playing an important role in the generation of plant proteome diversity (31).

Computational analysis of AS in plants was not available until recently. Ner-Gaon et al (32) identified 436 alternatively spliced genes in Arabidopsis by EST-pair alignment. The fraction of IntronR in their study was as high as 64%. A sampling of the IntronR events were confirmed by RT-PCR using polyribosome-RNA, demonstrating that these IntronR events are not the byproduct of incomplete splicing (32). Iida et al (33) aligned 248,514 RAFL

(RIKEN Arabidopsis Full-Length) cDNA/EST sequences to the Arabidopsis genome using a BLAST-based method. They identified 15,214 transcription units (TUs) containing at least two sequences each and observed alternative splicing for 11.6% of these TUs (33). Three other studies using smaller collection of EST/cDNA data briefly reported fewer AS events in Arabidopsis (9, 34, 35). All these pioneer studies revealed that a low fraction of genes (5-10%) are alternatively spliced and IntronR is the most prevalent AS type in Arabidopsis. However, none of the above studies did detailed analysis on the position and outcome of AS.

Compared with the case in human, the lower fraction of AS gene in plants is possibly due to lower cDNA/EST coverage. Millions of ESTs were used in human AS analysis (9), while less than one-tenth of that were used in Arabidopsis (9, 32-35). The number of current cDNA/EST sequences increased dramatically since last previous study, it is likely that much more AS events will be identified using current collection. As rice genome sequence becomes available recently (36, 37), and differences exist in the splicing mechanism between monocot and dicot plants (19), it is of great interests to explore the AS events in rice and compare them with Arabidopsis. In this study, we applied the GeneSeqer spliced-alignment program (38) to map currently available Arabidopsis and rice full-length cDNAs and ESTs to their respective genome sequences and identified thousands of AS events by exhaustive comparison of the deduced transcription units. The AS genes were comparatively analyzed and a small portion of the AS events were found to be conserved in the two plants. We also constructed a user-friendly database to store and visualize these AS events and will frequently update it to reflect recent changes on cDNA/EST collections in both plants.

Materials and Methods

Data Sources. Our initial analysis was based on a total of 323,340 Arabidopsis ESTs that were downloaded from GenBank using the ENTREZ query "Arabidopsis[ORGN] AND gbdiv_est[PROP]". Arabidopsis full-length cDNAs were retrieved from GenBank (query "Arabidopsis[ORGN] AND (FLI_CDNA[KYWD] OR GSLT_cDNA [KYWD])", with some additional sequences obtained from AtGDB (http://www.plantgdb.org/AtGDB/resource.php) for a total of 62,009 sequences. For rice (*Oryza sativa*), sets of 298,857 ESTs and 32,136 cDNAs were obtained from GenBank using similar queries. Updates to our database of plant AS events include more recent sequence depositions, but the analysis results reported here reflect only the sequences available at the time. The five Arabidopsis chromosome sequences were obtained from GenBank (accessions NC_003070, NC_003071, NC_003074, NC_003075, and NC_003076). For rice, our analysis was based on TIGR Release 3.0 of the twelve pseudochromosome sequences (downloaded from http://rice.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml).

Identification of AS in *Arabidopsis* and Rice. Arabidopsis and rice full-length cDNAs and ESTs were spliced aligned against their corresponding genomic origins using the GeneSeqer program (29) with species-specific splice site models and minimum allowed exon and intron lengths set to two and 20, respectively. Unix shell scripts and Perl scripts were written to automate the following steps: (i) Loading of GeneSeqer alignment data, including the predicted exon and intron coordinates and similarity and splice site scores, into a MySQL database. (ii) Filtering of the alignment data. Only high-quality alignments (overall similarity score at least 0.8) were used. For ESTs/cDNAs matching multiple genomic loci,

only the best match (presumed cognate) was used, as described previously (18). Exons with individual similarity scores less than 0.8 and their flanking introns were removed from the remaining alignments. Introns with either local similarity score for the 50bp upstream or downstream flanking exon sequences less than 0.8 were also removed. Redundancy in the sets of qualifying exons and introns was eliminated. For any pair of terminal exons sharing the same exon/intron, border only the longer exon was retained. (iii) Identification of AS events. For each intron, its coordinates were compared with each overlapping exon and intron. Two overlapping introns with different 5'- and/or 3'-ends and all overlapping intron/exon pairs were considered as candidate AS events. For overlapping introns, the intron with the most cDNA/EST evidence was taken to be the constitutive intron. Several cases of AS were distinguished as follows. Use of an alternative donor/acceptor site (AltD / AltA) is defined as the case of an intron differing from the constitutive intron only in the donor or acceptor site. Alternative position (AltP) refers to an intron overlapping the constitutive intron but differing in both donor and acceptor site position. Exon skipping (ExonS) occurs when an exon in one transcript isoform is completely contained in an intron of another isoform (if this intron also qualified as an AltD, AltA, or AltP event, only the ExonS event was counted). Intron retention (IntronR) occurs in a transcript isoform that contains a sequence segment that is exactly spliced out in an alternative isoform (i.e., an intron in our database is contained in an exon in the database). (iv) Other transcript variants, including alternative terminal exons and alternative initial exons, were not considered. They may involve other mechanisms (such as transcription) coupling with splicing (39). Candidate exon skipping events involving first or last exons were also removed, because these also constitute examples of alternative initial or terminal exons, respectively.

Web Interface and Visualization of AS. We created a web interface for the MySQL database to access all the alignment data and graphically display the AS events using Perl-cgi, PHP, and Javascript. The site is referred to as Alternative Splicing In Plants (ASIP) and is accessible at http://www.plantgdb.org/ASIP/. All scripts were written in modular fashion such that the database can be easily updated and expanded to other species. All displays are integrated with the PlantGDB (40) Arabidopsis and rice genome browsers (http://www.plantgdb.org/AtGDB/ and http://www.plantgdb.org/OsGDB/, respectively), allowing to retrieve sequence records and to view the AS events in an expanded genome context.

Visualization of different AS types at ASIP is illustrated in Figure 1. For AltD, AltA, and AltP events, a vertical color bar at the exon/intron border denotes the alternative sites. Multiple alternative sites are distinguished by different colors. For ExonS events, a green box within in intron indicates the position of the skipped exon. To visualize IntronR events, the spliced intron is drawn in light green color, and the exon retaining that intron is drawn as a normal exon.

Derivation of Putative Transcript Isoforms (PTIs). In order to classify AS events with respect to their impact on the translation of the alternatively spliced transcript it is necessary to deduce the likely full-length alternative transcripts from overlapping ESTs (in addition to experimentally obtained full-length cDNA isoforms). For the purpose of this study, we make the assumption that all AS events within a transcription unit are mutually independent, and thus we derived all possible combinations of alternative exons to generate the set of PTIs for

each transcription unit. This was done by first clustering all exons and introns from out database into sets corresponding to distinct transcription units and then assembling complete gene structures by concatenating compatible exons and introns in 5'- to 3'-direction. This approach is different from TAP (8), which relies on a reference sequence structure, and also from PASA (35), which assembles only observed combinations of exons. In some cases of tightly spaced genes, assignment of some ESTs to a particular transcription unit may be ambiguous. However, this did not affect our evaluation of AS, which explicitly excluded events involving terminal exons. The positions of the start and stop codons for each PTI were determined by searching for the longest open reading frame (ORF) in the forward direction for PTIs with introns and in both directions for PTIs without introns.

Position and Outcome of AS Events. To classify AS events relative to their predicted effect on mRNA translation and stability, any two PTIs differing in only one AS event were compared. If multiple PTI pairs contain the same AS event, only the pair with the longest ORF was considered for classification of that AS event. The position of the AS event was determined as either 5'-UTR or 3'-UTR if the alternative spliced intron located outside of the putative coding region on both PTIs. Otherwise, the AS event was regarded as within ORF and classified as either "read-through" or altered-ORF". Read-through occurs if the two PTIs have the same start and stop codons and only differ by an internal stretch of in-frame codons. If the AS event changes the predicted start or stop codon, it is labeled altered-ORF.

To check if an AS event will produce a NMD candidate, the distance between the stop codon and last exon junction was calculated for both PTIs. If one isoform has distance greater than 50 nucleotides while the other isoform has distance smaller than 50, then the AS

event was regarded to produce a NMD candidate. If the AS event produces a premature stop codon (PTC) in one PTI relative to the other, this stop codon was used to calculate the NMD distance.

Gene Ontology (GO) Annotation. GO annotation for the alternative spliced genes was derived from the Arabidopsis GO annotation, downloaded from the TAIR website (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/). For rice, we transferred the GO terms of the most similar Arabidopsis gene product as tentative annotation. Statistical analysis was conducted in MicroSoft Excel.

Conserved Alternatively Spliced genes in *Arabidopsis* and Rice. To identify conserved AS events between Arabidopsis and rice, all annotated Arabidopsis proteins were matched against all annotated rice proteins, and vice versa, using BLAST (41). We used 10^{-20} as the E-value cut-off for BLAST and labeled the highest scoring hit as "uniquely best" if its E-value was at least 10^{-20} times lower than that of the next highest scoring hit. Reciprocal uniquely best hits were selected as close homolog pairs (potential orthologs). If there was evidence in both genes for the same type of AS event, this gene pair was characterized as conserving AS, although not necessarily the same position and outcome of AS. To further identify conserved AS position, we first derived the set of conserved introns between Arabidopsis and rice by matching all Arabidopsis introns plus 30 bp flanking sequences against rice intron and flanking sequences using tBLASTx (E-value less than 10^{-4} and requiring both flanking exons to have at least 10 bases as part of the BLAST hits). The subset of these introns that occur in close homolog pairs were defined as conserved introns

between Arabidopsis and rice. Conserved AS events are AS events occurring in conserved introns.

RNA Extraction and RT-PCR Validation. Total RNA was isolated from 0.1g each of different Arabidopsis tissues, including seedling, leaf, root, stem, flower, and silique, using RNeasy Plant Mini Kit (Qiagen, Valencia, CA) following the manufacturer's protocol. Then 2μg RNA of each sample was treated with RQ1 RNAase-free DNAase (Promega, Madison, WI), and half was used for cDNA synthesis using the SuperScript III First-Strand Synthesis System (Invitrogen, Carlsbad, CA). A mixture of cDNA aliquots was used for PCR amplification. Primer sequences are described in *Supporting Table 1*. PCR was conducted as follows: initial denaturation at 94 °C for 4 min, followed by 38 cycles for 94 °C for 1 min, 60 °C for 1 min, 72 °C for 2 min, and ending with a final extension at 72 °C for 8 min. PCR products were separated on 1.5% agarose gels and visualized by ethidium bromide staining. For sequence analysis, PCR products of the expected size were purified from the gel using QIAquick Gel Extraction kit (Qiagen, Valencia, CA). Direct sequencing was performed at the Iowa State University DNA facility.

Results and Discussion

Genome-wide EST/cDNA Alignments in Arabidopsis and Rice. 95.8% and 85.7% of the current Arabidopsis and rice EST/cDNA collections could be unambiguously aligned to their respective genomes using the GeneSeqer spliced alignment program (38). The unaligned ESTs/cDNAs are either from organelle genomes (chloroplast and mitochondria) or different

subspecies or are short and low-quality sequences. In total, 369,218 Arabidopsis ESTs/cDNAs were matched to the genome and produced 372,772 cognate alignments (see Materials and Methods). As shown in *Supporting Table 2*, less than 1% of the EST/cDNAs have multiple cognate alignments. 25,231 transcription units (TUs) were identified, 23,856 (94.6%) of which correspond to annotated gene regions and 1,375 (5.4%) to novel gene regions. The average number of ESTs/cDNAs per TU is 14.8. In rice, 283,816 ESTs/cDNAs produced 319,391 cognate alignments, with ~3% of the aligned EST/cDNAs showing multiple cognate alignments. The high proportion of multiple cognate alignments relative to Arabidopsis presumably reflects recent gene duplications in rice (42). We defined a total of 36,270 rice TUs, 87.7% of which overlap annotated genes and 12.3% are in novel regions. The average number of ESTs/cDNAs per rice TU is about 8.8.

Rice Introns Are Generally Longer and Have Higher GC-content Compared with Arabidopsis Introns. The GeneSeqer spliced alignments are associated with scores for each exon and intron that indicate the level of sequence similarity and predicted splice site probabilities, respectively (38). To reduce the possibility of misleading interpretations of erroneous alignments, we removed from the data set all exons with similarity scores less than 0.8 and also their flanking introns. The remaining entries for Arabidopsis comprised a total of 128,098 exons (including 7,605 single exons, 42,170 terminal exons, and 78,323 internal exons) and 94,342 introns. The average length of Arabidopsis introns in this set is 173bp, about the same size as internal exons (*Supporting Table 3).* Less than 1% of introns are longer than 1kb. Consistent with previous results (20, 43, 44), the Arabidopsis introns are U-rich (~41%), with an average GC-content of 32.7%, which is ~10% lower than the exon GC-

content. This compositional contrast between introns and exons has been shown to be critical to splicing efficiency and accuracy (19, 20, 45).

In rice, a total of 111,343 introns and 166,057 exons (including 15,372 single exons, 58,349 terminal exons, and 92,336 internal exons) were identified using the same criteria. The average size of rice introns is 433bp, ~2.5 times as long as Arabidopsis introns. More than 10% of rice introns are longer than 1kb. The GC-content of rice introns and exons is respectively ~5% higher than the GC-content of Arabidopsis introns and exons, while the difference between introns and exons is of about the same magnitude as in Arabidopsis. More detailed comparison of base composition reveals that Arabidopsis introns are more U-rich than rice introns, while A-composition is similar between the two plants (*Supporting Figure 1*). The size and GC-content differences between rice and Arabidopsis introns supports possible variations in the splicing mechanism between monocot and dicot plants (19).

About One Fifth of Expressed Genes Are Alternatively Spliced in Plants. Among the 21,641 Arabidopsis genes (including 1,375 novel genes) with EST/cDNA evidence, there are 4,768 (22.0%) that display a total of 8,423 AS events in Arabidopsis. Compared with recent estimates of 11.6% based on RAFL (RIKEN Arabidopsis Full-Length cDNA) sequences (33) and less than 5% indicated in a TIGR study (35) and other previous estimates (9, 34), our AS ratio is much higher. This may be because of the use of (i) more recent, larger EST/cDNA collections and/or (ii) a more sensitive AS detection method. Our Arabidopsis EST/cDNA collection includes with few exceptions all of TIGR's collection and about one third of RIKEN's dataset, as well as ~176,000 new sequences not included in the TIGR analysis and

~300,000 EST/cDNA sequences not included in the RIKEN data set. As shown in Figure 2, our list includes 852 of the 909 (93.7%) TIGR annotated alternatively spliced genes (excluding AS types involving terminal exons not discussed here) and 807 of the 1431 (56.4%) RIKEN annotated alternatively spliced genes showing alternative donor/acceptor site, exon skipping, or intron retention. 57 genes from the TIGR list are absent in our collection. Among these, only 19 genes showed reliable AS by our criteria after manual inspection, while the remaining 38 genes are likely false annotations given the lack of cognate EST/cDNA evidence (shown in *Supporting Table 4*). In addition, 10 exon skipping cases identified by TIGR lack EST/cDNA evidence. Other types of AS events were identified in ASIP for these genes (*Supporting Table 4*). We did not perform a detailed comparison between RIKEN's and our AS list, because the EST/cDNA collections are too different. Another study using pair-wise EST comparisons detected 436 alternatively spliced Arabidopsis genes (32), 418 of which are included in our database.

In rice, the overall AS ratio is very similar to that of Arabidopsis. About 21.5% (6,638 out of 30,917) of the expressed genes showed a total of 14,825 AS events. The TIGR rice genome annotation project identified 2,538 alternatively spliced genes (http://rice.tigr.org/tdb/e2k1/osa1/expression/alt_spliced.info.shtml) showing any of the five AS types discussed here or alternative initiation or termination. Because the gene list is not broken up by AS type, we can only estimate overlap with our determination. 2,014 (80%) of the TIGR AS annotated genes are included in our database. Assuming a similar frequency in rice as in Arabidopsis for the five AS types studied here (~74% of all alternatively spliced genes according to (35), using the same computational approach as for rice), we can be confident that also for rice the vast majority of the TIGR annotated alternatively spliced

genes are included in our database. Thus, overall we estimate that our AS determination criteria have a false negative error of less than 2%, with errors caused by wrong strand predictions and some erroneous splice site predictions for non-canonical introns using GeneSeqer default parameters. It is hard to determine the false positive error rate of our method without large-scale experimental proof, but because these errors would derive from the same problems with automated spliced alignment causing false negatives, it should be expected to fall within the same bounds. Thus, we estimate the occurrence of AS in both Arabidopsis and rice at about 20-25% of all genes.

Intron Retention is the Most Prevalent AS Type in both Arabidopsis and Rice. Among the five AS types in Arabidopsis and rice, Table 1 indicates intron retention (IntronR) as the most prevalent type (more than 50% of AS events), followed by alternative acceptor site use (AltA). Alternative position (AltP) is the least prevalent type. IntronR in Arabidopsis was recently demonstrated to be a bona fide AS event instead of merely the product of incomplete splicing in a large scale study using ribosome-associated mRNAs as template for RT-PCR (32). Compared with AS events in human, where ExonS is the most abundant (4), the high frequency of IntronR in plants may reflect distinct features of plant pre-mRNA splicing. In particular, the observed frequency differences are consistent with a model in which plant introns are primarily recognized by intron definition, whereas in mammalians exon definition is the predominant mechanism. Assuming failure of splice site recognition occurs at a single site, such failure would mostly lead to exon skipping in mammalians but to intron retention in plants. Comparing Arabidopsis and rice, we note a prevalence of relatively long introns in

rice and a ~5% higher incident rate of ExonS in rice. We propose that exon definition is a more prominent mechanism of intron recognition in rice than it is in Arabidopsis.

Position of Alternatively Spliced Introns and the Effects of AS Events on mRNA Translation. Depending on the position of the alternative splice sites, AS may have no, little, or dramatic effect on the translation product of the resulting transcript isoform relative to the constitutive transcript. Because our data are derived from EST/cDNA sampling, some of the observed AS events may represent inefficient splicing, with marginal if any physiological role. To estimate such effect, we derived putative transcript isoforms (PTIs) for each gene from overlapping EST/cDNA alignments (see Methods and Materials). Comparison of these (full-length) PTIs yielded a reliable set of 4,922 and 6,089 AS events in Arabidopsis and rice, respectively, which were used in the following to classify AS events with respect to position, effect on translation, and conservation. In Arabidopsis, 746 (15.2%) of these AS events locate in the 5'-UTR and 315 (6.4%) in the 3'-UTR. The remaining 3,861 events (78.4%) locate within or overlap with the constitutive open reading frame (ORF). To facilitate this large-scale analysis, the constitutive ORF was defined as the longest ORF in the PTI with most abundant EST/cDNA evidence. In some cases, this definition is ambiguous, but this does not affect the classification of AS events because our classification is based on pairwise comparison of PTIs. Of the non-UTR AS events, 807 (16.4%) do not change the start and stop codon positions of the longest ORF in the two PTIs being compared ("read-through" events), whereas the remaining 3,054 events (62.0%) produce either an upstream stop codon (implicating a truncated protein product; 1526 events, 31%), a downstream stop codon (implicating an extended protein product; 217 events, 4.4%), an

upstream start codon (204 events, 4.1%), a downstream start codon (1,040 events, 21.1%), or change both codons (altered frame; 65 events, 1.3%). Note that the indication of an alternative downstream start codon merely refers to the position of the longest ORF in the PTI. Physiologically, presumably the native start codon is recognized and a truncated protein produced. Thus, in the following, all AS events that are not read-through are lumped as "altered-ORF" type. In rice, about 30% of the identified AS events locate in the UTRs, 14% are read-through, and the remaining 56% produce changes in the location of start or stop codons.

As shown in *Supporting Table 5*, IntronR produces the highest frequency of altered-ORF AS outcomes (71% in Arabidopsis and 60% in rice) and the lowest frequency of read-through (7% in Arabidopsis and 10% in rice), with other AS types producing 47%-56% altered-ORF and 12%-34% read-through outcomes. The higher ratio of altered-ORF outcomes in IntronR is unlikely to result from false positive IntronR cases due to genomic contamination and incomplete splicing. A subset of reliable Arabidopsis IntronR cases (where both spliced and retained isoforms are each supported by more than two ESTs/cDNAs) also showed ~70% altered-ORF outcomes (data not shown). Interestingly, the read-through ratio for IntronR was higher in rice than in Arabidopsis, while other rice AS types had lower read-through ratio.

More than One Third of AS Events may be Coupled with NMD. More than half of the observed AS events generate a premature termination codon (PTC), which renders the alternative mRNA isoforms possible candidates for nonsense-mediated decay (NMD). PTCs in human were defined as in-frame stop codons residing more than 50 bp upstream of the 3'-

most exon-exon junction (46). *Supporting Figure 2* shows a histogram of the distances between the stop codon and last exon-exon junction in a set of 4,868 Arabidopsis genes, which have their ORF annotation fully supported by EST/cDNA evidence. 4,675 of these genes (96%) have their stop codon in the last exon. For 117 of the remaining 193 genes, the distance between stop codon and the last exon-exon junction is less than 50bp. Thus, overall more than 98% of Arabidopsis genes fit the human pattern of stop codon positioning close to or downstream of the last exon-exon junction. Arabidopsis homologs of all human proteins involved in NMD and the exon junction complex have been identified (22). Therefore, we treat as potential substrates for NMD all AS isoforms producing PTCs defined as in-frame stop codons >50 bp upstream of the 3'-most exon-exon junction.

As shown in *Supporting Table 5*, about 42% Arabidopsis AS events and 36% rice events produce NMD candidates. These frequencies are similar to frequencies observed in human (11). Among the five types of AS events, AltP has the lowest incidence of NMD (18%-22% of all AltP events), IntronR has the highest (40-48%), and AltA, AltD and ExonS have rates around 34%-43%. NMD is a surveillance mechanism that removes erroneous mRNAs containing PTCs (47). If indeed more than one third of AS events will generate NMD candidates in plants as in humans, AS will produce much less protein diversity than might be expected. The NMD mechanism can also be routinely used by the cell to regulate gene expression (48), therefore coupling of AS and NMD could be an important post-transcriptional regulation to adjust the level of transcript isoforms (2).

Gene Ontology (GO) Analysis of Alternatively Spliced Genes. Approximate functional categorization of the AS genes was based on Arabidopsis gene ontology annotations (49).

4,661 (98%) of the 4,768 Arabidopsis alternatively spliced genes have GO annotation. For rice, all predicted rice proteins were first matched against the Arabidopsis proteins using BLASTP. The GO annotation of the most significant Arabidopsis hit for each rice gene was used in the analysis, yielding GO assignments for a total of 23,403 of the rice genes (40.4%) and 5,055 of the alternatively spliced genes (76%). TAIR also gives GOslim terms to summarize the GO annotation (49). We calculated the AS frequency for protein sets grouped by GOslim and GO terms to identify any functional categories with unexpected AS frequencies.

In cell component annotations, the over-represented and under-represented GOslim terms are very similar in Arabidopsis and rice. The most over-represented terms are "Golgi apparatus", "plastid", "other cellular components", "chloroplast" and "cytosol". Their AS ratios in Arabidopsis are 29.7%, 28.0%, 25.6%, 23.3% and 22.4%, respectively. In rice, all these five terms are also the most over-representeded, with AS ratios of 29.9%, 27.7%, 29.5%, 28.4% and 28.2%, respectively. On the other hand, "cell wall" and "extracellular" are the most under-represented terms in both Arabidopsis and rice. As shown in *Supporting Figure 3*, the overall trend in the AS ratios of different terms is consistent between Arabidopsis and rice, with a correlation coefficient of 0.91. Among the five AS types, the alternative acceptor site (AltA) ratio in different GO terms is the most consistent between the two species, with correlation coefficient of 0.92.

