Assessing the impact of alternative splicing on the diversity and evolution of the proteome in plants

Edouard I. Severing

Thesis committee

Thesis supervisor

Prof. dr. W.J. Stiekema
Professor of Genome Informatics
Wageningen University, the Netherlands
Professor of Bioinformatics
University of Amsterdam, the Netherlands

Thesis co-supervisor

Dr. R.C.H.J. van Ham
Vice-President Bioinformatics and Modeling
Keygene N.V., Wageningen, the Netherlands

Other members

Prof. dr. G.C. Angenent, Wageningen University
Prof. dr. M.A. Huijnen, Radboud University, Nijmegen
Dr. B. Snel, Utrecht University
Dr. J.E. Kammenga, Wageningen University

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences.

Assessing the impact of alternative splicing on the diversity and evolution of the proteome in plants

Edouard I. Severing

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday, 7 December 2011
at 11.00 a.m. in the Aula.

Edouard I. Severing
Assessing the impact of alternative splicing on the diversity and evolution of the proteome in plants

Thesis, Wageningen University, Wageningen, NL (2011)
With references, with summaries in Dutch and English
ISBN: 978-94-6173-096-1

This thesis is dedicated to my dear parents Harold and Diana.

Contents

Chapter 1

General Introduction

RNA processing

It is now well established that the details of the information flow from DNA to proteins (central dogma (Crick, 1970)) differ between prokaryote and eukaryote systems. While transcription and translation co-occur in prokaryotes (Miller et al., 1970), in eukaryotes they are spatially separated in the nucleus and cytoplasm, respectively. The primary mRNA molecule (pre-mRNA) that is produced by transcription of a eukaryotic gene is processed before being exported into the cytosol. Two of these processing steps involve the addition of a methyl cap (Gingras, 2009) to the N-terminus of the pre-mRNA molecule and modification of the 3' terminus through the addition of a poly-A tail (Li and Hunt, 1997) (Figure 1). Many pre-mRNA molecules can be divided into regions called exons and introns that are included and excluded from the mature mRNA, respectively (Gilbert, 1978). The removal of intronic sequences from the pre-mRNA, a process called splicing, is often essential in order for an mRNA molecule to serve as a template for a polypeptide (Figure 1). Splicing reactions in the nucleus are catalyzed by spliceosomes which are large complexes of several proteins and small nuclear RNAs that are actively assembled on pre mRNA molecules (Matlin et al., 2005).

Not long after the initial discovery of introns and splicing (Berget et al., 1977; Chow et al., 1977) it was realized that variation can exist in the regions of pre-mature mRNAs from the same gene that were recognized as introns (e.g. Alt et al., 1980; Early et al., 1980). This variability of pre-mRNA splicing, which is now widely known as Alternative Splicing (AS), enables the production of multiple mRNA- and protein molecules from a single gene.

Genes and complexity

Sequencing of different genomes has revealed that the number of protein-coding genes does not correlate with the complexity of the corresponding organism (Blencowe, 2006). If the number of protein coding genes were an indicator for complexity than the nematode *Caenorhabditis elegans* (~19,000) (The C. elegans Sequencing Consortium, 1998) would be more complex than the fruit fly Drosophila (~13,600) (Adams et al., 2000), almost equally complex as humans (~20,000) (Clamp et al., 2007) and less complex than the plant species Arabidopsis (~26,000) and rice (~41,000) (Sterck et al., 2007).