In molecular functions and biological process annotation, however, the over-represented and under-represented terms are not consistent between Arabidopsis and rice. Some over-represented terms in one species are the most under-represented terms in another species. For instance, the most under-represented biological process term in Arabidopsis is "DNA or

RNA metabolism", with an AS ratio of only 9%. In rice, however, it is the most over-represented term with a 27.7% AS ratio. Similarly, the molecular function terms "receptor binding or activity" and "nucleic acid binding" are the most under-represented in Arabidopsis but the most over-represented terms in rice. No clear correlation was found between Arabidopsis and rice, partly due to the existence of many GO terms with multiple GOslim term assignments.

Detailed comparisons for each GO term revealed many consistent groups between Arabidopsis and rice. For the over-represented groups, the cell-component GO term "nuclear speck" and "light-harvesting complex" identify the most over-represented groups in both Arabidopsis and rice. Their AS ratios are about three-fold higher than the overall AS ratio. Interestingly, all the 13 genes targeted to Arabidopsis nuclear speck are SR proteins functioning in splicing, and about 10 of them can be alternatively spliced. For the under-represented groups, the GO terms without alternative splicing events were not considered. "DNA transposition" in biological process and "ribonuclease H activity" in molecular function annotation are the most under-represented groups in both Arabidopsis and rice. More than 60-250 genes belong to these groups, with only one gene showing AS in each group. *Supporting Figure 4* displays the most over- and under-represented biological terms that are consistent in Arabidopsis and rice. Interestingly, many metabolism and biosynthesis processes are over-represented, while response to pathogen is under-represented.

Conserved Alternatively Spliced Genes. The GO annotation analysis indicated possible conservation of AS in plants. By searching Arabidopsis and rice close homolog pairs generated from reciprocal BLAST best hits (see Methods and Materials), 1,988 (41.7%) of

the 4,768 alternatively spliced genes in Arabidopsis were found to have close homologs in rice that are also alternatively spliced. Among the five AS types, IntronR is most conserved, with about 30% of the intron-retaining genes in Arabidopsis also showing IntronR in rice (Table 2). By construction, the conserved alternatively spliced genes do not necessarily conserve the same AS event. For instance, Arabidopsis gene At1g28570 and rice gene LOC_Os01g46120 (GDSL-motif lipases) have similar gene structure and both have ExonS in the ORFs. In the Arabidopsis gene, the second exon is skipped, while in the rice gene the fourth exon is skipped. Details about these conserved genes are available from the ASIP database (http://www.plantgdb.org/ASIP/EnterDB.php).

To identify conserved AS events, we first searched for conserved intron pairs as described in Methods and Materials. A total of 4,142 genes containing 7,640 conserved introns were identified in Arabidopsis, including 352 genes (445 introns) that are alternatively spliced. The frequency of genes containing conserved introns being alternatively spliced is about 8.5%, much lower than the overall AS frequency (~22%). As shown in Table 2 and *Supporting Table 6*, within this set of genes we could identify 41 Arabidopsis genes with AS events that are conserved in rice (same type of AS event for the conserved intron pair). 11 of these events (26.8%) are read-through, and 22 (53.6%) produce NMD-candidates.

The intron-pair method can only identify conserved introns in coding regions where the flanking exons are well conserved between Arabidopsis and rice. Therefore, we also visually compared a large number of homologous gene structures in Arabidopsis and rice and selected those with the same intron alternatively spliced. A small portion of additional conserved AS events were identified in this way, including U1-70K (U1 snRNP – 70K, At3g50670 and

LOC_Os10g02630), Tra2/SFRS10 (Transformer-2-beta, At1g07350 and LOC_Os03g15890), and PTB (polypyrimidine tract binding protein, At3g01150, At5g53180 and LOC_Os01g43170). For most of the homologous gene pairs with conserved AS events, the AS event has similar effects on translation. For example, both Arabidopsis and rice IDH2 genes have three introns and can utilize an alternative acceptor site 12 bases downstream for splicing the first intron. A protein with four amino acid residues (VITK) removed will be generated from the alternative mRNA isoform in both Arabidopsis and rice. The PTB genes in both rice and Arabidopsis show transcripts, which include an additional exon in their third introns and produce PTC-containing mRNA isoforms. The mammalian PTB gene can autoregulate its splicing by promoting the skipping of exon 11, which introduces a PTC and leads to NMD (50). This regulation pathway may be conserved in plants.

Interestingly, 30 conserved intron-pairs and many more conserved gene pairs have different AS types in Arabidopsis and rice. For instance, At3g60370 and LOC_Os07g30800 (FKBP-type peptidyl-prolyl cis-trans isomerase) are orthologs with similar gene structure. Splicing of the third intron in At3g60370 can use an alternative acceptor site that generates a PTC-containing isoform. Splicing of the third intron in LOC_Os07g30800 alternatively includes an additional exon, which also produces a NMD candidate. Although the two genes have different AS types, it may be that the function of their AS events is conserved.

Experimental Validation of Alternative Splicing. A total of 12 genes with previously unreported alternative splicing events were checked by RT-PCR. A mixture of total RNA from root, leaf, stem, flower, and silique tissues was used to reduce the possibility of under-reporting AS due to tissue-specific splicing patterns. Primers were designed either from the

AS region (atExp2 and IDH2) or from the flanking regions. PCR products were submitted to direct sequencing. We were able to confirm AS events in nine of the 12 genes. As shown in Figure 3, for the first two genes (atExp2 and IDH2, lane 1-4), PCR was successful using splicing-specific primers. Multiple bands of expected size were observed in another seven genes (PGAT1, PTB2a, SRp40, SCL30, SRM300, ZNF-RING, and KH-NOVA). Sequencing results confirmed the expected alternative splicing events. It is interesting to note the existence of extra bands in many genes, which may rise from other splicing isoforms or from non-specific amplification. In fact, extra bands of expected size were also observed for Exp1 and UDP-GT, but we were not able to direct sequence them due to their low abundance.

Conclusions

Clearly, not all of the thousands of AS events suggested by large-scale EST/cDNA alignments will be biologically functional. Aberrant splicing will inevitably occur during the complicated dynamic splicing process that is happening continuously and under varied physiological conditions. Most aberrant splicing (splicing errors) will be removed by mRNA surveillance mechanisms such as NMD, and thus are neutral to the organism. In the course of evolution, some splicing variants may be selectively beneficial and thus can be fixed as functional AS events.

How to distinguish splicing errors from biologically functional AS events is an unresolved question so far. Interruption of protein coding sequences and/or production of PTC are not good landmarks for splicing error, as the coupling of AS and NMD may be regulated and add another level of regulation to gene expression. Conservation of AS events

seems to be a good indication of functional AS events. Several papers addressed conserved cassette exons between human and mouse (12, 15, 17). We have shown that one fifth of genes undergo alternative splicing in both Arabidopsis and rice and identified a small portion of conserved events. The web-interfaced ASIP database can serve as a starting point for the community to identify functional AS events in plants.

Acknowledgments

We would like to thank Wei Huang for help with statistical analysis and Robert Fluhr for critical reading of the manuscript. This work was supported in part by NSF grants DBI-0110189 and DBI-0321600 and Research Grant No. IS-3454-03 from BARD, the United States – Israel Binational Agricultural Research and Development Fund.

References

1. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A. & Soreq, H. (2005) *Gene* **344,** 1-20.
2. Lareau, L. F., Green, R. E., Bhatnagar, R. S. & Brenner, S. E. (2004) *Curr Opin Struct Biol* **14,** 273-82.
3. Black, D. L. (2003) *Annu Rev Biochem* **72,** 291-336.
4. Gupta, S., Zink, D., Korn, B., Vingron, M. & Haas, S. A. (2004) *Bioinformatics.*
5. Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001) *Nucleic Acids Res* **29,** 2850-9.
6. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. & Bork, P. (2000) *FEBS Lett* **474,** 83-6.

7. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. (1999) *Genome Res* **9,** 1288-93.

8. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. (2001) *Genome Res* **11,** 889-900.

9. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. (2002) *Nat Genet* **30,** 29-30.

10. Modrek, B. & Lee, C. (2002) *Nat Genet* **30,** 13-9.

11. Lewis, B. P., Green, R. E. & Brenner, S. E. (2003) *Proc Natl Acad Sci U S A* **100,** 189-92.

12. Sorek, R., Shamir, R. & Ast, G. (2004) *Trends Genet* **20,** 68-71.

13. Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R. & Blencowe, B. J. (2005) *Trends Genet* **21,** 73-7.

14. Thanaraj, T. A., Clark, F. & Muilu, J. (2003) *Nucleic Acids Res* **31,** 2544-52.

15. Sugnet, C. W., Kent, W. J., Ares, M., Jr. & Haussler, D. (2004) *Pac Symp Biocomput,* 66-77.

16. Nurtdinov, R. N., Artamonova, II, Mironov, A. A. & Gelfand, M. S. (2003) *Hum Mol Genet* **12,** 1313-20.

17. Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. (2005) *Proc Natl Acad Sci U S A.*

18. Lorkovic, Z. J., Wieczorek Kirk, D. A., Lambermon, M. H. & Filipowicz, W. (2000) *Trends Plant Sci* **5,** 160-7.

19. Reddy, A. S. N. (2001) *Critical Rev Plant Sci* **20,** 523-571.

20. Ko, C. H., Brendel, V., Taylor, R. D. & Walbot, V. (1998) *Plant Mol Biol* **36,** 573-83.

21. Brown, J. W., Smith, P. & Simpson, C. G. (1996) *Plant Mol Biol* **32,** 531-5.

22. Wang, B. B. & Brendel, V. (2004) *Genome Biol* **5,** R102.

23. Golovkin, M. & Reddy, A. S. (1996) *Plant Cell* **8,** 1421-35.

24. Lazar, G. & Goodman, H. M. (2000) *Plant Mol Biol* **42,** 571-81.

25. Montag, K., Salamini, F. & Thompson, R. D. (1995) *Nucleic Acids Res* **23,** 2168-77.

26. Macknight, R., Bancroft, I., Page, T., Lister, C., Schmidt, R., Love, K., Westphal, L., Murphy, G., Sherson, S., Cobbett, C. & Dean, C. (1997) *Cell* **89,** 737-45.

27. Dinesh-Kumar, S. P. & Baker, B. J. (2000) *Proc Natl Acad Sci U S A* **97,** 1908-13.

28. Werneke, J. M., Chatfield, J. M. & Ogren, W. L. (1989) *Plant Cell* **1,** 815-25.

29. Baga, M., Glaze, S., Mallard, C. S. & Chibbar, R. N. (1999) *Plant Mol Biol* **40,** 1019-30.

30. Zhou, Y., Zhou, C., Ye, L., Dong, J., Xu, H., Cai, L., Zhang, L. & Wei, L. (2003) *Genomics* **82,** 584-95.

31. Kazan, K. (2003) *Trends Plant Sci* **8,** 468-71.

32. Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. & Fluhr, R. (2004) *Plant J* **39,** 877-85.

33. Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A. & Shinozaki, K. (2004) *Nucleic Acids Res* **32,** 5096-103.

34. Zhu, W., Schlueter, S. D. & Brendel, V. (2003) *Plant Physiol* **132,** 469-84.

35. Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L. & White, O. (2003) *Nucleic Acids Res* **31,** 5654-66.

36. Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., Jia, P., Zhao, Q., Ying, K., Yu, S., Tang, Y., Weng, Q., Zhang, L., Lu, Y., Mu, J., Zhang, L. S., Yu, Z., Fan, D., Liu, X., Lu, T., Li, C., Wu, Y., Sun, T., Lei, H., Li, T., Hu, H., Guan, J., Wu, M., Zhang, R., Zhou, B., Chen, Z., Chen, L., Jin, Z., Wang, R.,

Yin, H., Cai, Z., Ren, S., Lv, G., Gu, W., Zhu, G., Tu, Y., Jia, J., Chen, J., Kang, H., Chen, X., Shao, C., Sun, Y., Hu, Q., Zhang, X., Zhang, W., Wang, L., Ding, C., Sheng, H., Gu, J., Chen, S., Ni, L., Zhu, F., Chen, W., Lan, L., Lai, Y., Cheng, Z., Gu, M., Jiang, J., Li, J., Hong, G., Xue, Y. & Han, B. (2002) *Nature* **420,** 316-20.

37. Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., Antonio, B. A., Kanamori, H., Hosokawa, S., Masukawa, M., Arikawa, K., Chiden, Y., Hayashi, M., Okamoto, M., Ando, T., Aoki, H., Arita, K., Hamada, M., Harada, C., Hijishita, S., Honda, M., Ichikawa, Y., Idonuma, A., Iijima, M., Ikeda, M., Ikeno, M., Ito, S., Ito, T., Ito, Y., Iwabuchi, A., Kamiya, K., Karasawa, W., Katagiri, S., Kikuta, A., Kobayashi, N., Kono, I., Machita, K., Maehara, T., Mizuno, H., Mizubayashi, T., Mukai, Y., Nagasaki, H., Nakashima, M., Nakama, Y., Nakamichi, Y., Nakamura, M., Namiki, N., Negishi, M., Ohta, I., Ono, N., Saji, S., Sakai, K., Shibata, M., Shimokawa, T., Shomura, A., Song, J., Takazaki, Y., Terasawa, K., Tsuji, K., Waki, K., Yamagata, H., Yamane, H., Yoshiki, S., Yoshihara, R., Yukawa, K., Zhong, H., Iwama, H., Endo, T., Ito, H., Hahn, J. H., Kim, H. I., Eun, M. Y., Yano, M., Jiang, J. & Gojobori, T. (2002) *Nature* **420,** 312-6.

38. Brendel, V., Xing, L. & Zhu, W. (2004) *Bioinformatics* **20,** 1157-69.

39. Kornblihtt, A. R. (2005) *Curr Opin Cell Biol* **17,** 262-8.

40. Dong, Q., Schlueter, S. D. & Brendel, V. (2004) *Nucleic Acids Res* **32** Database issue, D354-9.

41. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res* **25,** 3389-402.

42. Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Huang, X., Su, Z., Tong, W., Tong, Z., Ye, J., Wang, L., Lei, T., Chen, C., Chen, H., Huang, H., Zhang, F., Li, N., Zhao, C., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Hu, W., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wong, G. K. & Yang, H. (2005) *PLoS Biol* 3, e38.

43. Goodall, G. J. & Filipowicz, W. (1991) *Embo J* **10,** 2635-44.

44. Arabidopsis Genome Initiative (2000) *Nature* **408,** 796-815.

45. Latijnhouwers, M. J., Pairoba, C. F., Brendel, V., Walbot, V. & Carle-Urisote, J. C. (1999) *Plant Mol Biol* **41,** 637-44.

46. Nagy, E. & Maquat, L. E. (1998) *Trends Biochem Sci* **23,** 198-9.

47. Baker, K. E. & Parker, R. (2004) *Curr Opin Cell Biol* **16,** 293-9.

48. Lejeune, F. & Maquat, L. E. (2005) *Curr Opin Cell Biol* **17,** 309-15.

49. Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. & Rhee, S. Y. (2004) *Plant Physiol* **135,** 745-55.

50. Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. (2004) *Mol Cell* **13,** 91-100.

Figure Legends and Tables

Figure 1. Visualization of five alternative splicing types. The top black line represents the genome sequence. Filled boxes and arrows indicate exons. Thin lines connecting the boxes indicate introns. The green lines represent introns that may be retained. The open green box represents a skipped exon. Vertical bars represent the alternative donor/acceptor sites. In the ASIP database, different donor and acceptor sites are denoted by different colors.

Figure 2. Comparison among the TIGR, RIKEN, and our (ASIP) data sets of alternatively spliced genes. Numbers represent sizes of the indicated gene sets. TIGR represents data from (35), and RIKEN represents data from (33). ASIP data are from this study.

Figure 3. RT-PCR validation of AS events in Arabidopsis. Primers were designed either from the splicing junctions of each isoform (atExp2 and IDH2) or from flanking exons (remaining genes). Arrows point to the PCR products of the target size. Direct sequencing was performed to confirm the AS events.

Table 1. Alternative splicing events and genes in Arabidopsis and rice

AS type	Arabidopsis		Rice	
	Events	Genes	Events	Genes
AltD	873 (10.4%)	745 (3.44%)	1,642 (11.1%)	990 (3.2%)
AltA	1,828 (21.7%)	1,464 (6.76%)	2,201 (14.9%)	1,699 (5.5%)
AltP	316 (3.7%)	207 (0.96%)	921 (6.2%)	562 (1.8%)
ExonS	687 (8.2%)	390 (1.80%)	2,004 (13.5%)	999 (3.2%)
IntronR	4,719 (56.0%)	3,139 (14.5%)	8,057 (54.3%)	4,626 (14.9%)
Total	**8,423**	**4,768 (22.0%)**	**14,825**	**6,638 (21.5%)**

AltD: Alternative donor site; AltA: Alternative acceptor site; AltP: Alternative position; ExonS: Exon skipping; IntronR: Intron retention. Percentages in the events columns represent the proportion of certain AS types relative to the total number of AS events. Percentages in the genes columns indicate the frequency of the AS type in all expressed genes studied (21,641 for Arabidopsis and 30,917 for rice).

Table 2. Conserved alternatively spliced genes in Arabidopsis and rice.

AS type	*Conserved genes#*	*Conserved introns**	*Conserved events* [$]
AltD	108 (14.5%)	61	5
AltA	259 (17.7%)	117	5
AltP	9 (4.3%)	21	5
ExonS	47 (12.1%)	4	1
IntronR	951 (30.3%)	263	29
Total	1154 (24.2%)	445	41

Conserved genes indicate the number of Arabidopsis genes showing any type of AS which have rice homologs also showing the same AS type. The percentages in parentheses indicate the fractions of genes with certain AS type relative to the total number of 4,768 Arabidopsis alternatively spliced genes.

* Numbers in the conserved introns column indicate the numbers of conserved introns between Arabidopsis and rice showing a certain type of AS in Arabidopsis.

[$] Conserved events represent the numbers of conserved introns in Arabidopsis showing the same type of AS in rice.

Figure 1

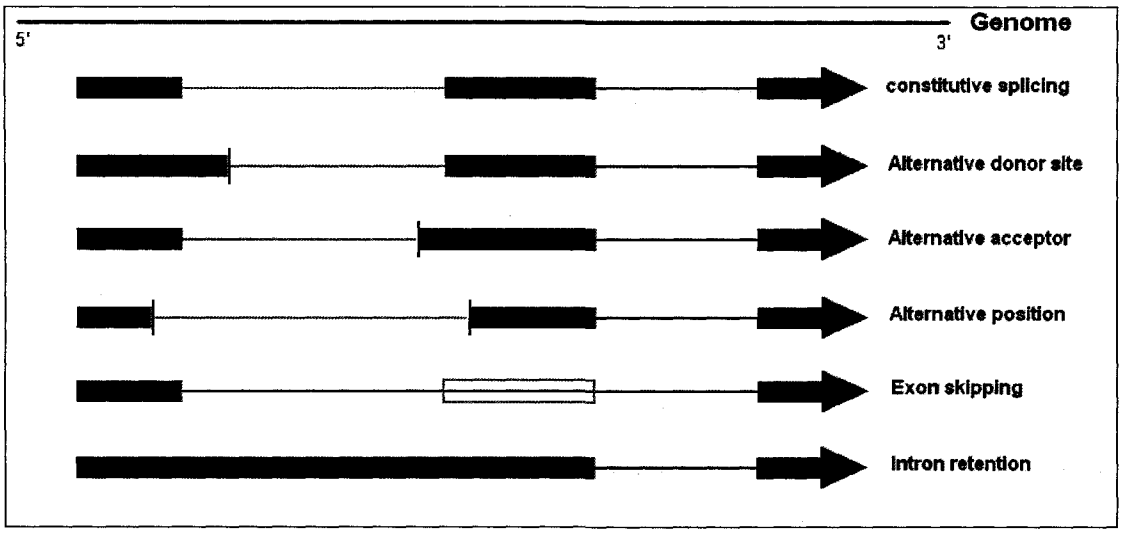

Genome
5'
3'
constitutive splicing
Alternative donor site
Alternative acceptor
Alternative position
Exon skipping
Intron retention

Figure 2

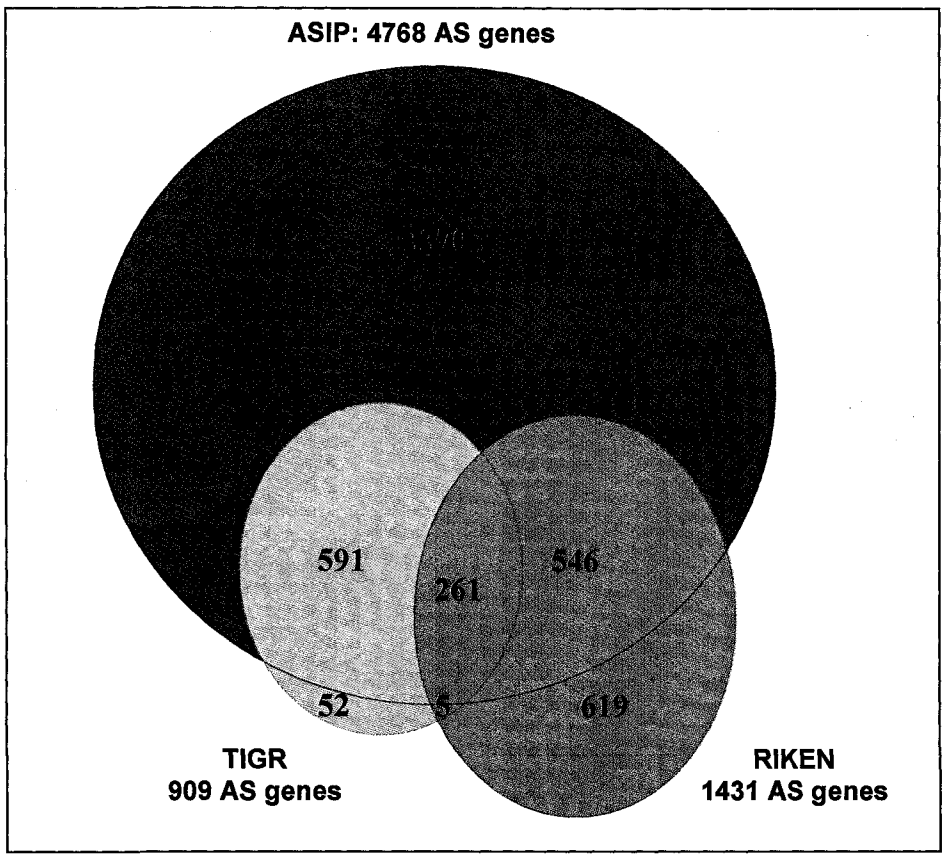
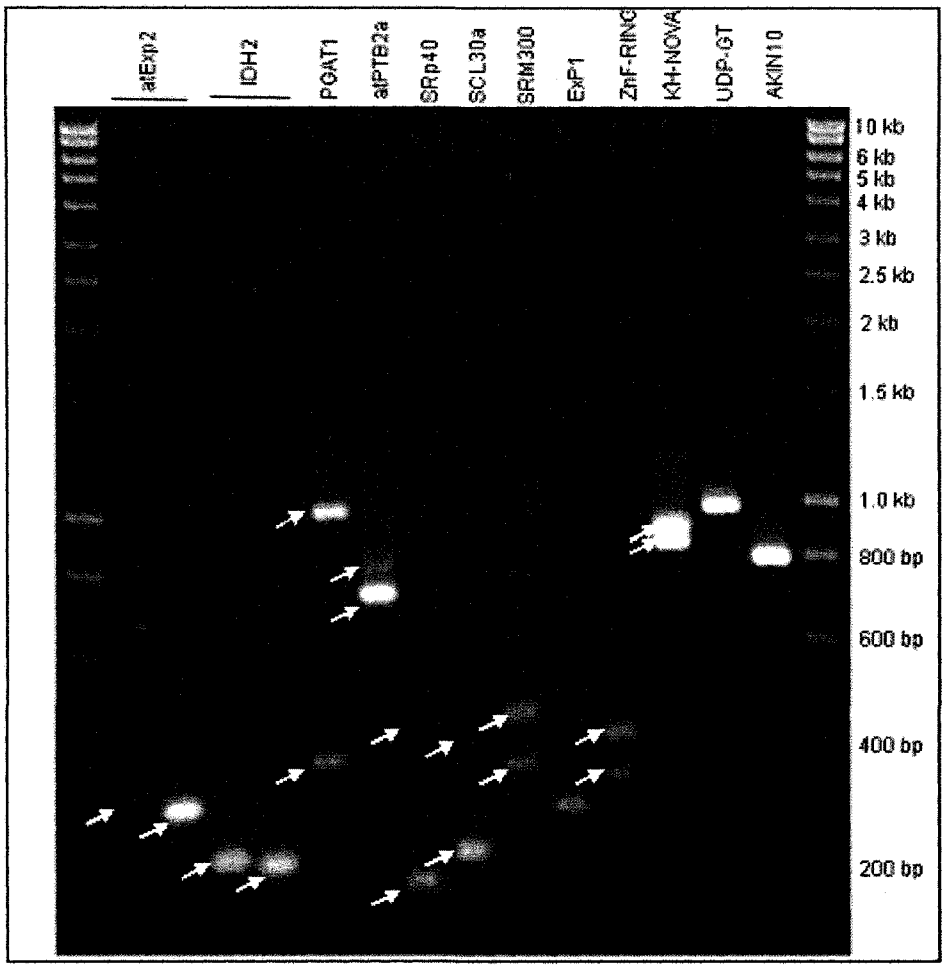
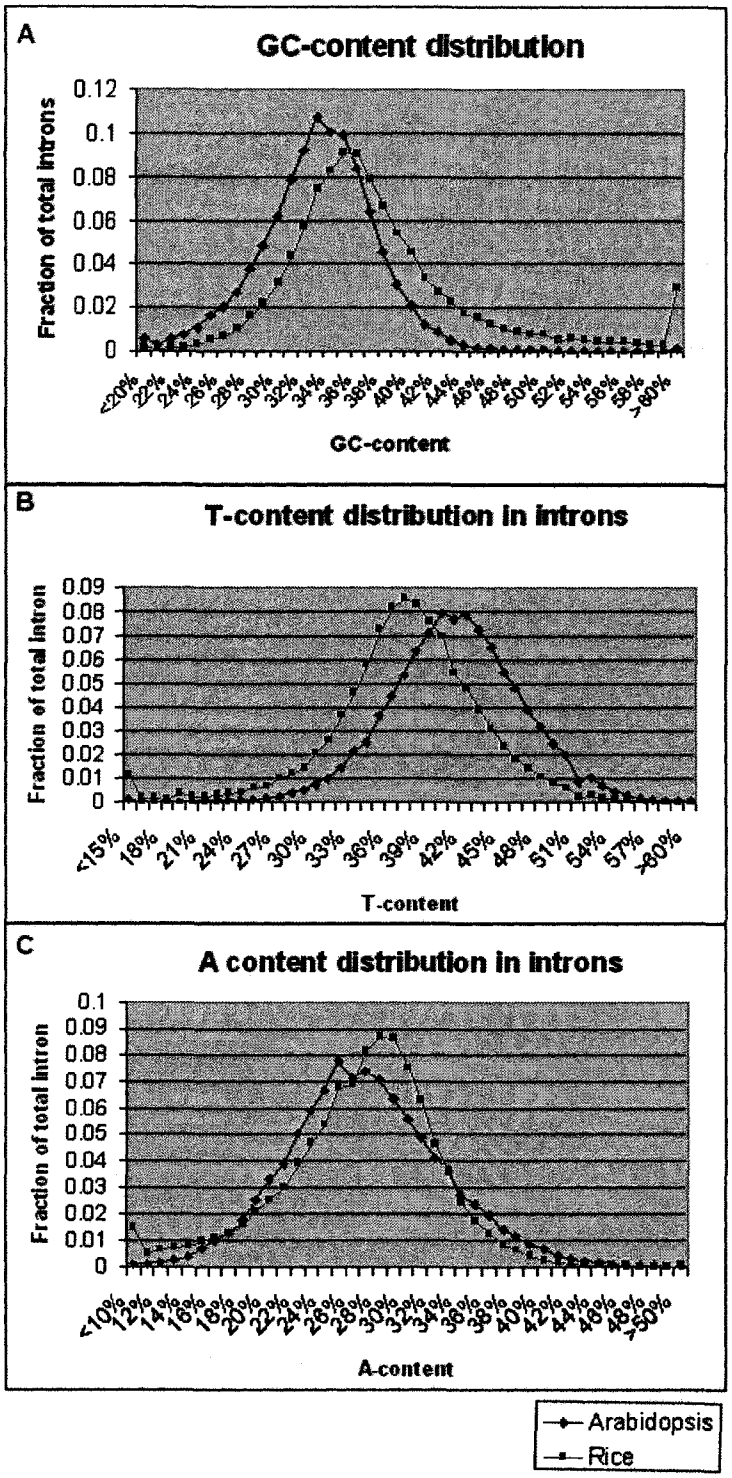

ASIP: 4768 AS genes
591
261
546
52
5
619
TIGR
909 AS genes
RIKEN
1431 AS genes

Figure 3


atExp2
ICH2
PGAT1
atPTB2a
SRp40
SCL30a
SRM300
ExP1
ZnF-RING
KH-NOVA
UDP-GT
AKIN10
10 kb
6 kb
5 kb
4 kb
3 kb
2.5 kb
2 kb
1.5 kb
1.0 kb
800 bp
600 bp
400 bp
200 bp

Supporting Data

Supporting Figure 1. Arabidopsis introns are more U-rich than rice introns.

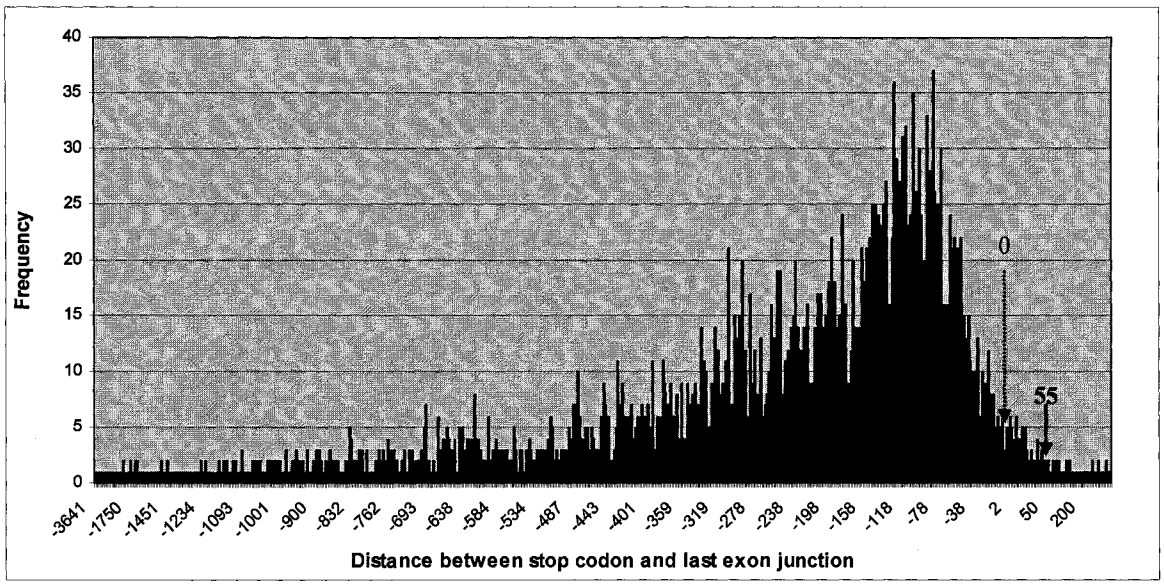


A. The median GC-content of Arabidopsis introns is ~5% lower than that of rice introns.

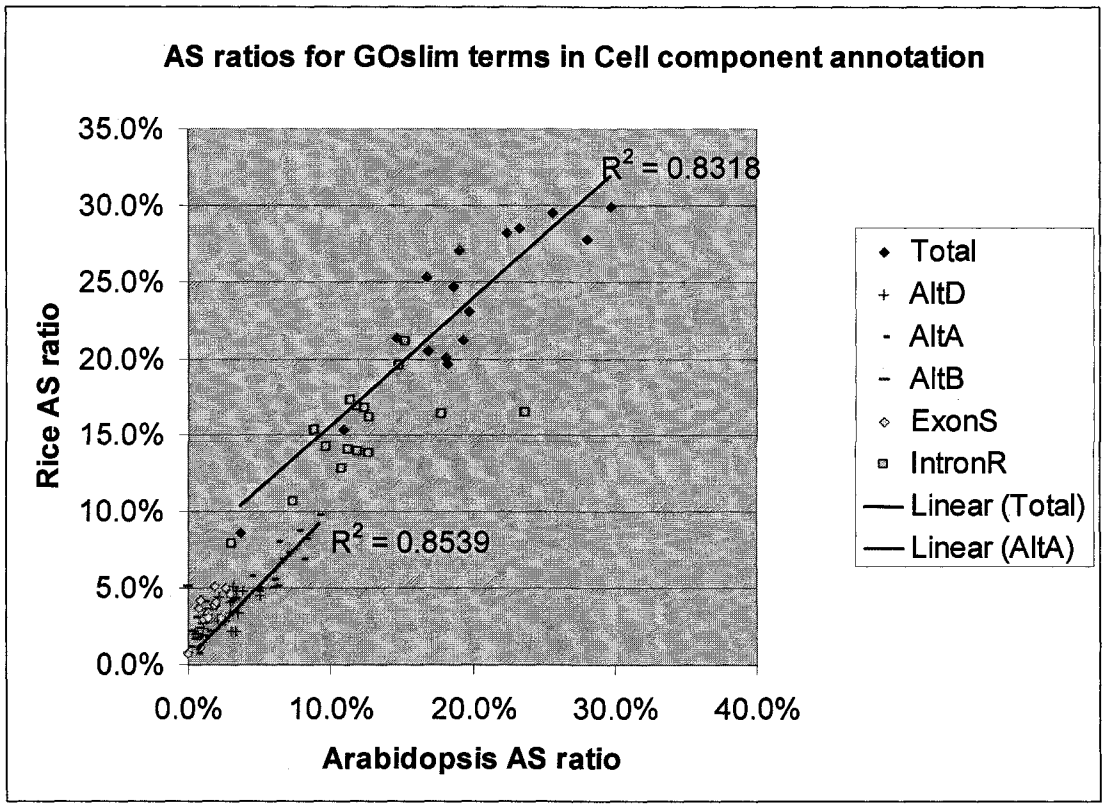
B. The T-content of Arabidopsis introns is higher than that of rice introns, demonstrating that Arabidopsis introns are more U-rich.

C. The A-content is similar in the two plants.

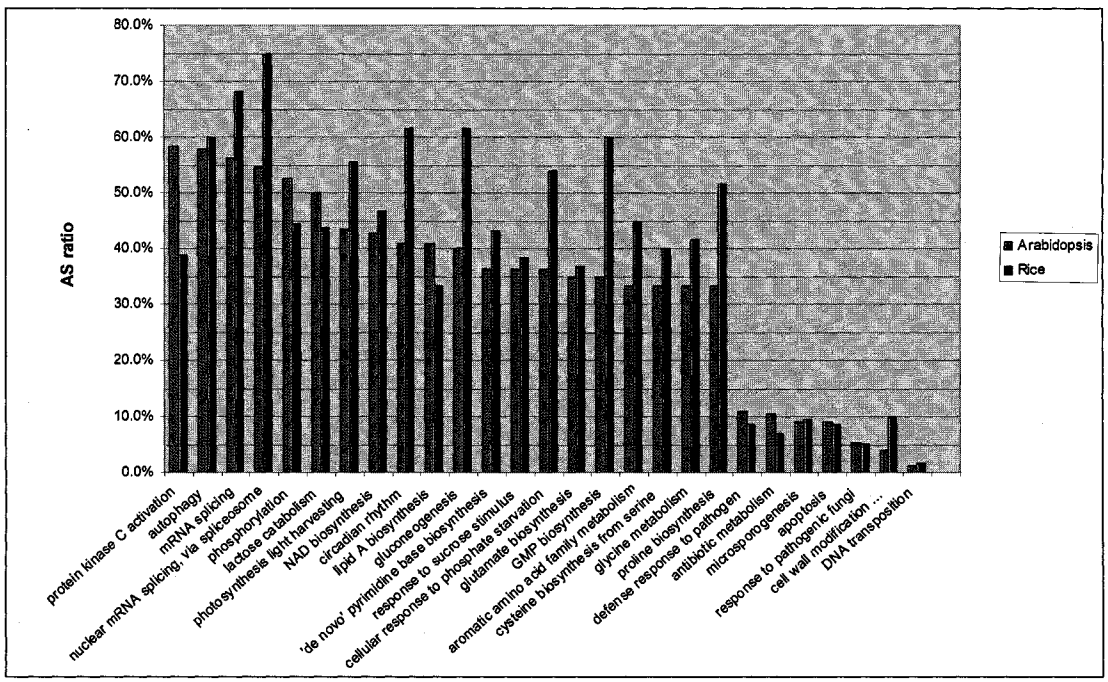
Supporting Figure 2. Frequency distribution of distances between the stop codon and the last exon-exon junction in Arabidopsis genes. The x-axis (Distance) represents different distances between the stop codon and the last exon-exon junction in about 5,000 selected Arabidopsis genes. The y-axis (Frequency) represents the number of genes with distances in the indicated range. Red arrows indicate distances discussed in the text.



Supporting Figure 3. Cell component GOslim term annotation of AS genes in Arabidopsis and rice. AS ratio indicates the frequency of alternatively spliced genes relative to the total number of genes in a certain GOslim term category. Ratios are plotted for each AS type and overall AS events (Total). A trend line is drawn for categories with significant linear regression (coefficient of determination indicated as R^2).



Supporting Figure 4. Over- and under-represent biological process GO terms in both Arabidopsis and rice AS genes. The selected GO-terms of biological process are shown along the x-axis. The y-axis indicates the fraction of alternatively spliced genes relative to all genes in the corresponding GO-term group. The rightmost seven GO-terms are under-represented terms, and the remaining groups are over-represented terms for alternatively spliced genes in both plant species.



Supporting Table 1. RT-PCR primers for detecting alternative splicing.

Name	Gene	Annotation	SItype	LeftPrimer (-F)	RightPrimer (-R)	Size
atExp2	At1g28100	Expressed protein	AltD	F1:TTTCAAAAGTTAGTG ACACTTTGTTC F2:AAGATTTTCAAAAA ACATTACTGC	ATACAACCGCAGG TCTCACC	345 326
IDH2	At2g17130	isocitrate dehydrogenase subunit 2 / NAD+ isocitrate dehydrogenase subunit 2	AltA	F1:GAGCCTTAAGGTGA TTACAAAG F2:GAGAGCCTTAAGTT CTGTTC	TGACATCGAATTG CTCTGGT	248 238
atPGAT1	At1g75020	phospholipid/glycerol acyltransferase family protein	ExonS	CCTCTTTTGTCTTCGCC AAC	TTCCCCTGAGCAT TGAAATC	383 (981)
atPTB2a	At3g01150	polypyrimidine tract-binding protein	AltA, ExonS	TCCCTGGAAATGTCCT CTTG	TGCCACAAGTAGG CTTAGATCA	824 (777, 722, 675)
atSRp40	At4g25500	arginine/serine-rich splicing factor RSP40 (RSP40)	ExonS, IntronR	ACTACGCCTGCCAAAA TCAT	ATTCAAAGCGGTC AAGTGCT	210 (413, 992)
atSCL30a	At3g13570	SC35-like splicing factor, 30a kD (SCL30a)	IntronR, ExonS	TGGTTCGCAACTTACG TCAT	AGTTGGCTTCTTCC GGTTTT	245 (406, 970)
SRM300	At1g07350	transformer serine/arginine-rich ribonucleoprotein, putative	ExonS, IntronR	CTTATGACAAGCGTCG TGGA	CAGCCCCAAGTAC TTTCCTG	368 (464, 768)
Exp1	At4g05590	Expressed protein	IntronR	CGTTCAAGTGGGGTAT AAGCA	TGCAAGAAAGATG GGGTTTT	300 (415)
ZnF-RING	At3g23280	Zinc finger (C3HC4-type RING finger) family protein / ankyrin repeat family protein	ExonS	GAACCAAAAGCGAAG CAGTC	AAGGCACTTGCTT TTCCTGA	347 (419)
KH-NOVA	At5g04430	KH-domain-containing protein NOVA, putative	AltA	GAAGCACCGATACGAC GATT	CGTCCAAGCACCA ATCCTAT	862 (925)
UDP-GT	At1g24100	UDP-glucoronosyl/UDP-glucosyl transferase family	AltB	TCCTTGGGGACTTGAA GTTG	TCACTCATTGCCTT CACAGC	322, (981)
AKIN10	At3g01090	Snf1-related protein kinase (AKIN10)	AltD	GAATTTTCTCCTCCGCC TTT	CTCGTTTCATGGG GTCAACT	812 (1137)

Supporting Table 2. EST/cDNAs alignments against the Arabidopsis and rice genomes.

	Arabidopsis	Rice
EST/cDNAs/Total	323,340 / 62,009 / **385,349**	298,857 / 32,136 / **330,993**
Aligned to genome	307,510 / 61,708 / **369,218**	253,758 / 30,058 / **283,816**
Not aligned	15,830 / 301 / 16,131	45,099 / 2,078 / 47,177
Total EST/cDNA cognate alignments	**372,772**	**319,391**
ESTs/cDNAs with Single cognate alignment	366,687 (99.3%)	275,624 (97.1%)
ESTs/cDNAs with Multiple cognate alignments	2,531 (0.7%)	8,192 (2.9%)
Transcript Units (TU)	25,231	36,270
In annotated region	23,856	31,804
In novel region	1,375	4,466
Average EST/cDNAs per TU	14.8	8.8
Expressed genes*	21,641	30,917

*Expressed genes denote all genes with EST/cDNA evidence. We compared the GenBank and TIGR annotated genes with our transcription units (TUs). If multiple TUs overlap with an annotated gene model in the same direction, these TUs are thought to belong to one gene.

Supporting Table 3. Intron and exon statistics for Arabidopsis and rice.

	Arabidopsis	Rice
Introns	94,342	111,343
Average length (bp)	173	433
Median length (bp)	101	160
GC-content	32.7%	37.3%
Long introns (>1kb)	762 (0.8%)	11,541 (10.4%)
Exons	128,098	166,057
Average length of internal exons (bp)	172	193
GC-content	42.4%	48.6%

Supporting Table 4. AS events annotated by TIGR but missed in our data set (ASIP).

Light gray highlights indicate AS events missed by our method. Dark gray highlights indicate AS events found in other genes by our method.

GeneID	TIGR AS_Type	Description	Note/Reason	TIGR-URL	ASIP-URL	ASIP-AS
19 AS genes missed in ASIP						
At2g32160	AltDonAccpt	hypothetical protein	check EST 19846670	TIGR	ASIP	Not found
At2g41240	AltDonAccpt	bHLH protein	check EST 20127106	TIGR	ASIP	Not found
At2g44060	AltDonAccpt	similar to late embryogenesis abundant proteins	check EST 2759340	TIGR	ASIP	Not found
At3g17240	AltDonAccpt	dihydrolipoamide dehydrogenase 2, mitochondrial (lipoamide dehydrogenase 2) (mtlpd2)	check EST 19740800	TIGR	ASIP	Not found
At3g62120	AltDonAccpt	multifunctional aminoacyl-tRNA ligase-like protein	check EST 872012	TIGR	ASIP	Not found
At4g22230	AltDonAccpt	hypothetical protein	check EST 24762212	TIGR	ASIP	Not found
At4g33740	AltDonAccpt	unknown protein	check EST 19877217	TIGR	ASIP	Not found
At5g20720	AltDonAccpt	chloroplast Cpn21 protein	check EST 4127455	TIGR	ASIP	Not found
At5g46160	AltDonAccpt	ribosomal protein L14p family	check EST 16604451	TIGR	ASIP	Not found
At5g62470	AltDonAccpt	MYB96 transcription factor-like protein	check EST 5823334	TIGR	ASIP	Not found
At1g07820	IntronRetain	histone H4	Single Exon Direction	TIGR	ASIP	Not found
At1g48760	IntronRetain	delta-adaptin, putative	Single Exon Direction	TIGR	ASIP	Not found
At1g66160	IntronRetain	expressed protein	Single Exon Direction	TIGR	ASIP	Not found
At3g08940	IntronRetain	putative chlorophyll a/b-binding protein	Single Exon Direction	TIGR	ASIP	Not found
At3g57340	IntronRetain	DnaJ protein family	Single Exon Direction	TIGR	ASIP	Not found
At4g14960	IntronRetain	tubulin alpha-6 chain (TUA6)	GeneSeqer direction wrong	TIGR	ASIP	Not found
At5g06530	IntronRetain	ABC transporter family protein	Single Exon Direction	TIGR	ASIP	Not found
At5g53160	IntronRetain	putative protein	Single Exon Direction	TIGR	ASIP	Not found
At5g58540	IntronRetain	putative protein	Single Exon Direction	TIGR	ASIP	Not found
38 TIGR-AS genes are possible false positive						
At1g18090	AltDonAccpt	expressed protein	No reliable EST	TIGR	ASIP	Not found
At1g59520	AltDonAccpt	expressed protein	No reliabe EST	TIGR	ASIP	Not found
At1g70830	AltDonAccpt	Csf-2-related	No reliable EST	TIGR	ASIP	Not found
At1g74530	AltDonAccpt	unknown protein	No reliable EST	TIGR	ASIP	Not found
At1g74910	AltDonAccpt	ADP-glucose pyrophosphorylase family	No reliable EST	TIGR	ASIP	Not found
At1g79920	AltDonAccpt	heat shock protein hsp70, putative	No reliable EST	TIGR	ASIP	Not found
At2g18240	AltDonAccpt	putative integral membrane protein	No reliable EST	TIGR	ASIP	Not found
At2g21870	AltDonAccpt	putative ATP synthase	No reliable EST	TIGR	ASIP	Not found
At2g39670	AltDonAccpt	expressed protein	No reliable EST	TIGR	ASIP	Not found
At3g02200	AltDonAccpt	expressed protein	No reliable EST	TIGR	ASIP	Not found
At3g02360	AltDonAccpt	6-phosphogluconate dehydrogenase, putative	Terminal exon, Ignored	TIGR	ASIP	Not found
At3g15690	AltDonAccpt	putative acetyl-CoA carboxylase biotin-containing subunit	No reliable EST	TIGR	ASIP	Not found
At3g18860	AltDonAccpt	WD-40 repeat protein family	No reliable EST, 19861911 Alignment	TIGR	ASIP	Not found
At3g19570	AltDonAccpt	hypothetical protein	No reliable EST	TIGR	ASIP	Not found

At3g47450	AltDonAccpt	putative protein	No reliable EST	TIGR	ASIP	Not found
At3g52780	AltDonAccpt	purple acid phosphatase-like protein	No reliable EST	TIGR	ASIP	Not found
At4g13345	AltDonAccpt	Expressed protein	No reliable EST	TIGR	ASIP	Not found
At4g24550	AltDonAccpt	clathrin coat assembly like protein	No reliable EST	TIGR	ASIP	Not found
At4g31780	AltDonAccpt	1,2-diacylglycerol 3-beta-galactosyltransferase (UDP-galactose:diacylglycerol galactosyltransferase) (MGDG synthase) (MGD1), putative	No reliable EST	TIGR	ASIP	Not found
At5g14660	AltDonAccpt	expressed protein	No reliable EST	TIGR	ASIP	Not found
At5g14850	AltDonAccpt	mannosyltransferase, putative	No reliable EST	TIGR	ASIP	Not found
At5g22050	AltDonAccpt, IntronRetain	protein kinase-related	No reliable EST	TIGR	ASIP	Not found
At5g25980	AltDonAccpt	glycosyl hydrolase family 1	No reliable EST	TIGR	ASIP	Not found
At5g26000	AltDonAccpt	glycosyl hydrolase family 1, myrosinase precursor	No reliable EST	TIGR	ASIP	Not found
At5g39785	AltDonAccpt	Expressed protein	No reliable EST	TIGR	ASIP	Not found
At5g48000	AltDonAccpt	cytochrome p450 family	Terminal exon, Ignored	TIGR	ASIP	Not found
At5g51630	AltDonAccpt	disease resistance protein (TIR-NBS-LRR class), putative	No reliable EST	TIGR	ASIP	Not found
At5g63620	AltDonAccpt	zinc-binding dehydrogenase-related	No reliable EST	TIGR	ASIP	Not found
At5g65010	AltDonAccpt	asparagine synthetase (gb|AAC72837.1)	No reliable EST	TIGR	ASIP	Not found
At4g14310	ExonSkip	hypothetical protein	No EST	TIGR	ASIP	Not found
At4g35740	ExonSkip	putative protein	No reliable EST	TIGR	ASIP	Not found
At1g09280	IntronRetain	expressed protein	No reliable EST	TIGR	ASIP	Not found
At2g14720	IntronRetain	spot 3 protein and vacuolar sorting receptor homolog AtELP2b	Intron too short?	TIGR	ASIP	Not found
At3g63500	IntronRetain	putative protein	No reliable EST	TIGR	ASIP	Not found
At4g13850	IntronRetain	glycine-rich RNA-binding protein (AtGRP2)	No reliable EST, filtered	TIGR	ASIP	Not found
At4g25390	IntronRetain	protein kinase family	No EST	TIGR	ASIP	Not found
At5g22220	IntronRetain	E2FB transcription factor	Intron too short?	TIGR	ASIP	Not found
At5g58320	IntronRetain	contains similarity to unknown protein (gb|AAB63087.1)	No reliable EST	TIGR	ASIP	Not found
10 TIGR-ExonS are possible false positive. Other AS events were identified for these genes in ASIP.						
At1g52730	ExonSkip	WD-40 repeat protein family	No EST	TIGR	ASIP	AltA; IntronR
At2g38880	ExonSkip	putative CCAAT-binding transcription factor subunit	No reliable EST	TIGR	ASIP	AltA; AltD; IntronR
At3g59970	ExonSkip	methylenetetrahydrofolate reductase MTHFR1	No reliable EST	TIGR	ASIP	IntronR
At4g01610	ExonSkip	cathepsin B-like cysteine protease, putative	No reliable EST	TIGR	ASIP	IntronR
At4g09970	ExonSkip	putative protein	No EST	TIGR	ASIP	AltA
At4g32470	ExonSkip	ubiquinol-cytochrome c reductase - like protein	No reliable EST	TIGR	ASIP	AltA
At4g36210	ExonSkip	putative protein	No EST	TIGR	ASIP	IntronR
At4g38240	ExonSkip	alpha-1,3-mannosyl-glycoprotein beta-1,2-N-acetylglucosaminyltransferase, putative	No EST	TIGR	ASIP	IntronR
At5g19330	ExonSkip	putative protein	No EST	TIGR	ASIP	AltA
At5g34940	ExonSkip	putative protein	No EST	TIGR	ASIP	IntronR
3 genes annotation changed, AS found in ASIP						
At1g55680	IntronRetain	WD-40 repeat protein family	[redacted]	TIGR	ASIP	IntronR
At3g04605	IntronRetain	Mutator-like transposase	[redacted]	TIGR	ASIP	IntronR
At3g14800	IntronRetain	expressed protein	[redacted]	TIGR	ASIP	IntronR

Supporting Table 5. Outcome of alternative splicing events.

	AS type	Cases checked	AS position[$] (5-UTR/ORF/3-UTR)	AS outcome (ORF)[#] (NoEffect / ReadThrough /AltORF)	NMD candidates
At	AltD	695	150 / 502 / 43	193 (27.8%) / 109 (15.7%) / 393 (56.5%)	297 (42.7%)
	AltA	1529	219 / 1242 / 68	287 (18.8%) / 434 (28.4%) / 808 (52.8%)	560 (36.6%)
	AltP	144	17 / 118 / 9	26 (18.1%) / 49 (34%) / 69 (47.9%)	32 (22.2%)
	ExonS*	227	47 / 171 / 9	56 (24.7%) / 43 (18.9%) / 128 (56.4%)	97 (42.7%)
	IntronR	2327	313 /1828 / 186	499 (21.4%) / 172 (7.4%) /1656 (71.2%)	1110 (47.7%)
	Total	4922	746 / 3861 / 315	1061 (**21.6%**) / 807 (**16.4%**) / 3054 (**62%**)	2096 (**42.6%**)
Rice	AltD	898	205 / 570 / 123	328 (36.5%) / 110 (12.2%) / 460 (51.2%)	300 (33.4%)
	AltA	1549	235 / 1154 / 160	395 (25.5%) / 327 (21.1%) / 827 (53.4%)	532 (34.3%)
	AltP	345	53 / 257 / 35	88 (25.5%) / 72 (20.9%) / 185 (53.6%)	65 (18.8%)
	ExonS*	537	98 / 377 / 62	160 (29.8%) / 80 (14.9%) / 297 (55.3%)	186 (34.6%)
	IntronR	2760	361 / 1942 /457	818 (29.6%) / 281(10.2%) / 1661 (60.2%)	1103 (40.0%)
	Total	6089	952 / 4300 / 837	1789 (**29.4%**) / 870 (**14.3%**) / 3430 (**56.3%**)	2186 (**35.9%**)

*Only perfect exon skipping is considered here. Perfect exon skipping refers to the case where only one exon is skipped while the surrounding exons are identical between different isoforms.

[$]AS position is determined by checking the position of alternative spliced introns relative to the open reading frame. If the intron is located in the UTR region in both isoforms, the AS position is assigned as either 5-UTR or 3-UTR. In other cases, the AS position is assigned as ORF.

All AS events in UTR are thought to have no effect on ORF (NoEffect). If AS events do not change the start and stop codons, they are regarded as read-through (ReadThrough). All other cases are classified as AltORF.

Supporting Table 6. Arabidopsis genes bearing conserved alternative splicing (AS) events. ORFnote: outcome of AS on ORF; NT, not read-through; NT-truncated, not read-through and will produce truncated protein; NT-changeStart, not read-through, possibly change the start codon; NMDnote: Coupling of AS with NMD; NMD, AS will produce non-sense mediated decay candidates; No-NMD, AS will not produce NMD candidates;

Arabidopsis					Rice				
geneID	intronID	AStype	ORFnote	NMDnote	geneID	intronID	AStype	ORFnote	NMDnote
At1g15110	5874	IntronR	NT-truncated	NMD	Os05g48060	79455	IntronR	NT-truncated	NMD
At1g20620	7995	IntronR	NT	No-NMD	Os06g51150	89402	IntronR	NT-truncated	No-NMD
At1g32200	11674	IntronR	NT-truncated	NMD	Os10g42720	21343	IntronR	NT-truncated	Ignored
At1g44446	12884	IntronR	NT-truncated	NMD	Os10g41780	21081	IntronR	NT-truncated	NMD
At1g53800	15452	AltD	NTchangeStart	NMD	Os10g12360	16801	AltD	NTchangeStart	NMD
At1g67300	19380	AltA	NT-truncated	NMD	Os02g17500	37791	AltA	NT-truncated	NMD
At1g70760	20631	IntronR	NT-truncated	No-NMD	Os05g28090	75379	IntronR	NT-truncated	No-NMD
At1g79820	24068	AltA	NT-elongated	NMD	Os02g17500	37791	AltA	NT-truncated	NMD
At2g17130	27107	AltA	RT-AddOn	No-NMD	Os04g40310	65962	AltA	RT-Deleted	No-NMD
At2g18960	27756	IntronR	NT-truncated	NMD	Os04g56160	70236	IntronR	NT-truncated	NMD
At2g21660	28950	IntronR AltB	RT-AddOn	No-NMD	Os03g46770	56832	IntronR AltB	NT-truncated	No-NMD
At2g25605	30096	IntronR	NT-truncated	NMD	Os05g49910	80019	IntronR	NT-truncated	NMD
At2g28550	31403	IntronR	NT-truncated	NMD	Os04g55560	70047	IntronR	NT	NMD
At2g30860	32094	IntronR	RT-Deleted	No-NMD	Os01g55830	10661	IntronR	RT-AddOn	No-NMD
At2g40110	35876	IntronR	NT-truncated	NMD	Os03g49150	57312	IntronR	NT-truncated	NMD
At2g44680	37805	AltD	RT-AddOn	No-NMD	Os10g41520	20992	AltD	RT-Deleted	No-NMD
At3g01910	39661	AltD	NT-changeStart	NMD	Os08g41830	104573	AltD	NTchangeStart	NMD
At3g02070	39732	IntronR	NT-truncated	NMD	Os04g32970	64431	IntronR	NT-truncated	NMD
At3g24520	48953	IntronR	NT-truncated	No-NMD	Os03g53340	58596	IntronR	NT-truncated	No-NMD
At3g27770	50049	IntronR	NT-changeStart	NMD	Os02g03790	34418	IntronR	NTchangeStart	NMD
At3g46130	51408	IntronR	NT-changeStart	NMD	Os11g47460	27063	IntronR	NTchangeStart	NMD
At3g58010	55552	IntronR	NT-changeStart	NMD	Os04g57020	70548	IntronR	NTchangeStart	Ignored
At3g62310	57166	AltA	NT-truncated	NMD	Os03g19960	52741	AltA	NTchangeStart	NMD
At4g02890	58778	IntronR	RT-AddOn	No-NMD	Os02g06640	35340	IntronR	RT-AddOn	No-NMD
At4g05320	59456	IntronR AltB	RT-AddOn	No-NMD	Os02g06640	35340	IntronR AltB	RT-AddOn	No-NMD
At4g16520	62428	IntronR	NT-truncated	NMD	Os08g09240	99895	IntronR	NT-truncated	NMD
At4g17150	62737	AltA	NTchangeStart	NMD	Os02g10440	36450	AltA	NTchangeStart	NMD
At4g19600	63686	IntronR	NT-truncated	NMD	Os11g05850	22607	IntronR	NT	No-NMD
At4g23400	64944	AltB	NT-truncated	No-NMD	Os02g44630	42437	AltB	NT	NMD
At4g23470	64990	IntronR	NT-truncated	No-NMD	Os10g39100	20335	IntronR	NT	No-NMD
At4g27960	67023	IntronR	NT-truncated	No-NMD	Os01g46930	8270	IntronR	NT	No-NMD
At4g31860	68548	IntronR	NT-truncated	NMD	Os06g44210	87456	IntronR	NT-truncated	NMD
At4g33050	69187	IntronR	RT-AddOn	No-NMD	Os10g27170	18010	IntronR	NT	NMD
At4g35950	70450	AltA	NT-truncated	No-NMD	Os02g58730	47040	AltA	NT	No-NMD
At4g36980	70897	AltD	NT-truncated	NMD	Os03g27840	54497	AltD	NTchangeStart	NMD
At4g39260	71814	IntronR AltB	RT-AddOn	No-NMD	Os03g46770	56830	IntronR AltB	RT-AddOn	No-NMD
At5g03240	72844	IntronR AltB	RT-AddOn	No-NMD	Os02g06640	35340	IntronR AltB	RT-AddOn	No-NMD
At5g19770	79478	ExonS	RT-Deleted	No-NMD	Os07g38730	95362	ExonS	RT-Deleted	No-NMD
At5g26610	81921	IntronR	NT-truncated	NMD	Os04g02500	62008	IntronR	NT-truncated	NMD Ignored
At5g41700	84558	IntronR	NT-truncated	No-NMD	Os06g30970	85728	IntronR	NT	No-NMD
At5g59780	91235	IntronR AltD	NT-changeStart	Ignored	Os11g47460	27063	IntronR AltD	NTchangeStart	NMD

Chapter 4: Molecular characterization and phylogeny of U2AF1 homologs in plants

A paper to be submitted to *Plant Physiology*

Bing-Bing Wang[1] and Volker Brendel[1,2]

[1]Department of Genetics, Development and Cell Biology and [2]Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

Abstract

U2AF is an essential splicing factor with critical roles in recognition of the 3'-splice site. In animals, the U2AF small subunit (U2AFS) can bind to the 3'-AG intron border and promote U2 snRNP binding to the branchpoint sequences of introns through interaction with the U2AF large subunit. Two copies of U2AFS-encoding genes (U2AF1) were identified in *Arabidopsis* (AUSa and AUSb). Both are expressed in all tissues checked, with AUSa expressed at a higher level than AUSb in most tissues. Differences in the expression patterns of AUSa and AUSb in roots were revealed by a promoter::GUS assay, with AUSb expressed strongly in whole young roots and root tips and AUSa limited to root vascular regions. Altered expression levels of AUSa or AUSb cause pleiotropic phenotypes (including flowering time, leaf morphology, and flower and silique shape) and splicing pattern changes

for some pre-mRNA. U2AFS homologs were also identified from maize, rice and other plants with large-scale EST projects. A novel C-terminal domain (SERE) is highly conserved in all seed plant protein homologs, suggesting it may have an important function specific to higher plants.

Introduction

Splicing is an essential process in eukaryotic gene expression. The precise excision of introns from pre-mRNA requires a dynamically assembled RNA-protein complex (spliceosome). Many proteins participate in intron and exon definition prior to the assembly of U1 and U2 small nuclear RNP (snRNPs). U2AF is such a splicing factor. Before spliceosome assembly, U2AF binds to the polypyrimidine tract (Py-tract) between the intron branch point and 3' AG dinucleotide intron boundary to recruit U2 snRNP to the branch site sequence. The U2AF protein is composed of a large subunit ($U2AF^{65}$) and a small subunit ($U2AF^{35}$) (Zamore and Green, 1989), with $U2AF^{65}$ binding directly to the Py-tract (Zamore et al., 1992) and $U2AF^{35}$ binding to the 3' AG boundary (Merendino et al., 1999; Wu et al., 1999; Zorio and Blumenthal, 1999). $U2AF^{35}$ can promote the binding of $U2AF^{65}$ to the Py-tract by interacting simultaneously with the $U2AF^{65}$ and SR proteins (Zuo and Maniatis, 1996). It was also shown that the binding of $U2AF^{35}$ can trigger events in spliceosome assembly in addition to stabilizing $U2AF^{65}$ binding (Guth et al., 2001). *In vivo* studies in *Drosophila* revealed that $U2AF^{35}$ is an essential factor, because mutations in $U2AF^{38}$ caused lethality and development defects (Rudner et al., 1996). *In vitro* studies, however, suggested that some pre-mRNAs, including human β-globin pre-mRNA and adenovirus major late (AdML) pre-mRNA, do not

require $U2AF^{35}$ in splicing (Guth et al., 1999). These seemingly conflicting results indicate that $U2AF^{35}$ may function in a substrate-specific manner.

The gene encoding $U2AF^{35}$ was designated U2AF1. Single to multiple copies of U2AF1 were identified in fission yeast (Wentz-Hunter and Potashkin, 1996), worm (Zorio and Blumenthal, 1999), fly (Rudner et al., 1996), fish (Tassone et al., 1999), chicken (Pacheco et al., 2004), and mammals (Zhang et al., 1992; Tupler et al., 2001; Shepard et al., 2002). As the molecular weights of these homologous proteins are not always 35-KDa, we proposed to use the term U2AFS to generally refer to the protein product of U2AF1 genes. All U2AFS proteins contain a pseudo-RNA recognition motif (ψRRM), which functions as a novel protein recognition motif and thus was renamed as UHM (U2AF homology motif) (Kielkopf et al., 2001; Kielkopf et al., 2004). The ψRRM domain is flanked by two highly conserved $C_8C_5C_3H$ zinc fingers (Kielkopf et al., 2001). U2AFS also contains a C-terminal RS domain (Zhang et al., 1992), which is believed to mediate the protein interaction between $U2AF^{35}$ and SR proteins. The ψRRM and two zinc fingers are essential for the U2AFS function (Webb and Wise, 2004), while the RS domain was found to be dispensable in vivo (Rudner et al., 1998). In mammals, a recently-duplicated copy of U2AF1 encodes a 26-KDa protein ($U2AF^{26}$), which is nearly identical to $U2AF^{35}$ in the N-terminal region (Shepard et al., 2002). $U2AF^{26}$ lacks the C-terminal RS domain but is still able to functionally substitute for $U2AF^{35}$ in splicing (Shepard et al., 2002). A recent study demonstrated that an alternative splicing isoform of the U2AF1 pre-mRNA is conserved from fish to human (Pacheco et al., 2004). The isoform will produce a protein with seven amino acid differences located in the ψRRM. The protein is still be able to bind to $U2AF^{65}$ and promote U2AF splicing activity *in vitro*

(Pacheco et al., 2004). Multiple copies and isoforms of U2AF1 may contribute to the fine tuned control of pre-mRNA splicing in vertebrates (Pacheco et al., 2004). In yeast (*Saccharomyces cerevisiae*), however, no ortholog of $U2AF^{35}$ exists although a functional ortholog of $U2AF^{65}$ was identified (Mud2p) (Abovich et al., 1994). No SR proteins were found either, possibly because the conserved branch consensus sequence eliminates the requirements for these factors in yeast.

Mammalian introns can be classified into AG-dependent and AG-independent types (Reed, 1989). In the AG-dependent introns, the Py-tract is short (weak) and the adjacent 3'-AG boundary is essential for splicing (Reed, 1989). The function of $U2AF^{35}$ is important in these introns, as it stabilizes the binding of $U2AF^{65}$ with the weak Py-tract (Zuo and Maniatis, 1996). AG-independent introns, however, have a long (strong) Py-tract, and the AG is not required for splicing (Reed, 1989). $U2AF^{65}$ alone is sufficient for recruiting the U2 snRNP to the branchpoint sequence (Wu et al., 1999). Plant introns have neither conserved branchpoint sequences nor a Py-tract. Two $U2AF^{65}$ homologs isolated from *Nicotiana plumbaginifolia* can complement the *in vitro* splicing of adenovirus pre-mRNA in HeLa cell extracts depleted of U2AF factor (Domon et al., 1998). Previous results revealed that Arabidopsis has three copies of genes encoding the U2AF large subunit and one possible pseudogene (Wang and Brendel, 2004). These results suggest that the mechanism of 3'-ss recognition is conserved in plants. It is possible that U2AFS also plays a critical role in plant 3'-ss recognition, compensating for the lack of strong branchpoint and Py-tract motifs.

Currently very little is known about the detailed splicing mechanism in plants. Our recent

survey revealed that most metazoan splicing factors are conserved and more than half of them are duplicated in plants (Wang and Brendel, 2004).We proposed that plants share the general splicing mechanism with metazoans but have distinct regulatory mechanisms (Wang and Brendel, 2004). Compared with nine SR proteins in human, a total of 19 SR proteins were identified from Arabidopsis, including four families of novel SR proteins (Lazar et al., 1995; Lopato et al., 1996b; Lopato et al., 1996a; Golovkin and Reddy, 1998, 1999; Lopato et al., 1999a; Lopato et al., 1999b; Lopato et al., 2002; Kalyna and Barta, 2004; Wang and Brendel, 2004). Some novel SR-proteins were found to be conserved in maize (Gupta et al., 2005), suggesting that these genes are possibly conserved in the plant kingdom. Recent studies using fluorescent protein tags revealed that SR proteins are dynamically distributed in nuclear speckles (Ali et al., 2003; Docquier et al., 2004; Fang et al., 2004; Lorkovic et al., 2004; Tillemans et al., 2005). The overall nuclear localization of Arabidopsis SR proteins is similar to each other and to animal homologs, although differences also exist (Tillemans et al., 2005). Overexpression of atSRp30 and atRSZ33 changes the alternative splicing pattern of some endogenous genes and causes morphological and developmental abnormalities (Lopato et al., 1999b; Kalyna et al., 2003). It is likely that atSRp30 and atRSZ33 may have important roles in splice site selection. As SR proteins function in 3'-ss recognition by binding to exonic splicing enhancers (ESEs) (Tian and Maniatis, 1993), and $U2AF^{35}$ bridges SR protein binding to ESEs with $U2AF^{65}$ binding to the Py-tract to stabilize the interactions (Zuo and Maniatis, 1996), it is of great interest to characterize the $U2AF^{35}$ homologs to understand the mechanism of 3'-ss recognition in plants .

No U2AF1 homolog has been identified experimentally in plants. Database searches revealed

two copies of potential U2AF1 genes in Arabidopsis (Domon et al., 1998; Wang and Brendel, 2004). These two genes are highly conserved with their metazoan counterparts on the sequence level, indicating that their functions may also be similar. In this study, we report the experimental characterization of the two genes. Expression pattern differences and functional divergences were found to distinguish the genes. Computational identification of U2AFS homologs in other plants revealed a highly conserved C-terminal domain specific to the plant clade of U2AFS homologs.

Results

Identification of *Arabidopsis* U2AF Small subunit homologs (AUS)

We recently reported the computational identification of two U2AF1 homologs (At1g27650 and At5g42820) in Arabidopsis in a genome-scale analysis (Wang and Brendel, 2004). At1g27650 maps to the short arm of chromosome 1 and encodes a predicted polypeptide of 296 amino acids. At5g42820 maps to the long arm of chromosome 5 and encodes a predicted polypeptide of 283 amino acids. There is a third gene (At1g10320) showing significant similarity to human U2AFS but is presumably the ortholog of a mammalian U2AF1 related sequence, U2AF1-RS1 (Kitagawa et al., 1995). The gene namenclature conventions of previous genome-scale studies were followed, and AUSa and AUSb were used to represent At1g27650 and At5g42820, respectively (Arabidopsis U2AF small subunit a and b). As shown in Supplementary Figure 1, The AUSa and AUSb proteins contain most of the conserved domains of hsU2AF[35], including the ψRRM, one RS domain, two zinc fingers, and the two regions for interacting with U2AF large subunit (Kielkopf et al., 2001). Both proteins

lack the stretch of glycines existing in hsU2AF^{35}. The overall sequences of the two *Arabidopsis* proteins are 83% identical to each other. Each *Arabidopsis* homolog shows approximately 69% similarity to hsU2AF^{35} on the amino acid level (BLAST2 alignment using default parameters).

At the moment of this study, 16 EST sequences and eight cDNA sequences could be aligned to the AUSa region (see displays at http://www.plantgdb.org/AtGDB/). Spliced alignment of the full-length cDNAs reveals a 402nt intron in the 5'-UTR region of AUSa. Two ESTs (gi2757034, gi5839839) indicate that the 5'-UTR intron may be retained in some tissues. For AUSb, only three ESTs were available, and all matched to the 3'-end of the gene. One of the EST clones (Accession: AI997531) was ordered (Genome Systems Inc., St. Louis, MO) and sequenced from both directions. The full-length sequence identified the 5'-end of the AUSb gene, which includes a 277nt intron in the 5'-UTR region. It also revealed a poly (A) tail at the 3'-end (which was missing in the original EST sequence). The sequence was deposited to GenBank (Accession: AF409140). There is about 80% identity between the protein-coding sequence of AUSa and AUSb. Based on the EST library information and difference in EST numbers between AUSa and AUSb, we concluded that AUSa is expressed in all major tissues and probably at a higher level than the AUSb gene.

Expression patterns and alternative splicing of AUS

To verify the gene annotation and check the expression patterns of the AUSa and AUSb genes, specific primers were designed from the 5'- and 3'-UTRs of both genes. RT-PCR was

conducted using RNAs extracted from *Arabidopsis* 7-days seedling, leaf before flowering (LeafBF), leaf after flowering (LeafAF), meristem after flowering (MeriAF), root after flowering (RootAF), stem, flower and silique tissues. As shown in Figure 1A, both AUSa and AUSb genes express in all these tissues. RT-PCR using total RNAs from root before flowering and meristem before flowering revealed the same expression pattern (data not shown). No clear intron retention product can be identified for either AUSa or AUSb, indicating that the 5'-UTR intron is spliced efficiently.

Intriguingly, AUSa seems to have non-canonical introns in addition to the 5'-UTR intron. Several smaller bands in addition to the main product were observed in the RT-PCR for AUSa. Sequencing results revealed that two additional segments in the 3'-end could be removed from the main transcript. These additional segments are possible introns and named AltIntron1 and AltInron2 (AltIntron stands for alternative intron). AltIntron1 (287nt) and AltIntron2 (345nt) overlap with each other. The position of these introns is shown in Figure 1B. Neither of the alternative introns is canonical. Both have a repeat region flanking the intron/exon junction (AltIntron1: AGGAGCA; AltIntron2: AAAAC), thus their real borders are difficult to determine. The GC-content of AltIntron1 and AltIntron2 is 55% and 53.6%, respectively, a little higher than the overall GC-content of the mRNA (50%). Independent RT-PCRs using different transcriptase and RNAs confirmed the existence of additional products. Splicing out of AltIntron1 and AltIntron2 will remove the coding sequences for the C-terminus of AUSa protein. The truncated Arabidopsis U2AFS proteins retain the conserved N-terminal domains and a shortened SR domain. This domain structure is similar to U2AF26, a duplicated copy of U2AF35 in human and mouse (Shepard et al., 2002).

Whether the truncated AUSa proteins really exist is unknown. Non-canonical introns were also reported in other plant splicing factors (Gupta et al., 2005).

AUSa has higher expression levels than AUSb in most tissues

To compare the expression level of AUSa and AUSb in different tissues, real time RT-PCR was employed using primers designed from the 5'-UTR region for both genes. As the two genes are very similar to each other at the nucleotide level, it is not feasible to design gene specific probes for Northern analysis. To eliminate possible DNA contamination, the reverse primers were designed from the exon junctions of the 5'-UTR intron for both genes. A validation experiment using different dilution of cDNAs confirmed that the primer for AUSa and AUSb gene have similar amplification efficiency (data not shown). The expression of AUSa in LeafBF was arbitrarily selected as calibrator and the other samples were compared with it to get relative expression levels.

As shown in Figure 2, real time RT-PCR revealed that AUSa transcript is more abundant than AUSb in all tissues tested. In flower, stem, silique, RootAF and MeriAF, the AUSa transcript level is significantly higher than the AUSb level (t-test, $P<0.05$). AUSa expresses in a level over two-fold higher than AUSb in these tissues. In seedlings and leaf tissues (LeafAF and LeafBF), the difference between AUSa and AUSb is not statistically significant. The AUSa level is less than twice that of AUSb in leaves. It seems that both AUSa and AUSb express at a relatively stable level in different tissues. Compared with the expression level in LeafBF, no significant difference was observed for AUSb in other tissues. For AUSa,

only flower and MeriAF have significantly higher expression. From these results, we concluded that AUSa expresses in a level similar to or a little higher than AUSb in most tissues before flowering. After flowering, both genes have increased level in leaves. The expression of AUSa increases significantly in meristem and flower, while expression of AUSb seems to be decreased in roots.

Promoter::GUS assays reveal similarities and differences between the expression patterns of AUSa and AUSb

As RT-PCR results only revealed the expression patterns of AUSa and AUSb in major tissues, a promoter::GUS assay was employed to identify the expression patterns in more detail. For both AUSa and AUSb, two promoter constructs (the long and short one, AUSa: 876bp and 1358bp; AUSb: 555bp and 982bp, see Materials and Methods) revealed similar GUS staining patterns, indicating that the short promoter region is a functional unit. The strong promoter control (CaMV35S) showed strong GUS expression throughout the plants. No GUS activities were detected in the negative control transgenic plants (data not shown).

As shown in Figure 3, the GUS staining patterns are consistent with RT-PCR results and demonstrate that both AUSa and AUSb genes express in most tissues. The spatial and temporal expression of the two genes is similar in most tissues. Strong GUS activities are detected in 2-5 days seedlings (Figure 3, A, F, G). Shoot meristem, leaf primordial and young leaves including trichromes show intense GUS activity (Figure 3, B, C, and I-K). In large leaves, GUS expression is decreased and localized around vascular tissues of leaves (Figure

3I). Adult leaf blades show weak GUS activity, while the vasculature and petioles have stronger GUS expression (Figure 3, B, I). The most intense GUS activity is found in flowers. Flower buds, sepals, stamens, anthers, pollens, stigmas and the basal region of flowers all show strong GUS activity (Figure 3, D, L-N). Petals show weak but detectable GUS expression. In siliques, the placenta and funiculus have strong GUS expression (Figure 3E).

Differences between the expression of AUSa and AUSb were also discovered in flowers and young roots. As shown in Figure 3M-Q, strong GUS activities were detected on the tops of pistils in control plants (CaMV35S). The AUSa transformants also have detectable GUS expression. AUSb transformants, however, have no clear GUS activities. A distinct expression pattern was discovered in young roots. For AUSa, the expression in young roots is limited to vascular regions (Figure 3P). Root tips and hairs do not show clear GUS expression (Figure 3S). The AUSb and CaMV35S promoter drive strong GUS activities on the whole root (Figure 3Q, R), with the strongest expression in root tips (Figure 3T, U).

Both AUSa and AUSb proteins localize to the cell nucleus

As pre-mRNA splicing takes place in the nucleus, the AUSa and AUSb gene products should have nuclear localization if they are indeed splicing factors. AUSa and AUSb ORF sequences were fused in-frame downstream of the GFP coding sequence driven by CaMV35S promoter in gateway vector pMDC43 (Curtis and Grossniklaus, 2003). As shown in Figure 4, both AUSa and AUSb proteins are clearly enriched in nuclei. No differences were detected between the two genes. Strong green florescence was detected in the nuclei of root cells

(Figure 4AB), root hair cells (Figure 4B), leaf cells, guard cells (Figure 4DE) and trichomes (Figure 4F). GFP protein alone targets to both nucleus and cytoplasm, as indicated by the florescence in whole roots and relatively weak florescence in nucleus (Figure 4C). Detailed study using confocal microscopy revealed that the distribution of AUSa and AUSb protein in nucleus is not even (Figure 4AB, insets). They both are likely organized in nuclear speckles, a pattern similar to known SR proteins (Tillemans et al., 2005).

Plants with altered expression levels of AUSa or AUSb genes show pleiotropic morphological changes

One T-DNA insertion line (SALK_050678) was identified for AUSa (ordered from ABRC, Arabidopsis Biological Resource Center)(Alonso et al., 2003). The T-DNA is inserted into the 5'-UTR intron. The AUSa gene still expresses in the homozygous mutants. The intron (402nt) with T-DNA insertion (~4.4 kb) is spliced out. Real time RT-PCR revealed that the full-length transcript level of AUSa is down regulated by 2.5 fold (Figure 5). To knock down AUSb gene, antisense and RNAi vectors were constructed based on pCAMBIA1301. Both constructs were transformed into Arabidopsis. The AUSb-RNAi plant has the AUSb gene knocked-down to seven-fold lower expression (Figure 5). The AUSb level in the antisense plant is not clearly down-regulated. Surprisingly, AUSa instead is up-regulated, with transcript abundance in the mutant about twice the wild type level. The ratio of AUSb to AUSa expression decreased, as the AUSb/AUSa ratio is about 0.5 in the mutant and 1.0 in wild type plants at the same stage. The AUSa T-DNA insertion plant and AUSb transgenic plants show similar as well as distinct morphological phenotypes. As shown in Figure 6, they

are all late flowering. Under continuous light condition, wild type Arabidopsis (Col-CS60000) flowers at 20-25 days with 11-12 leaves. The AUSa T-DNA plants flower at 25-30 days with 12-15 leaves. The AUSb RNAi and antisense plants flower at 26-32 days with 13-15 leaves. All three mutants have shorter flowers with an enlarged bottom part compared with wild type (Figure 6). The leaf morphology is also changed in the three mutants. Compared with wild type, the rosette leaves in AUSa-TDNA plants are larger with flat surface. Leaf color in AUSa-TDNA is lighter and yellowish. Rosette leaves in AUSb-RNAi and antisense plants, however, are smaller and dark green. The leaf surface is less flat and leaf edge is more serrated than wild type. Cauline leaves in these mutants have similar changes as the rosette leaves. The shape of siliques is also altered in AUSa-TDNA plants. As shown in Figure 6, normal siliques are cylindrical, tapering at the ends. In contrast, the siliques in AUSa-TDNA plants are flattened, and widened at the distal end. Silique number is increased at the terminus of the stem in AUSa mutants. In some AUSa mutants as well as some AUSb transgenic plants, we found that the primary stem stopped growing at a certain stage (Figure 6D, indicated by a red arrow), with two to three branches below the terminus growing normally.

Some constitutive splicing but not alternative splicing patterns are affected in mutants

As AUSa and AUSb presumably function in splicing, we checked the splicing pattern of 18 genes in the mutants by RT-PCR. 12 of the 18 genes were predicted to be alternatively spliced (AS), and nine of the 12 have been validated as AS by RT-PCR in a separate study (Wang and Brendel, unpublished). The remaining 6 genes include the AUSa and AUSb

genes, three genes (FLC, FCA and FPA) involved in the flowering pathway (of interest because the mutants are late flowering), and AtDBR1 (At4g31770) as a constitutively spliced multiple-intron gene. No differences were observed for the splicing patterns of the 12 AS genes among AUSa and AUSb mutants and wild type, indicating that both AUSa and AUSb genes may not function in regulating splice site selection. One example of the 12 AS genes (KH-NOVA, At5g04430) is shown in Figure 7, upper right panel. The FCA gene can also be alternatively spliced, with the γ-isoform encoding full-length protein to promote flowering, and the β-isoform using an alternative polyadenine site in the 3rd intron (Macknight et al., 1997). The γ-isoform expresses in low levels in our mutants and wild type plants. Clear differences exist for β-isoform expression levels, which are the highest in wild type plants and lowest in AUSb RNAi and antisense plants. The expression level for flowering repressor FLC gene is also changed in mutants. Consistent with the late flowering phenotype, the AUSb-antisense plants have high level FLC expression (Figure 7, bottom panel). No difference was found for the FPA gene. The results for the AUSa and AUSb genes are consistent with the real time RT-PCR results of Figure 2. Interestingly, for AtDBR1, we found extra bands in AUSa and AUSb mutants. As shown in the left upper panel in Figure 7, wild type plants produce a single band (~1.2kb), indicating no alternative splicing for AtDBR1. In AUSa-TDNA plants, however, an additional band of smaller size (~1kb) is produced. In AUSb-RNAi and antisense plants, extra bands of larger size (1.3-1.4kb) are produced. These extra bands indicate that novel splicing/alternative splicing will be produced when the AUSa or AUSb expression level is changed.

Maize has at least two U2AF1 homologs

Two maize U2AF1 homologs were found by matching the AUSa sequence against maize ESTs using the PlantGDB BLAST server (http://www.plantgdb.org/cgi-bin/PlantGDBblast/). 26 maize ESTs were found to give significant hits. Two contigs were constructed from these ESTs. One is a full-length contig and named ZUSa. The other one is a partial sequence. One EST (AI491620) from the latter contig was ordered and sequenced from both ends. The sequence turns out to be full-length, and this gene is named ZUSb. The predicted ZUSa and ZUSb protein sequences share 70% identity. Like their counterparts in *Arabidopsis*, they both have all domains conserved in human U2AF[35]. Specific primers were designed from the 5' and 3'-UTR. Through genome PCR, RT-PCR and EST/cDNA alignments, the gene structures of ZUSa and ZUSb were identified. As shown in Figure 8, both maize genes have introns in the 5'-UTR region. ZUSa has two introns, one is 437nt, and the other is 706nt, with an exon of 88nt between them. RT-PCR and sequencing results revealed that the exon can be skipped in some transcripts. For ZUSb, the single 5'-UTR intron also undergoes alternative splicing. A 165nt intron is revealed by spliced alignment of the full-length EST (AI491620) and two other EST sequences (gi33466507 and gi33468026) against the ZUSb genome sequence. Three other EST sequences (gi18655115, gi18655114 and gi18662135) suggest an alternative acceptor site located 290nt downstream, which produces a 455nt intron. The additional 290nt segment has GT-AG borders, indicating that it is a possible intron. The 455nt intron can be spliced either in one step by using the downstream acceptor site or possibly in two steps by first splicing out the 165nt intron and then the 290nt additional intron. The two-step splicing works just like there is an exon of 0nt between the two introns.

Two alleles were identified for ZUSb by sequencing genomic clones from maize inbred lines B73 and W64. ZUSb-B73 is identical to the gene on maize genomic survey sequence (GSS) contig (ZmGSStuc11-12-04.13932.1) assembled by PlantGDB. ZUSb-W64 gene has a 9nt deletion and one base change (C→A) in the 3'end of ORF, causing a three amino acids insertion and one amino acid change. Variations were also observed in the 3'-UTR and in the 5'-UTR intron between ZUSb-B73 and ZUSb-W64. For ZUSa, at least three recent duplications exist in the genome. We sequenced three genomic clones from inbred line B73. Two of them miss an 8nt segment (CTCTCCGT) and a downstream 16nt segment (TGTTGCTTAGCTCCGG), which are present in the 3'-UTR of the third B73 clone. The 8nt and 16nt segments are close to each other, with 19nt between them. Three genomic clones from inbred line W64 have the same sequence. We named the first two clone sequences ZUSa-1 and the third clone sequence ZUSa-2. ZUSa-1 also has several nucleotide changes in the coding region compared with ZUSa-2, but their deduced protein sequences are identical to each other. Current genomic survey sequences (GSS) from B73 also demonstrate that both ZUSa-1 and ZUSa-2 are present in the B73 genome, and our cDNA clones from RT-PCR demonstrate that both ZUSa-1 and ZUSa-2 are expressed. Interestingly, some cDNA clones have an extra 24nt segment (CACCGTGACCGTGATGACTACCAC) in the coding region, located in the 3'-end of the coding region. A segment consisting of the first 21nt of the extra segment exists in all cDNA and genome clones. The 24nt segment is absent in all our genomic clones. Three GSS sequences (gi34246858, gi32023337, gi32014742), however, contain the 24nt segment, suggesting another active copy of ZUSa in B73 (ZUSa-3). There may be additional ZUSa copies/alleles, as suggested by the presence of an extra 8nt segment

(ATTCAGGA) in three cDNA clones and two ESTs (gi13149630 and gi6012605). None of the current B73 GSS sequences contain the segment.

Rice U2AF1 homologs

Two U2AF1 homologs were already identified in rice (Domon et al., 1998). As rice genome sequences are available now (Feng et al., 2002; Sasaki et al., 2002; Yu et al., 2002), both the indica and japonica rice genome sequences were searched, and the chromosome location of the two rice homologs were identified in both subspecies. The osU2AF35a gene is located on chromosome 9 (BAC clone: OSJNBb0052C07; Accession: AC108762), and the osU2AF35b gene is located on chromosome 5 (LOC_Os05g48960). TIGR rice pseudomolecule release 3.0 did not include BAC AC108762, therefore no TIGR locus name was assigned to the osU2AF35a gene. Slight differences exist between the indica and japonica genomic sequences for both osU2AF35a and osU2AF35b, while the protein sequences are identical. In addition to the active copies, one and two pseudogenes were identified in japonica and indica rice, respectively. In japonica rice, the pseudogene (osU2AF35p-J) is located in the region of the TIGR annotated locus LOC_Os01g47750 on chromosome 1. In indica rice, two copies of the osU2AF35p (osU2AF35p-I1, osU2AF35p-I2) are dispersed on chromosome 1, with 30kb between them. These pseudogenes are nearly identical to each other and have over 85% similarity to osU2AF35b on the nucleotide level. Compared with the active copies, the pseudogenes have multiple nucleotide deletions interrupting the open reading frame in the central region.

EST/cDNA alignments revealed that the gene structures and splicing patterns of osU2AF35a and osU2AF35b are very similar to those of ZUSa and ZUSb. As shown in Figure 8, osU2AF35a has a 5'-UTR intron (500nt), which could be spliced out either in two steps (176nt intron and 324nt intron) or in one step directly from pre-mRNA. About 15 ESTs/cDNAs (such as gi32984690) represent the product of first step splicing (or utilizing the upstream donor site, transcript 1). Three EST/cDNAs (such as gi33682532) represent the product either spliced from the first step product or from pre-mRNA by using the downstream donor site (transcript 2). For osU2AF35b, two introns are present in the 5'-UTR, separated by a 56nt exon. The first intron is 108nt, and the second intron is 1,231nt. The 1,231nt intron could also be spliced either in two steps (564nt and 667nt) or in one step. One EST (gi29641592) and seven ESTs/cDNAs (such as gi3298022) represent transcripts 1 and 2, respectively. EST gi29685226 represents an additional transcript (transcript 3), which can be generated from transcript 1 by removing 119nt from the 3'-end of the 667nt intron. Compared with transcript 2, transcript 3 includes part (547nt) of the 667nt intron. The splicing pattern of osU2AF35b is similar to the pattern in ZUSa, where an 88nt exon can be included. The sequence of the 88nt exon is conserved in a similar position in the osU2AF35b intron. Based on sequence similarity, gene structure and splicing pattern, we conclude that osU2AF35a is the ortholog of ZUSb, and osU2AF35b is the ortholog of ZUSa. There are conserved sequence segments between the orthologs intron pairs, including the region flanking the alternatively spliced intron-exon junction and the central region of the alternative exon. These conserved parts may contribute to the alternative inclusion/exclusion of part of the 5'-UTR intron.

Domain features of plant U2AF1 homologs

By searching the EST sequences in plant species except Arabidopsis, rice and maize, 15 more full-length homologs were identified in nine other species, including three homologs in wheat, two homologs in barley, soybean, tree cotton and potato, and one homolog in Medicago, tomato, pine and unicellular green algae. In addition, 11 partial homologs were also identified in the above species and other species including sorghum, lotus, upland cotton, and rye. The partial proteins all show high similarity to the U2AF1 homologs in their closely related species. Sequences of these plant homologs are available in Supplementary File 1. All the plant full-length U2AF1 proteins were aligned with homologs from human (hmU2AF^{35}), Drosophila (dmU2AF^{38}), *C. elegans* (ceU2AF^{35}) and fission yeast (spU2AF^{23}). Alignment of the N-terminal regions is shown in Supplementary Figure 1. All the U2AF1 homologs contain a degenerate RNA binding domain, two CCCH type zinc fingers and one RS domain with variable length. The two zinc fingers and part of the RNP-1 region are also conserved in lmU2AF23 (Accession: AAF27955, the U2AF1 homolog in parasite *Leishmania major*) and ecU2AF38 (Accession: CAD27114, from parasite *Encephalitozoon cuniculi*), indicating that the ancient U2AF1 protein contained these domains and that their functions are critical among eukaryotic organisms (data not shown). Interestingly, the two segments involved in direct interaction of U2AF^{35} and U2AF^{65} in human (Kielkopf et al., 2001) are not well conserved in plants. The critical tryptophan in the interacting region was replaced by either phenylalanine in most plant homologs or tyrosine in osU2AF^{35}a and ZUSb.

A higher-plant specific domain was observed in a multiple sequence alignment of the C-terminal regions as shown in Figure 9. All U2AF1 homologs from monocot plants, dicot plants and loblolly pine contain a conserved domain of 23 amino acids in their C-terminus. The consensus sequence for the domain is RSPVREGSEERRA(K/R)IEQWNRERE, where underlined amino acids are completely conserved in all higher plant homologs. For the (K/R) site, all monocot homologs except osU2AF^{35}a have K, while most dicot homologs have R. We named this C-terminal domain SERE (short for SEERRAIQWRE). Four U2AF1 homologs in wheat, barley, potato and tomato (taU2AFSa, hvU2AFSp, stU2AFSb and leU2AFSb, respectively) have two SERE domains in their C-terminus. The U2AF1 homolog in unicellular green algae (crU2AFSa) does not contain the SERE domain, indicating the function of the domain may be involved in higher plant specific splicing mechanisms.

Phylogeny of the U2AF1 homologs

A phylogenetic tree was constructed based on a multiple sequence alignment of the N-terminal regions. Six additional animal homologs (hsU2AF^{26}, msU2AF^{26}, msU2AF^{35}, ggU2AF^{35}, tkU2AF^{35} and msU2AFS^{35}-like) and lmU2AF^{23} and ecU2AF^{38} were also included in the analysis. As shown in Figure 10, lmU2AF^{23} and ecU2AF^{38} are outgroups in the phylogenetic tree. All the plant homologs are clustered into one clade. All the animal homologs except msU2AFS^{35}-like gene are clustered into another clade, where all vertebrate homologs are clustered together. The two U2AF^{26} homologs from mouse and human are separated from the vertebrate U2AF^{35} group, indicating that the U2AF^{26} is duplicated from

U2AF^{35} at least before the divergence of human and mouse, possibly before the vertebrate divergence. Later in human, three more duplications happened. But the additional copies turn out to be pseudogenes (our unpublished results and (Tupler et al., 2001)). The msU2AFS^{35}-like homolog is deduced from a RIKEN cDNA sequence (BC003883, gi13278054). Surprisingly it is clustered in the Dicot II clade and showed high similarity to plant homologs. Mouse genome was searched and no genomic region was found for the cDNA sequence. No plant-like U2AF1 homolog could be found in human, Drosophila, *C. elegans* and fission yeast genomes. It is likely that the mouse cDNA sequence results from contamination. The fission yeast spU2AF^{23} is also clustered in the animal group.

In the plant group, the seed plant homologs are clustered into four clades, including two monocot and two dicot clades. The green algae homolog is classified as outgroup. ZUSa (maize), osU2AF35b (rice), hvU2AFSa (barley), taU2AFSb and taU2AFSc (wheat) are clustered into Monocot I clade, and ZUSb, osU2AF35a, hvU2AFSp and taU2AFSa are clustered into Monocot II clade. For dicot U2AF1 homologs, AUSa, AUSb, gmU2AFSb (Soybean) and gaU2AFSb (tree cotton) are clustered into Dicot I clade, and the remainder are clustered into Dicot II clade. Interestingly, the Dicot I clade is clustered in a big group with Monocot clades and ptU2AFSa (from lobolly pine), indicating the ancient form of U2AFS in the ancestor of seed plants. There might be another ancient U2AFS gene, as suggested by the separation of the Dicot II clade from the Dicot I–Monocot–ptU2AFSa group. The second form was lost in monocot and some dicot (Arabidopsis) lineages. After the divergence of monocot and dicot plants, individual duplications of U2AFS genes happened in the ancestor of monocot plants, nightshade family (potato and tomato), Triticeae (wheat and barley) and

Arabidopsis. Recently, the U2AF1 was duplicated again in monocot plants, as suggested by the pseudogenes in rice and three copies of ZUSa in maize.

Discussion

Expression and function of U2AF1 homologs in Arabidopsis

Two homologs of U2AF1 were characterized in Arabidopsis in this study. Both AUSa and AUSb express in all tissues checked, with the expression level of AUSa always higher than that of AUSb. In some tissue (for example, root tip), however, AUSb expresses strongly while AUSa is barely expressed, as suggested by our promoter::GUS assay. The divergence of expression pattern for this duplicated gene pair suggests divergent functions of the two genes.

The sequences and domain structure of the AUSa and AUSb gene products are very similar to human U2AF^{35}, indicating that the two proteins have exchangeable functions similar to their mammalian homologs. What could be their functional divergence? Because U2AF^{35} functions in a substrate-specific manner, it is likely that AUSa and AUSb may have different pre-mRNA substrates. The pre-mRNAs produced only in root tips are good candidates for AUSb substrates. For most pre-mRNAs, AUSa and AUSb may function simultaneously, as they are both expressed in all major tissues. Our data show that altered expression levels of either AUSa or AUSb cause similar phenotypes, including late flowering or leaf morphology changes, suggesting that the two genes may have common substrates. In our RT-PCR analysis on AUSa and AUSb mutants, extra products were observed for the AtDBR1 gene.

These products very likely represent aberrant splicing isoforms generated by the altered level of AUSa or AUSb.

Non-canonical alternative splicing of U2AF1 and other splicing factors

RT-PCR on AUSa genes revealed two extra bands in addition to the constitutively spliced product. Non-canonical alternative splicing events were identified by sequencing the extra products, which removes a segment each with repeated borders from the second exon of AUSa gene. The two isoforms encode similar proteins that retain only the N-terminal domains of AUSa. In our AUSa knock-down mutants, we also found a band smaller than expected in RT-PCR for the AtDBR1 gene. Very likely the reduction of AUSa level will lead to the usage of non-canonical sites in some pre-mRNAs. As the expression level of AUSa changes dynamically during growth and development, it is likely that AUSa can autoregulate the level of full functional protein by these alternative splicing events. In vertebrates, U2AF1 gene can also be alternatively spliced by inclusion of an additional exon, producing an isoform with seven amino acid differences in the ψRRM (Pacheco et al., 2004). Two maize SR proteins (ZmSRp31A and ZmSRp31B) also show non-canonical alternatively spliced introns (Gupta et al., 2005), strongly suggesting that many splicing factors can post-transcriptionally regulate their expression by non-canonical alternative splicing.

Evolutionary history of U2AFS domain structure

From the sequence alignment and phylogenetic analysis, it is clear that the U2AF1 gene would have existed in the ancestor of eukaryotic organisms. The ancient U2AFS contained at least a ψRRM and two CCCH zinc-fingers. It may also contain a run of glycines because the

animal homologs and some monocot plant homologs have this motif. But this motif was lost in some homologs. In addition, several amino acid residues N-proximal to the two CCCH zinc fingers are highly conserved in all known U2AFS proteins (as shown in Supplementary Figure 1), indicating their existence and functional importance in the ancestor. In the plant kingdom, a C-terminal domain (SERE) was acquired after the divergence of green algae and probably before the divergence of seed plants. A tandem copy of the SERE domain was maintained in several plant U2AF1 homologs. The SERE domain may have plant-specific functions, such as recognizing plant-specific splicing signals or interacting with plant-specific SR proteins.

How do plants recognize weak introns?

Plant introns have neither conserved branch sequences nor a Py-tract. The 3'-ss recognition in plants will therefore rely more on U2AF than in mammalians. What's the exact role of plant U2AF homologs and how they achieve their functions are challenging questions. It is likely that various splicing enhancer exist in either intron or exon in plants. As multiple copies of SR proteins also exist in plants, and some of them are plant-specific (Kalyna and Barta, 2004; Wang and Brendel, 2004), we propose here that plant use different U2AF and SR protein combinations to recognize introns with weak splicing signals. Similar to mammalian introns, some plant introns may be AG-independent and may not require U2AFS for correct splicing. Multiple copies of U2AFS exist in nearly all higher plant genomes, and they may preferably interact with different U2AF large subunits to form different U2AF heterodimers, as suggested by variations in the interacting regions of U2AFS from the same species as well as different species (Supplementary Figure 1). The RS domain and its surrounding regions of

U2AFS contain many variations (not shown in the alignments), indicating the flexibility of interaction between U2AFS and variable SR proteins. It is likely that some SR proteins may preferably interact with one of the U2AFS protein, which in turn interacts with specific U2AF large subunits. Different SR proteins may bind to different splicing enhancer elements. Therefore, plant may recognize weak 3'-splice sites by combining different U2AF heterodimers with different SR proteins that bind to splicing enhancer elements.

Materials and Methods

Identification of maize, rice and other plant U2AF1 homologs

Maize and rice homologs were identified by matching the AUSa sequence against the maize GSS assembly (http://www.plantgdb.org/PlantGDB-cgi//blast/PlantGDBblast?db=Zeamays+GSScontig) and rice genome sequences (for japonica rice: http://rice.tigr.org/; for indica rice: http://rise.genomics.org.cn/rice/), respectively. EST collections were also searched on the PlantGDB BLAST server (http://www.plantgdb.org/PlantGDB-cgi/blast/PlantGDBblast). Two maize contigs were obtained through EST analysis. One contig encodes a full-length protein and was named ZUSa. For another contig, EST clone AI491620 was ordered from Stanford University and sequenced from both ends. The sequence revealed another full-length protein and was named as ZUSb. Other plant homologs were identified by matching *Arabidopsis* homologs against plant EST sequences by BLAST. All plant ESTs were downloaded from NCBI Plant Genomes Central (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html). Hits with E-value less than 10^{-04} were regarded as significant hits. All hits were then used to match the EST sequences

again. The same criterion was used to retrieve all related ESTs. CAP3 (Huang and Madan, 1999) was then used to cluster these related ESTs and make contigs. The contigs were matched against all known U2AF1 sequences and against Arabidopsis proteins using BLASTx. Contigs were regarded as U2AF1 homologs if (1) they had an E-value of less than e^{-15} when searching against U2AF1 homologs; and (2) their best hit in Arabidopsis is a U2AF1 homolog. Putative proteins were translated from the contigs by the NCBI-ORF (Open Reading Frame) finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

Sequence alignment and phylogenetic analysis

Multiple sequence alignments of the U2AFS proteins were generated with ClustalW using default parameters (Thompson et al., 1994). The alignments were visualized using the BioEdit program (version 5.0.9 http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Phylogenetic analysis of the sequences was conducted using the MEGA software (version 2.1; http://www.megasoftware.net/; (Kumar et al., 2001)). The phylogenetic trees were constructed using the neighbor-joining method with bootstrap test. The distance model used was Kimura 2-parameter. All other parameters used were default.

***Arabidopsis* growth conditions and RNA extraction**

Arabidopsis seeds were sown in soil and grown at 4 °C grown for four days, then the plants were moved to a growth room and grown at 22 °C with continuous light. Total plant RNA was isolated using either TriZol reagent (Invitrogen) or Plant RNeasy Mini kit (Qiagene, Valencia, USA) from 0.1- 0.2g of different tissues. The manufacture's protocol was followed. For *Arabidopsis*, root, leaf, meristem, stem and flower tissues from wild type ecotype

Columbia were used. For maize, root, shoot and endosperm tissues from inbred line B73 were used. Total RNA was dissolved in 30 µl DEPC-treated H_2O and saved at -20 °C.

RT-PCR and Real-Time RT-PCR

Total RNA was treated by RQ1 RNAase free DNAase according to manufacturer's protocol (Promega, Madison, WI). 2 µg treated RNA were then used for first strand synthesis and PCR according to manufacture's protocol (Invitrogen). A mixture of treated RNAs was used as no-RT control. For real time RT-PCR, PRIMER EXPRESS V2.0 software (Applied Biosystems) was used to design oligonucleotide primers. cDNAs were prepared as described above and diluted 600-fold for amplification of 18S ribosome RNA gene and 3-fold for other genes. 1 µl of diluted cDNA was used in a 25- µl reaction with SYBR Green Master Mix (Applied Biosystems). All reactions were performed in triplicate by using a Prism 5700 Sequence Detection System (Applied Biosystems). The experiments were replicated twice using different RNA samples. Primer efficiency was checked for each primer pair by constructing a standard curve using an equal mixture of all cDNAs (Applied Biosystems). The expression level of each gene was calculated based on $2^{-\Delta\Delta ct}$ method described in user bulletin #2 (Applied Biosystems). The relative amount of calculated message was normalized to the level of 18S rRNA gene.

Promoter::GUS assay

Two potential promoter regions together with 5'-UTR region were checked for both AUSa and AUSb. For AUSa, promoter 1 (PGaa1) is the genomic region from 876nt before the ATG start codon, and promoter 2 (PGaa2) is from 1,358nt before the start codon. For AUSb,

promoter 1 (PGab1) is from 555nt before the start codon, and promoter 2 (PGab2) is from 982nt before the start codon. As shown in Supplementary Figure 2, both longer promoters for AUSa and AUSb include part of the first exon of the neighboring gene. These tentative promoters were amplified from *Arabidopsis* genome DNA. Primers are shown in Supplementary Table 1. PCR products were purified and ligated to vector pCMABIA1381z. The vectors were subjected to sequencing from both ends to make sure the insertions were correct. In addition to the two promoters for each gene, the CaMV35S promoter was linked with GUS gene and used as a strong promoter control (PGxx). The empty pCAMBIA1381z (no promoter) was used as no promoter control (negative control, PG_{00}). These constructs are all shown in Supplementary Figure 2. The right vectors were used for *Arabidopsis* transformation by methods described below. 3-5 individual transgenic plants from each transformation were subjected to histochemical GUS assays, following the protocol described in (Weigel and Glazebrook, 2002).

Antisense and RNAi vector construction

Antisense vectors were constructed based on binary vector pCAMBIA1301. The vector diagrams are shown in Supplementary Figure 3. The ORF of AUSb was amplified by PCR using primers described in Supplementary Table 1. The PCR product was isolated, digested and inserted into the downstream of CaM35S promoter in the reverse direction. To construct an RNA interference (RNAi) vector, the AUSb-ORF was amplified using the primer set shown in Supplementary Table 1. The PCR product and antisense vector were digested by MluI and BsrGI (New England Biolabs, Beverly, MA), then ligated by T4-DNA ligase (Promega, Madison WI). The RNAi construct uses the AUSb-ORF to replace part of the

antisense vector sequence in the sense direction. The resulting transcript forms a hairpin structure that triggers silencing of the endogene.

Arabidopsis transformation

Different vectors were transformed into Agrobacterium by electroporation methods. *Arabidopsis* ecotype Columbia was transformed by Agrobacterium using the floral dip method (Weigel and Glazebrook, 2002). Seeds were screened at 0.8% *Arabidopsis* selective medium containing 50μg/ml hygromycin for seven days, then transformed to 1.5% plates for another seven day. Resistant plants were transferred to soil and analyzed.

Acknowledgments

We would like to thank Wei Huang, Tiffanie Kuhn and Zhen Ni Li for help with experiments. We are also grateful to Wei Huang, Drs. Philip Becraft, Robert Fluhr, and Thomas Peterson for critical reading of the manuscript. Microscopy pictures were taken in the laboratories of Drs. Shuizhang Fei and Jo Anne Powell-Coffman and in the ISU Confocal Microscopy Facility. This work was supported in part by NSF grants DBI-0110189 and DBI-0321600 and Research Grant No. IS-3454-03 from BARD, the United States – Israel Binational Agricultural Research and Development Fund.

Literature Cited

Abovich, N., Liao, X.C., and Rosbash, M. (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. Genes Dev **8,** 843-854.

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

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C.C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D.E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W.L., Berry, C.C., and Ecker, J.R. (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science **301,** 653-657.

Curtis, M.D., and Grossniklaus, U. (2003). A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol **133,** 462-469.

Docquier, S., Tillemans, V., Deltour, R., and Motte, P. (2004). Nuclear bodies and compartmentalization of pre-mRNA splicing factors in higher plants. Chromosoma **112,** 255-266.

Domon, C., Lorkovic, Z.J., Valcarcel, J., and Filipowicz, W. (1998). Multiple forms of the U2 small nuclear ribonucleoprotein auxiliary factor U2AF subunits expressed in higher plants. J Biol Chem **273,** 34603-34610.

Fang, Y., Hearn, S., and Spector, D.L. (2004). Tissue-specific expression and dynamic organization of SR splicing factors in Arabidopsis. Mol Biol Cell **15,** 2664-2673.

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., Jia, P., Zhao, Q., Ying, K., Yu, S., Tang, Y., Weng, Q., Zhang, L., Lu, Y., Mu, J., Zhang, L.S., Yu, Z., Fan, D., Liu, X., Lu, T., Li, C., Wu, Y., Sun, T., Lei, H., Li, T., Hu, H., Guan, J., Wu, M., Zhang, R., Zhou, B., Chen, Z., Chen, L., Jin, Z., Wang, R., Yin, H., Cai, Z., Ren, S., Lv, G., Gu, W., Zhu, G., Tu, Y., Jia, J., Chen, J., Kang, H., Chen, X., Shao, C., Sun, Y., Hu, Q., Zhang, X., Zhang, W., Wang,

L., Ding, C., Sheng, H., Gu, J., Chen, S., Ni, L., Zhu, F., Chen, W., Lan, L., Lai, Y., Cheng, Z., Gu, M., Jiang, J., Li, J., Hong, G., Xue, Y., and Han, B. (2002). Sequence and analysis of rice chromosome 4. Nature **420,** 316-320.

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

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

Gupta, S., Wang, B.B., Stryker, G.A., Zanetti, M.E., and Lal, S.K. (2005). Two novel arginine/serine (SR) proteins in maize are differentially spliced and utilize non-canonical splice sites. Biochim Biophys Acta **1728,** 105-114.

Guth, S., Martinez, C., Gaur, R.K., and Valcarcel, J. (1999). Evidence for substrate-specific requirement of the splicing factor U2AF(35) and for its function after polypyrimidine tract recognition by U2AF(65). Mol Cell Biol **19,** 8263-8271.

Guth, S., Tange, T.O., Kellenberger, E., and Valcarcel, J. (2001). Dual function for U2AF(35) in AG-dependent pre-mRNA splicing. Mol Cell Biol **21,** 7673-7681.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res **9,** 868-877.

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.

Kielkopf, C.L., Lucke, S., and Green, M.R. (2004). U2AF homology motifs: protein recognition in the RRM world. Genes Dev **18,** 1513-1526.

Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. Cell **106,** 595-605.

Kitagawa, K., Wang, X., Hatada, I., Yamaoka, T., Nojima, H., Inazawa, J., Abe, T., Mitsuya, K., Oshimura, M., Murata, A., and et al. (1995). Isolation and mapping of human homologues of an imprinted mouse gene U2af1-rs1. Genomics **30,** 257-263.

Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17,** 1244-1245.

Lazar, G., Schaal, T., Maniatis, T., and Goodman, H.M. (1995). Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF. Proc Natl Acad Sci U S A **92,** 7672-7676.

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

Lopato, S., Mayeda, A., Krainer, A.R., and Barta, A. (1996b). Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors. Proc Natl Acad Sci U S A **93,** 3074-3079.

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

Lorkovic, Z.J., Hilscher, J., and Barta, A. (2004). Use of fluorescent protein tags to study nuclear organization of the spliceosomal machinery in transiently transformed living plant cells. Mol Biol Cell **15,** 3233-3243.

Macknight, R., Bancroft, I., Page, T., Lister, C., Schmidt, R., Love, K., Westphal, L., Murphy, G., Sherson, S., Cobbett, C., and Dean, C. (1997). FCA, a gene

controlling flowering time in Arabidopsis, encodes a protein containing RNA-binding domains. Cell **89,** 737-745.

Merendino, L., Guth, S., Bilbao, D., Martinez, C., and Valcarcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. Nature **402,** 838-841.

Pacheco, T.R., Gomes, A.Q., Barbosa-Morais, N.L., Benes, V., Ansorge, W., Wollerton, M., Smith, C.W., Valcarcel, J., and Carmo-Fonseca, M. (2004). Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS. J Biol Chem **279,** 27039-27049.

Reed, R. (1989). The organization of 3' splice-site sequences in mammalian introns. Genes Dev **3,** 2113-2123.

Rudner, D.Z., Breger, K.S., and Rio, D.C. (1998). Molecular genetic analysis of the heterodimeric splicing factor U2AF: the RS domain on either the large or small Drosophila subunit is dispensable in vivo. Genes Dev **12,** 1010-1021.

Rudner, D.Z., Kanaar, R., Breger, K.S., and Rio, D.C. (1996). Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects. Proc Natl Acad Sci U S A **93,** 10333-10337.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., Antonio, B.A., Kanamori, H., Hosokawa, S., Masukawa, M., Arikawa, K., Chiden, Y., Hayashi, M., Okamoto, M., Ando, T., Aoki, H., Arita, K., Hamada, M., Harada, C., Hijishita, S., Honda, M., Ichikawa, Y., Idonuma, A., Iijima, M., Ikeda, M., Ikeno, M., Ito, S., Ito, T., Ito, Y., Iwabuchi, A., Kamiya, K., Karasawa, W., Katagiri, S., Kikuta, A., Kobayashi, N., Kono, I., Machita, K., Maehara, T., Mizuno, H., Mizubayashi, T., Mukai, Y., Nagasaki, H., Nakashima, M., Nakama, Y., Nakamichi, Y., Nakamura, M., Namiki, N., Negishi, M., Ohta, I., Ono, N., Saji, S., Sakai, K., Shibata, M., Shimokawa, T., Shomura, A., Song, J., Takazaki, Y., Terasawa, K., Tsuji, K., Waki, K., Yamagata, H., Yamane, H., Yoshiki, S., Yoshihara, R., Yukawa, K., Zhong, H., Iwama, H., Endo, T., Ito, H., Hahn, J.H., Kim, H.I.,

Eun, M.Y., Yano, M., Jiang, J., and Gojobori, T. (2002). The genome sequence and structure of rice chromosome 1. Nature **420,** 312-316.

Shepard, J., Reick, M., Olson, S., and Graveley, B.R. (2002). Characterization of U2AF(26), a splicing factor related to U2AF(35). Mol Cell Biol **22,** 221-230.

Tassone, F., Villard, L., Clancy, K., and Gardiner, K. (1999). Structures, sequence characteristics, and synteny relationships of the transcription factor E4TF1, the splicing factor U2AF35 and the cystathionine beta synthetase genes from Fugu rubripes. Gene **226,** 211-223.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22,** 4673-4680.

Tian, M., and Maniatis, T. (1993). A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. Cell **74,** 105-114.

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

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature **409,** 832-833.

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

Webb, C.J., and Wise, J.A. (2004). The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. Mol Cell Biol **24,** 4229-4240.

Weigel, D., and Glazebrook, J. (2002). Arabidopsis: A laboratory manual. (New York: Cold Spring Harbor Laboratory Press).

Wentz-Hunter, K., and Potashkin, J. (1996). The small subunit of the splicing factor U2AF is conserved in fission yeast. Nucleic Acids Res **24,** 1849-1854.

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

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296,** 79-92.

Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc Natl Acad Sci U S A **86,** 9243-9247.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. Nature **355,** 609-614.

Zhang, M., Zamore, P.D., Carmo-Fonseca, M., Lamond, A.I., and Green, M.R. (1992). Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. Proc Natl Acad Sci U S A **89,** 8769-8773.

Zorio, D.A., and Blumenthal, T. (1999). U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. RNA **5,** 487-494.

Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes Dev **10,** 1356-1368.

Figure Legends

Figure 1. RT-PCR analysis of AUSa and AUSb transcripts. Panel A shows the gel pictures of RT-PCR results. RNA tissues include seedling, leaf before flowering (LeafBF),

leaf after flowering (LeafAF), meristem after flowering (MeriAF), root after flowering (RootAF), stem, flower and silique. NoRT is the negative control, where equal amount of different tissue RNAs were mixed and used in the RT-PCR reaction without adding reverse transcriptase. Genome is the genomic DNA positive control. The arrows point to non-canonical alternative splicing isoforms. Panel B describes the non-canonical alternative splicing pattern. The solid boxes represent exons, and lines represent introns. Numbers indicate the sizes of the corresponding introns.

Figure 2. Expression levels of AUSa and AUSb. Real time RT-PCR was performed to compare the expression levels of AUSa and AUSb in different tissues. The RNA tissues used are described in the legend to Figure 1. The level of AUSa in leafBF was arbitrarily chosen as calibrator, and other expression levels were compared with it. Relative expression level is indicated by bar heights. The thin lines above the bars represent the standard error among three experimental repeats.

Figure 3. GUS staining patterns for AUSa and AUSb promoters. The tentative AUSa or AUSb promoters were linked to the GUS gene. A-E, M, P, S AUSa; F-L, N, Q, T, AUSb; O, R, U, CaMV35S promoter control; A,B,F,G, whole seedlings in different stages; C, I, J, leaves and trichomes; D, M-O, flowers and inflorescences; E, silique; K, leaf primordial and meristem; L, anther and pollens; H, P-U, primary roots.

Figure 4. Cellular localization of AUSa and AUSb proteins. The open reading frame of AUSa or AUSb was fused with GFP in frame. Nuclear localization was observed for both

AUSa and AUSb. A, D, AUSa; B, E, F, AUSb; C, GFP control; A-C, young roots; D-E, leaf epidermal and guard cells, with top picture taken without UV excitation and bottom picture with excitation; F: trichome. Inserts in the upper right corners of A-C are images taken by confocal microscopy and show the uneven distribution of AUSa and AUSb in the nucleus. GFP control is targeted to both nucleus and cytoplasm.

Figure 5. AUSa and AUSb expression levels in AUSa/b mutants. Real time RT-PCR results on AUSa/b mutants and wild type (WT). The level of AUSa in WT was chosen as calibrator. Bars indicate the relative expression levels. The thin lines above the bars represent standard errors.

Figure 6. Phenotypes of AUSa and AUSb mutants. Pleiotropic phenotypes are observed in AUSa/b mutants. A, all AUSa and AUSb mutants are late flowering; B, leaf morphology changes in mutants, with the upper panel showing the changes in rosette leaves, and the lower panel showing cauline leaves. C, flower shape changes in AUSa mutants as well as in AUSb mutants (not shown); D, main stem stops growing in some AUSa mutants. Some AUSb mutants also show the same phenotype. E, the shape of siliques is changed in AUSa mutants. See text for detailed descriptions of the phenotypes.

Figure 7. Effects of AUSa and AUSb on splicing patterns. Upper panel, RT-PCR reveals changes in constitutive splicing but not in alternative splicing patterns. The left panel shows the results for the AtDBR1 gene, with extra bands produced in AUSa and AUSb mutants (indicated by arrows). The right panel is an example of an alternatively spliced gene. M:

Marker. The lower panel shows the abundance level changes for the FCA-β isoform and FLC gene transcript.

Figure 8. Gene structure and alternative splicing patterns for maize and rice U2AF1 genes. Boxes represent exons. Lines between boxes represent introns. Green bars indicate the ORF. Numbers above the lines and exons represent the sizes of introns or exons. Splicing of certain introns is represented by two lines starting from the border of the intron and meeting at the resulting transcript isoform.

Figure 9. C-terminus alignment of plant U2AF1 homologs. The SERE domain is indicated by the red box. Conserved amino acid residues are highlighted by black background. Residues not identical but similar to the conserved one are highlighted by green background. Abbreviations: ZUSa and ZUSb, maize homologs; AUSa and AUSb, Arabidopsis homologs; os: rice; ts: wheat; hv: barley; gm: soybean; mt: Medicago; le: tomato; st: potato; ga: cotton; pt: pine; cr: unicellular algae; hm: human; dm: Drosophila melanogaster; ce: nematode; sp: fission yeast.

Figure 10. Phylogenetic tree of U2AF1 homologs. Homologs grouped in the same clade are highlighted by the same background color. Branch lengths indicate distances. Numbers on the branch are bootstrap values of confidence in the displayed branches (see Materials and Methods for details). Abbreviations: as in the legend to Figure 9; ms: mouse; tk: bony fish; gg: chicken; lm: *Leishmania major*; ec: *Encephalitozoon cuniculi*.

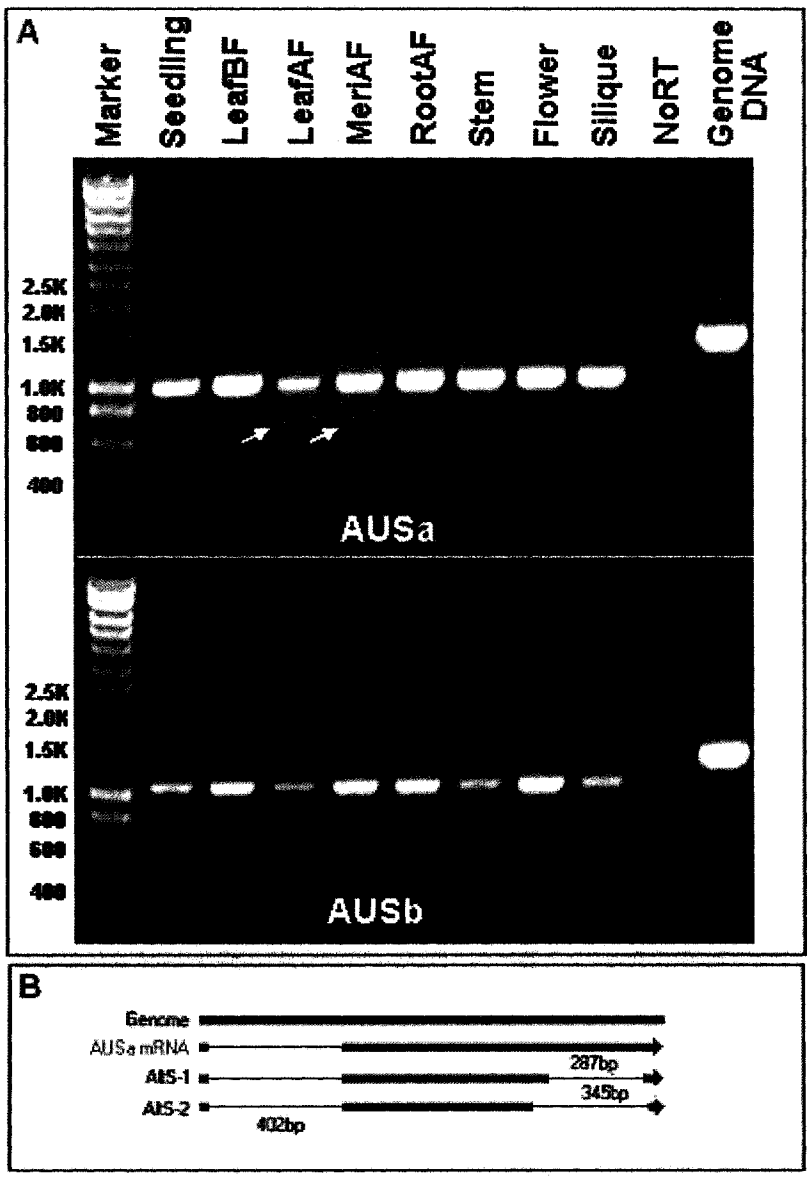

A
Marker
Seedling
LeafBF
LeafAF
MeriAF
RootAF
Stem
Flower
Silique
NoRT
Genome DNA
2.5K
2.0K
1.5K
1.0K
800
600
400
AUSa
AUSb
B
Genome
AUSa mRNA
AtS-1
AtS-2
287bp
345bp
402bp

Figure 1

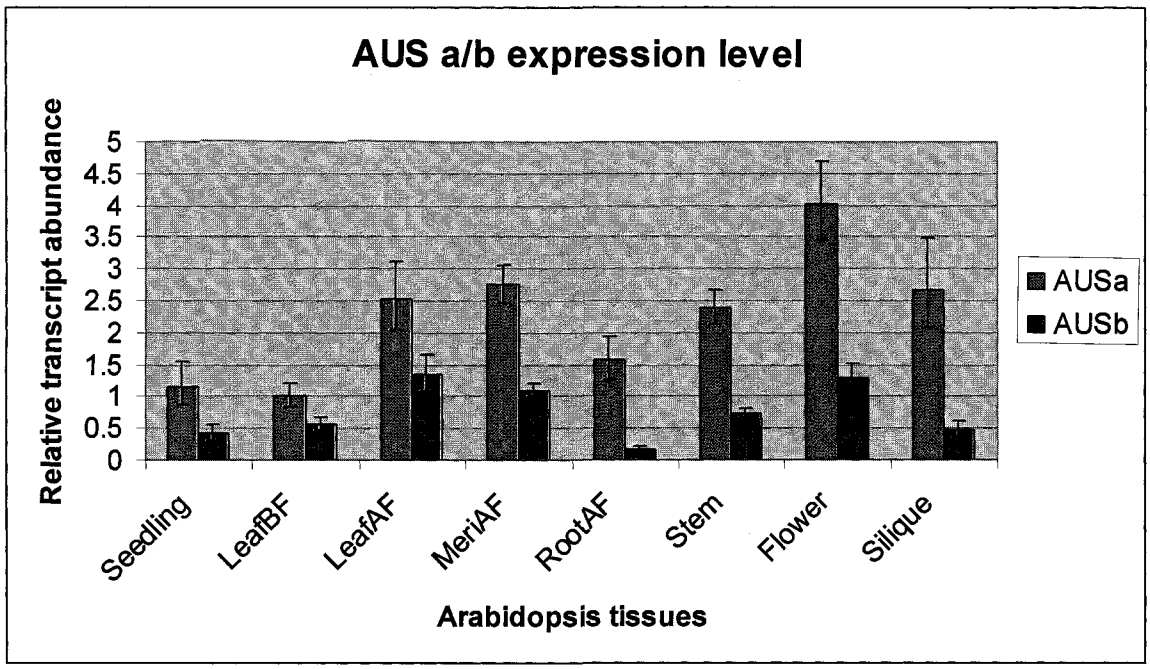

AUS a/b expression level
Relative transcript abundance
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Seedling
LeafBF
LeafAF
MeriAF
RootAF
Stem
Flower
Silique
Arabidopsis tissues
AUSa
AUSb

Figure 2

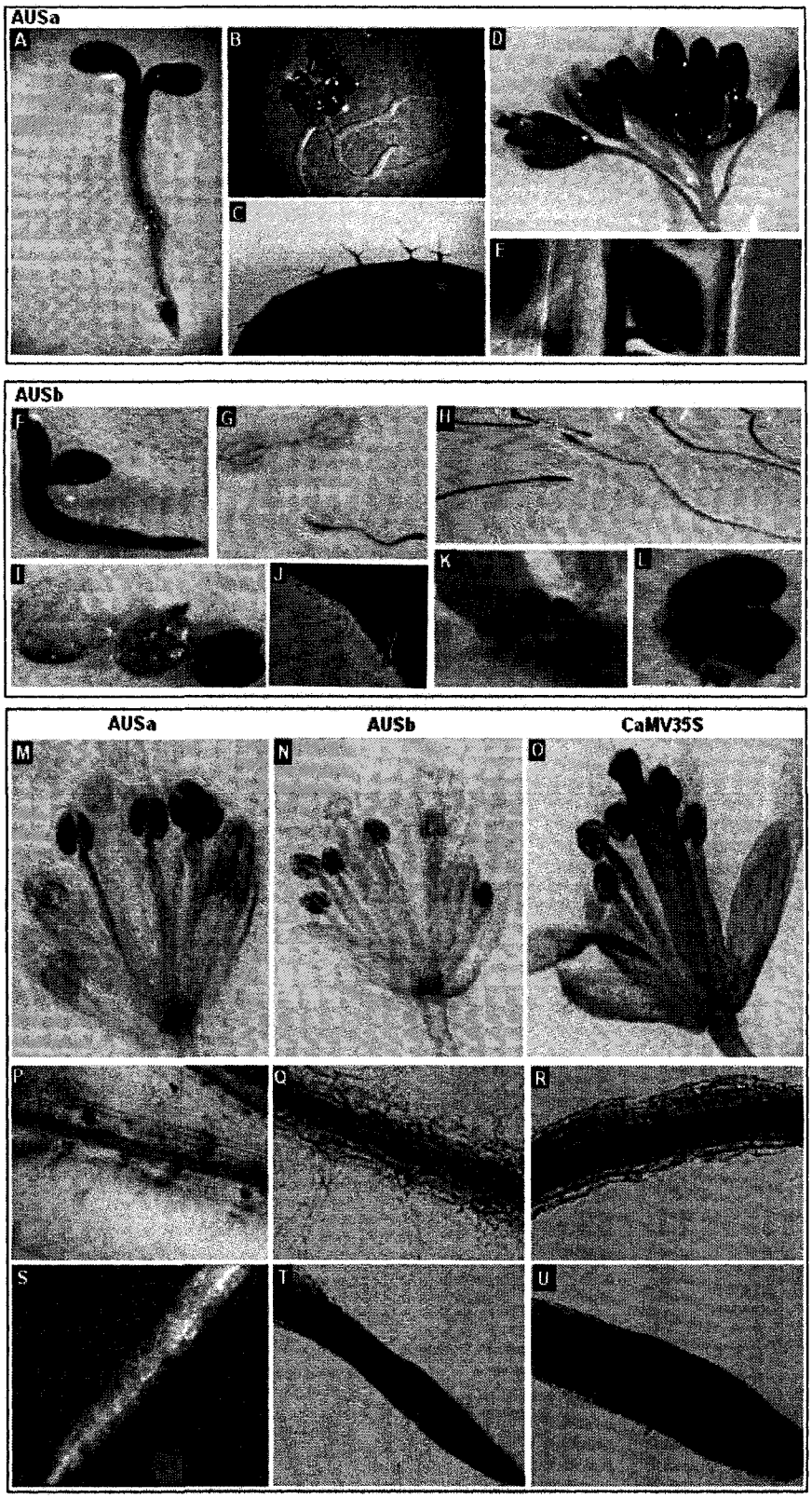


Figure 3

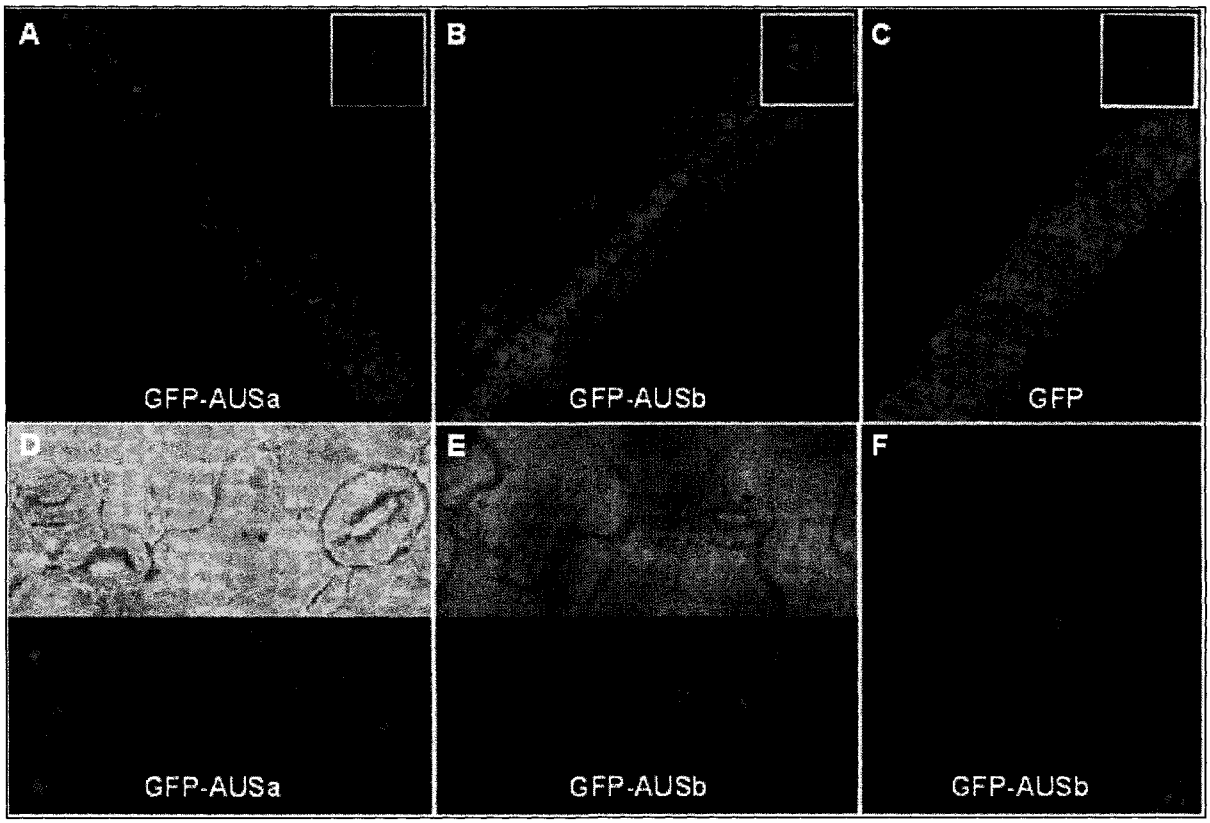


Figure 4

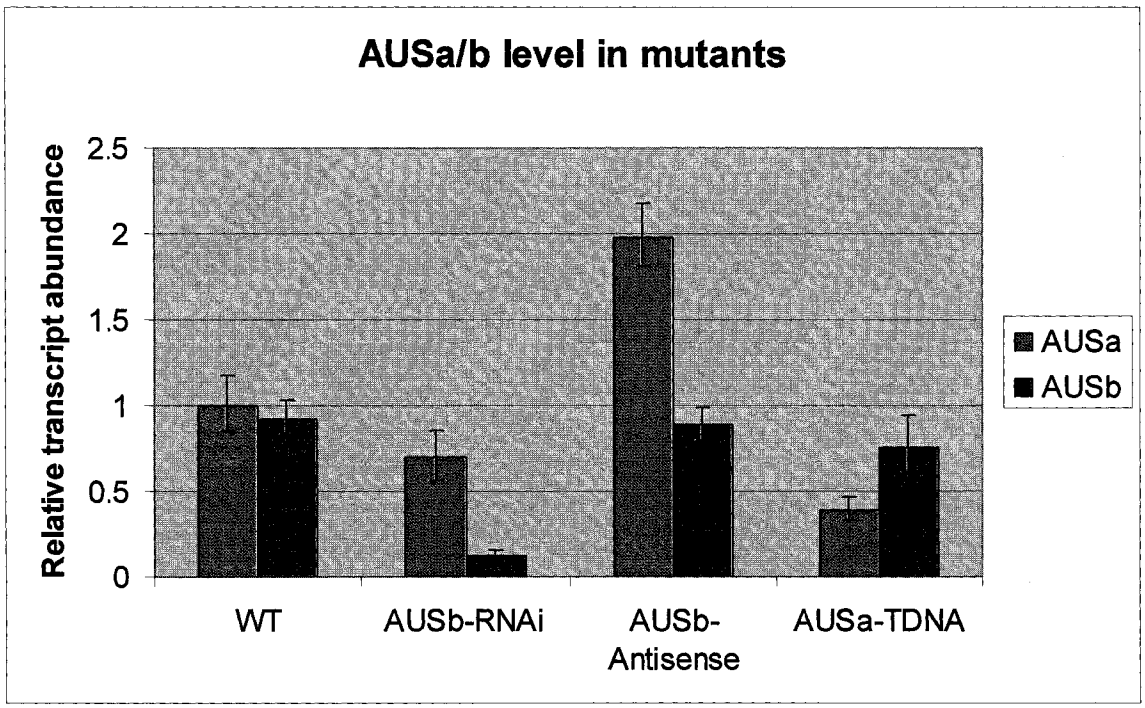

AUSa/b level in mutants
Relative transcript abundance
2.5
2
1.5
1
0.5
0
WT
AUSb-RNAi
AUSb-Antisense
AUSa-TDNA
AUSa
AUSb

Figure 5

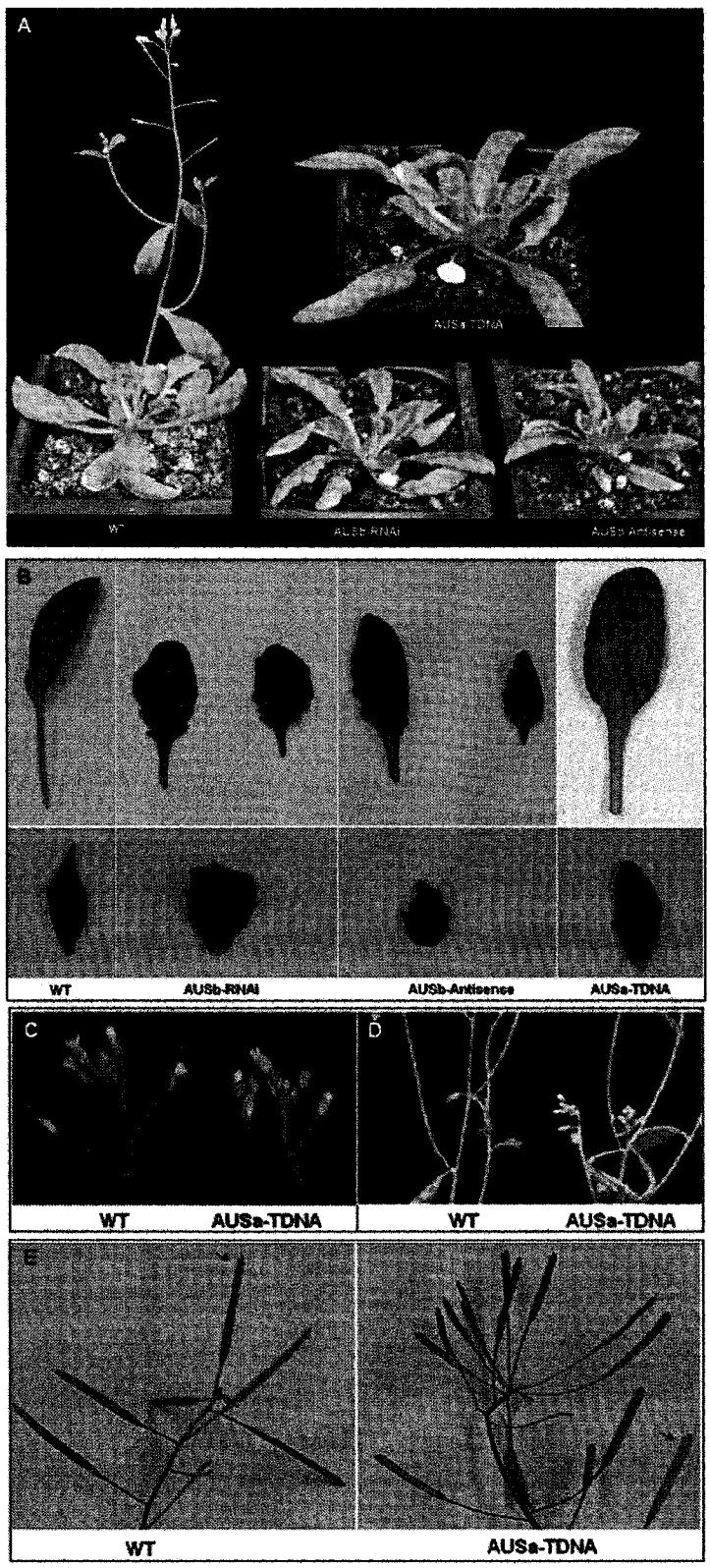

A
AUSa-TDNA
WT
AUSb-RNAi
AUSb-Antisense
B
WT
AUSb-RNAi
AUSb-Antisense
AUSa-TDNA
C
WT
AUSa-TDNA
D
WT
AUSa-TDNA
E
WT
AUSa-TDNA

Figure 6

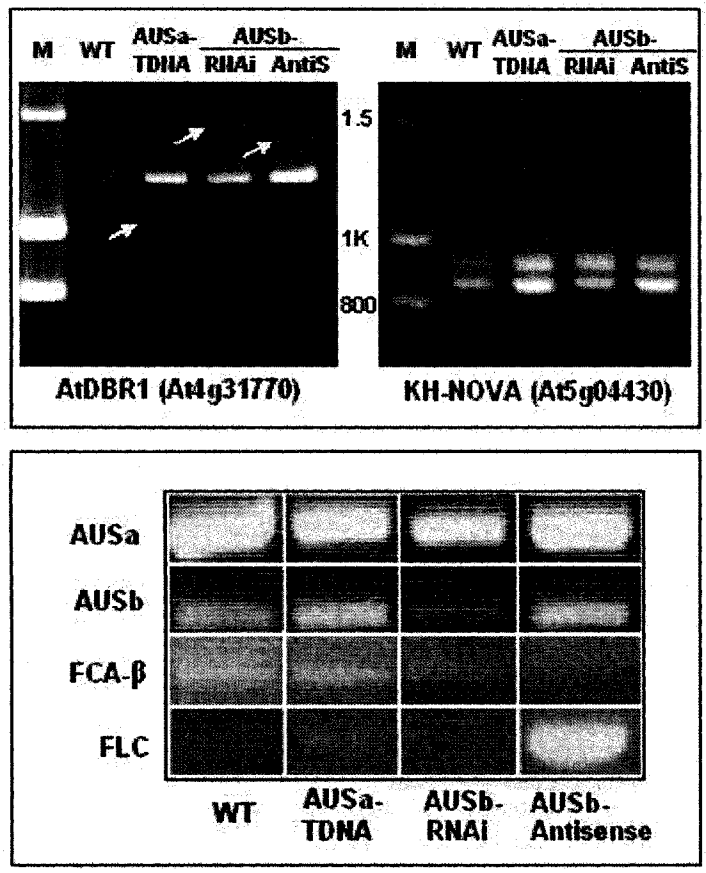

M
WT
AUSa-TDNA
AUSb-RNAi
AUSb-AntiS
1.5
1K
800
AtDBR1 (At4g31770)
KH-NOVA (At5g04430)
AUSa
AUSb
FCA-β
FLC
WT
AUSa-TDNA
AUSb-RNAi
AUSb-Antisense

Figure 7

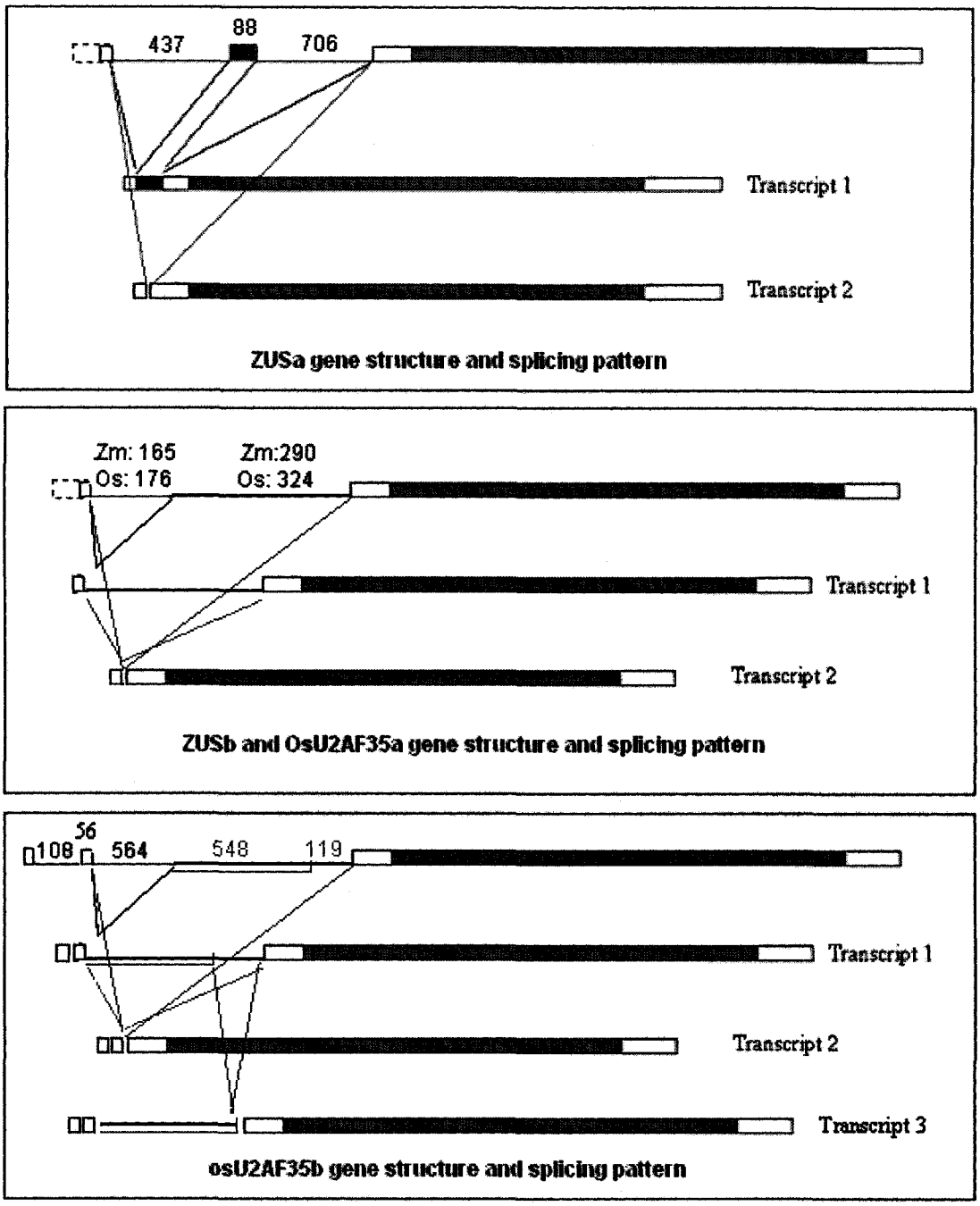

437
88
706
Transcript 1
Transcript 2
ZUSa gene structure and splicing pattern
Zm: 165
Os: 176
Zm:290
Os: 324
Transcript 1
Transcript 2
ZUSb and OsU2AF35a gene structure and splicing pattern
56
108
564
548
119
Transcript 1
Transcript 2
Transcript 3
osU2AF35b gene structure and splicing pattern

Figure 8

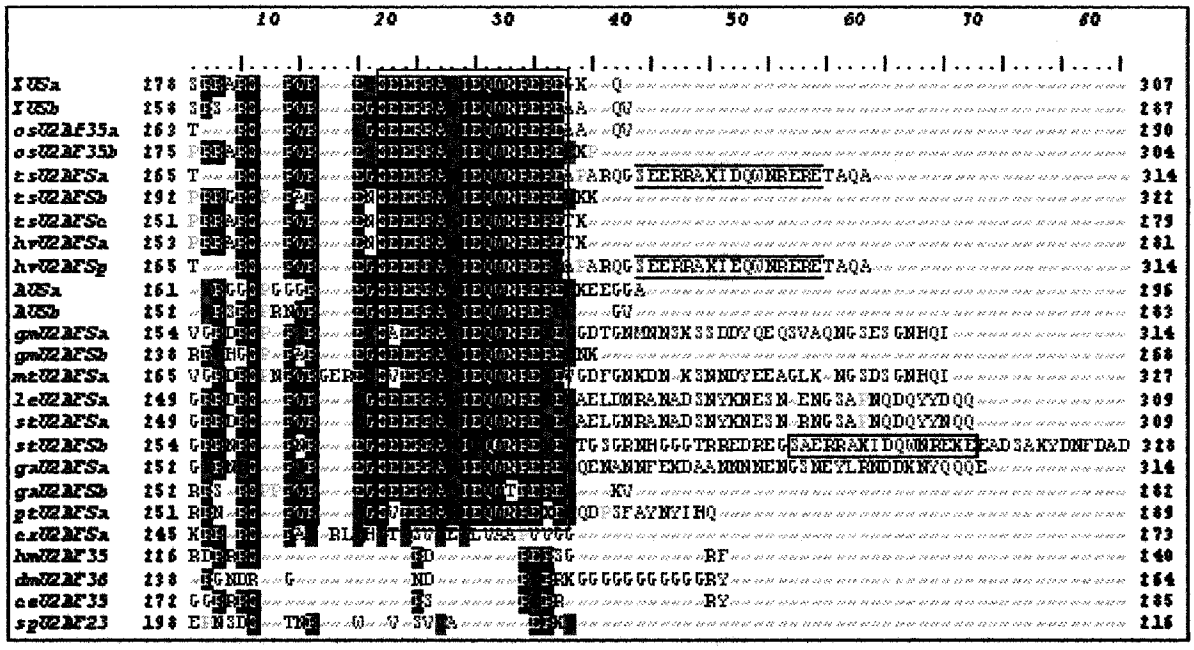


Figure 9

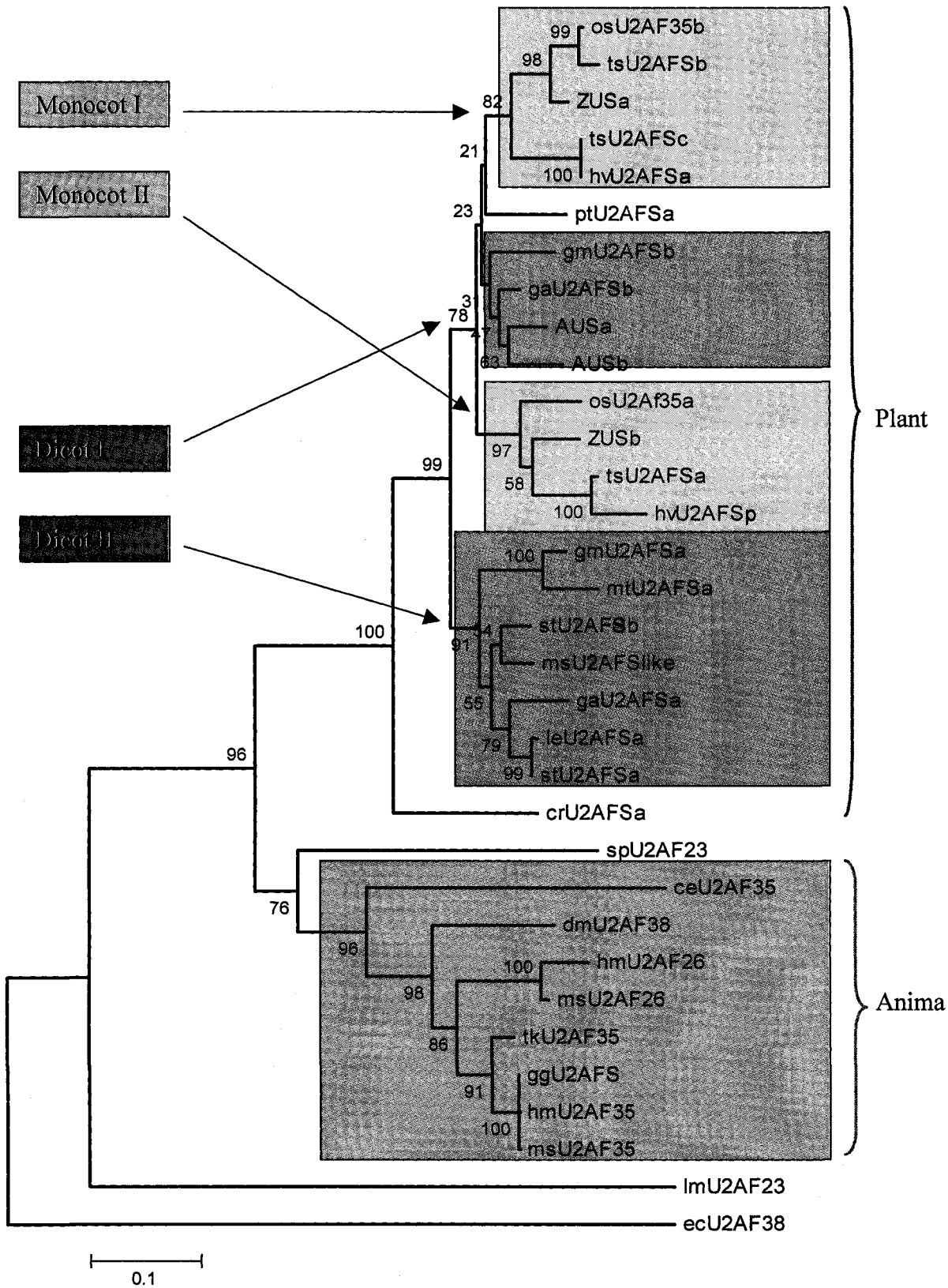

Monocot I
Monocot II
Dicot I
Dicot II
osU2AF35b
tsU2AFSb
ZUSa
tsU2AFSc
hvU2AFSa
ptU2AFSa
gmU2AFSb
gaU2AFSb
AUSa
AUSb
osU2Af35a
ZUSb
tsU2AFSa
hvU2AFSp
gmU2AFSa
mtU2AFSa
stU2AFSb
msU2AFSlike
gaU2AFSa
leU2AFSa
stU2AFSa
crU2AFSa
spU2AF23
ceU2AF35
dmU2AF38
hmU2AF26
msU2AF26
tkU2AF35
ggU2AFS
hmU2AF35
msU2AF35
lmU2AF23
ecU2AF38
Plant
Anima
0.1

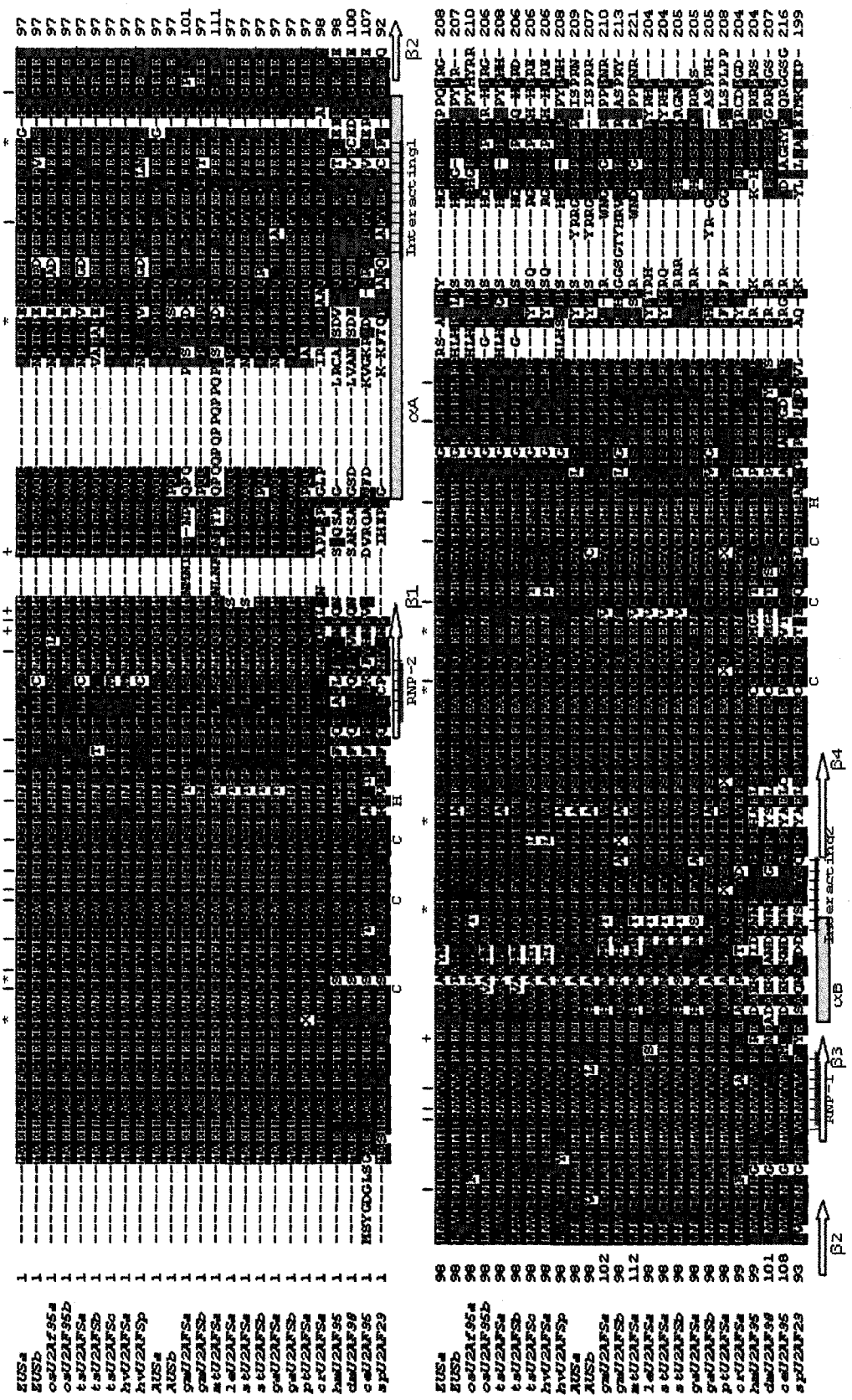
Figure 10

Supplementary Data

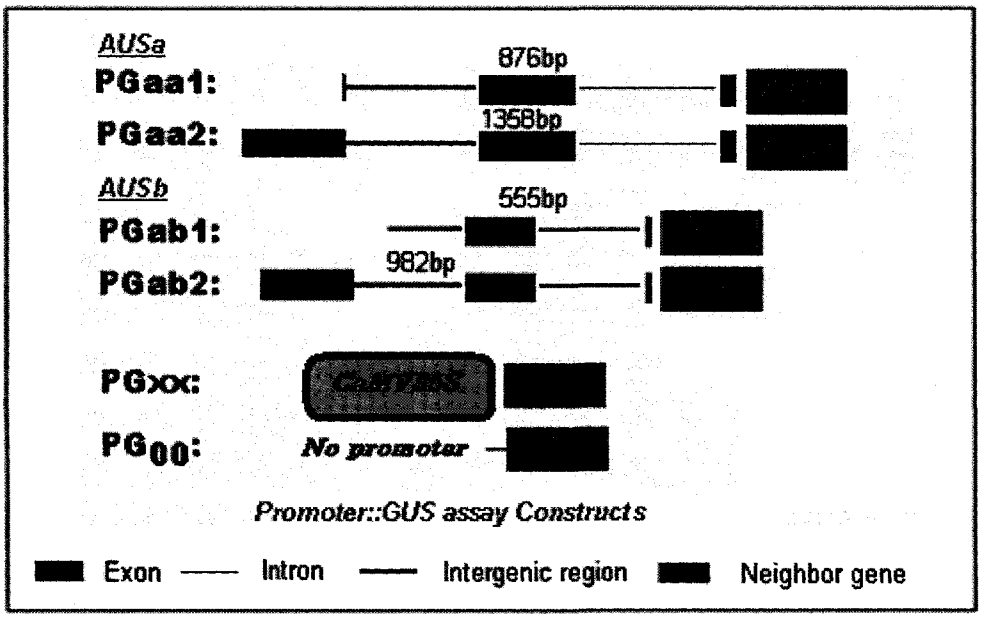
Supplementary Table 1. Primers used for RT-PCR and vector constructs. All primers used in RT-PCR, real time RT-PCR and vector constructions were as follows. The underlined letters show restriction enzyme digestion sites. Primers used in mutant RT-PCR to check alternative splicing pattern changes are listed in Supporting Table 1 in Chapter 3.

Usage	Genes (Construct)	Forward	Reverse
RT-PCR	AUSa	ATCCCACTCATCTCTGTAAC	GCTATGTGGTTTCTGCGTC
	AUSb	CGATAGCTTCTCTTCCACT	AGTTCGTGAGGCAAATGATG
Real Time	AUSa	GCGACCACTCACCCTCAGAT	GGCAGAATTTCCACCAGAATGT
	AUSb	ACCTCCTCGTCGGCGATAG	CTCTGCCATTCTCACCGAAGAA
	18S rRNA	ATTTCTGCCCTATCAACTTTCG	TGGATGTGGTAGCCGTTTCT
Promoter:: GUS	AUSa (PGaa1)	AAGCCCGGGGTTGTTTTGATTTTGACT	GCTCCTGCAGGGCAGAATTTCCACC
	AUSa (PGaa2)	AACCCCGGGATGTGCACACTGATAG	
	AUSb (PGab1)	CTTGAATTCCTCGTTATGTTATGAC	CTCCTGCAGTCTCACCTAAACATAACA
	AUSb (PGab2)	GATGAATTCGGCCAAACCTCTTATGA	
Antisense	AUSb	TTAGCCATGGATGGCAGAGCATTTAGCT	AACGCCATGGTTAAACTCCCTCATCACG
RNAi	AUSb	AGCTTAAGACGCGTATGGCRGAGCATTTRGCTTCA	TCCTTAAGTGTACATTAAACTCCCTCATCACGTTC
Mutant RT-PCR	FLC	GAACCCAAACCTGAGGATCA	TCCAGCAGGTGACATCTC
	FCA-γ isoform	GACGCTGGGAAGATGACACT	GCTTTTCCTCTCCTTGACTAAT
	FCA-β isoform		CTGGCAAGCATAAGCAATAAG
	FPA	TCTGGGATTGGAATTTTGGAT	AGGCAGAGGGTAGAAGAGAT
	AtDBR1	ATGGAGACCTAGACAATGTG	GCTTCCTCCATCTCTTCTAT

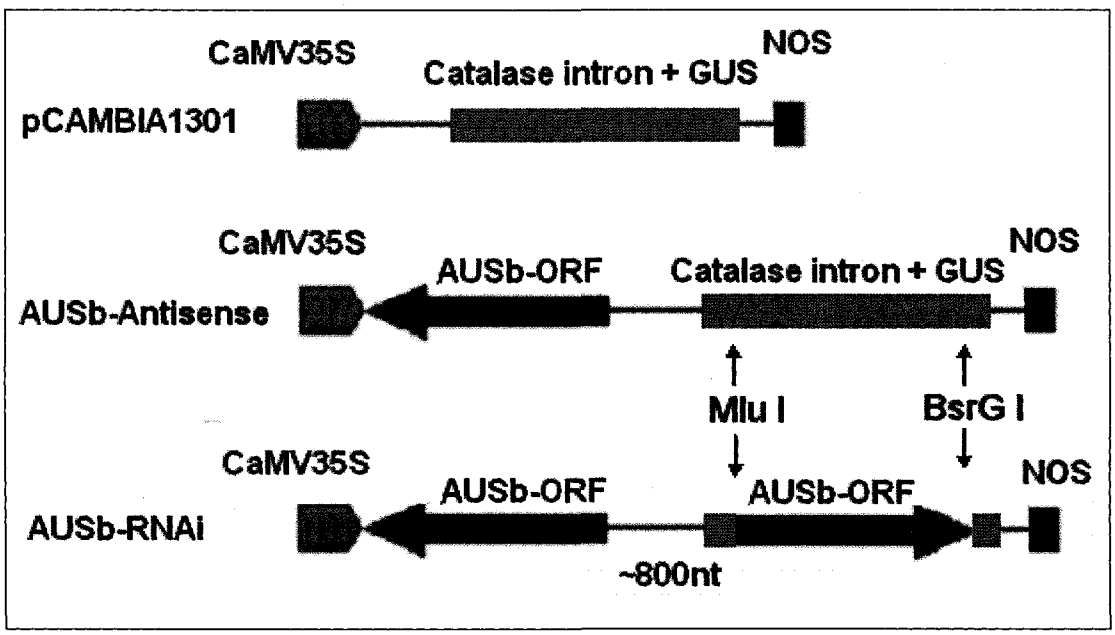
Supplementary Figure 1. N-terminal alignment of U2AFS proteins. Conserved domains are indicated at the bottom of the alignment. Boxes and arrows indicate the structure of human U2AF[35]. | indicate the fully conserved amino acids in all homologs including the ones not used in the alignment. * indicates plant specific residues; + indicates higher plant specific residues. Abbreviations: ZUSa and ZUSb: Maize homologs; AUSa and AUSb: Arabidopsis homologs; os: rice; ts: wheat; hv: barley; gm: soybean; mt: Medicago; le: tomato; st: potato; ga: cotton; pt: pine; cr: unicellular algae; hm: Human; dm: *Drosophila melanogaster*; ce: nematode; sp: fission yeast.


RNP-2
RNP-1
Interacting1
Interacting2
αA
αB
β1
β2
β3
β4

Supplementary Figure 2. Constructs for Promoter::GUS assays. PGaa: Promoter::GUS assay for AUSa; PGab: Promoter::GUS assay for AUSb. PGxx: Positive control construct; PG00: No promoter construct used as negative control. Solid boxes represent the predicted exons before the start codon of the AUSa or AUSb gene (the black boxes shows an exon from an upstream gene). Black (bold) lines depict intergenic regions. Blue (fine) lines delineate introns. The picture is not drawn to scale.



Supplementary Figure 3. Constructs for AUSb-antisense and AUSb-RNAi assays. The basal vector is pCAMBIA1301, which includes the CaMV35S strong promoter, GUS gene, and NOS terminator. To construct the AUSb-antisense vector, the AUSb-ORF was inserted in reverse direction immediately downstream the of CaM35S promoter. In the AUSb-RNAi vector, another AUSb-ORF replaces the central region of the GUS gene in the antisense vector after Mlu I and BsrG I digestion and follow-up ligation. The picture is not drawn to scale.



Chapter 5: General Conclusion

In this study, we systematically surveyed the splicing machinery in Arabidopsis, identified and compared alternative splicing events in both Arabidopsis and rice, and experimentally characterized U2AF1 homologs in Arabidopsis. Two databases were constructed for the community to use and will facilitate studies of plant splicing mechanisms.

In Arabidopsis, a total of 74 snRNA genes and 395 genes encoding splicing related proteins were identified by sequence comparison and motif searches, including the previously elusive U4atac snRNA gene. Most of the genes have not been experimentally studied. Our data show that about 50% of the splicing related genes are duplicated in Arabidopsis. The duplication ratios for splicing regulators are even higher, indicating that the splicing mechanism is generally conserved among plants, but that the regulation of splicing may be more variable and flexible, thus enabling plants to respond to their specific environments. Classification of these genes and detailed information on gene structure, alternative splicing, gene duplications, and phylogenetic relationships are made accessible as a comprehensive database of Arabidopsis Splicing Related Genes (ASRG) at http://www.plantgdb.org/SRGD/ASRG/.

Alternative splicing in plants was found here to be more prevalent than previously expected. Both Arabidopsis and rice have about 22% of the expressed genes being alternatively spliced, and in both about 55% AS events are intron retention. The consistent high frequency of IntronR suggests prevalence of splice site recognition by intron definition in plants. 40% of

Arabidopsis AS genes are also alternatively spliced in rice, with some examples strongly suggesting a role of the AS event as an evolutionary conserved mechanism of post-transcriptional regulation. We created a comprehensive web-interfaced database to compile and visualize the evidence for alternative splicing in plants (ASIP, available at: http://www.plantgdb.org/ASIP/).

Two copies of U2AF1 (AUSa and AUSb) were identified and characterized in *Arabidopsis*. AUSa expressed at a higher level than AUSb in most tissues. Differences in the expression patterns of AUSa and AUSb in roots were also revealed. Altered expression levels of AUSa or AUSb cause pleiotropic phenotypes (including flowering time, leaf morphology, flower and silique shape). Extra RT-PCR products (possibly novel splicing isoforms) were detected for some genes (AtDBR1) in both AUSa and AUSb mutants, indicating the importance of AUSa/b for correct splice site recognition of some genes. A novel C-terminal domain (SERE) is highly conserved in all seed plant protein homologs, suggesting it may have an important function specific to higher plants.

In conclusion, similarities as well as differences were revealed between the splicing mechanisms in plants and mammalians. Distinct features also exist between Arabidopsis and rice, demonstrating that each organism may have evolved special mechanisms to ensure the efficient and accurate splicing in different environments.

Acknowledgements

I am grateful to my major professor, Dr. Volker Brendel, for his guidance, patience and generosity throughout my graduate studies. What I like the most is his matrix management. Volker gave me enough freedom on developing research projects and learning necessary techniques, which will be extremely beneficial to my future career. I am also indebted to my committee members, Drs. Phil Becraft, Thomas Peterson, Xun Gu and Shashi Gadia, who are always around and helpful whenever I encounter problems.

During these years, the whole VB lab (i.e., the matrix) gave me lots of help. The matrix agents, Wei Huang (lab manager) and Mike Brekke (system support) helped me a lot in technical support. Many matrix elements, including Drs. Wei Zhu (TIGR), Shailesh Lal (Oakland Univ.), Qunfeng Dong, Carolyn Lawrence, Xiaokang Pan and graduate students Shannon Schlueter, Michael Sparks and Yuanbin Ru, interacted with me actively and gave me many thoughtful suggestions. Matrix assistant Zhen Ni (Jenny) Li and Tiffanie Kuhn did much tedious bench work for me.

I would like to give specially thank to my wife, Yanwen Xiong, for her continuous love and encouragement. My parents and brother understand and support me all the time, which is really the most precious treasure of mine. My friends Feng Zhang and Hongwu Jia discussed with me quite often about course studies, research design and career goals. Also Xueyuan Cao, Chuan Shen and Hongtao Qin helped me on some experiments.

Without assistance from these friends, it would not have been possible to have this thesis.