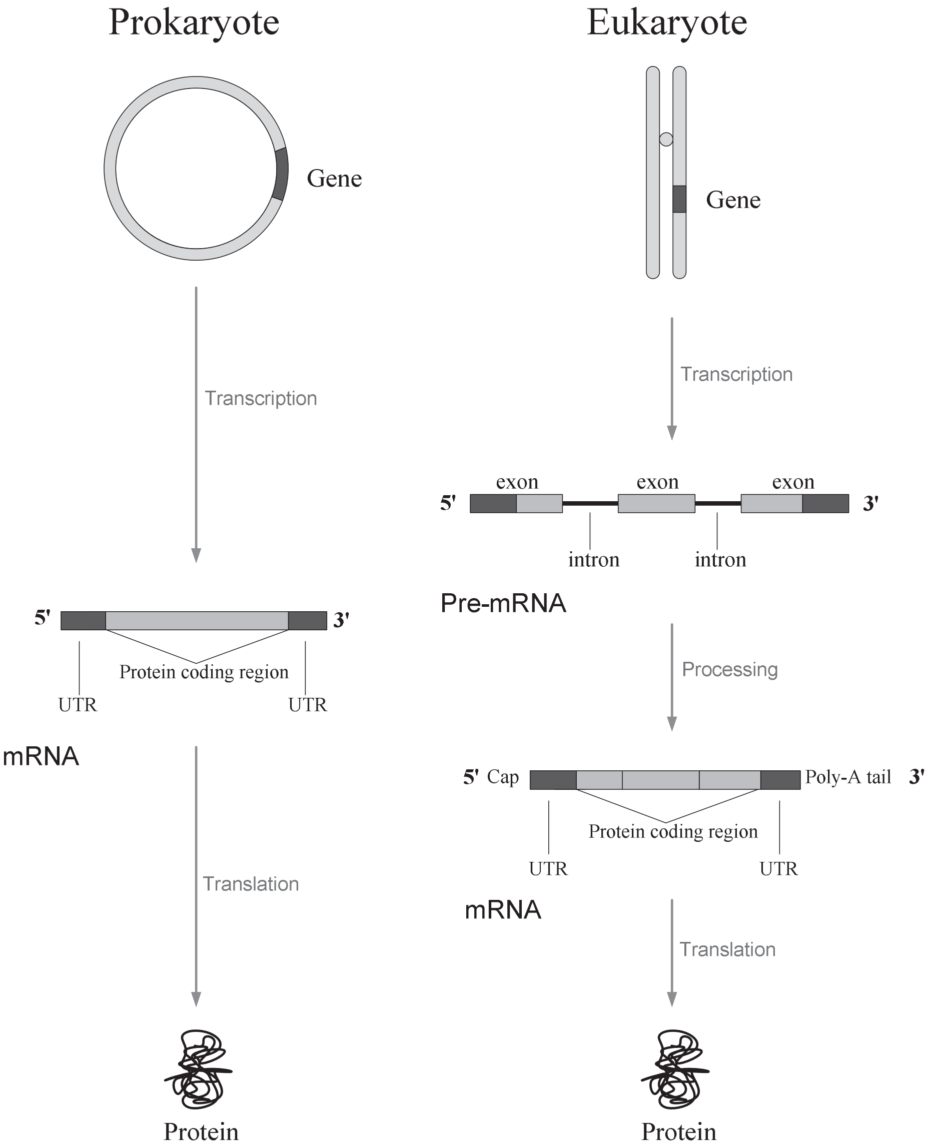

Figure 1. Genetic information processing. The mRNA molecule obtained by transcription of a prokaryotic protein coding gene contains a continuous protein coding region which can directly be translated into a protein. In eukaryotes, transcription results in a premature mRNA molecule which is further processed by for instance the addition of a 3’- poly-A tail and a 5’- cap. The protein coding region of a eukaryotic pre-mature mRNA is often divided into exons and introns that are included in - and excluded from the final mRNA molecule, respectively. After removal of the introns (splicing), the protein coding region becomes continuous.

Over the years, it became clear that in humans not just 5%, as previously thought (Sharp, 1994), but more than 90% (Pan et al., 2008; Wang et al., 2008) of multi-exon genes are subjected to AS. Mammals tend to have more alternative splicing than *C. elegans* and Drosophila (Nilson and Gravery 2010). In plants, around 20 to 40% of genes are capable of producing multiple transcripts (Campbell et al., 2006; Wang and Brendel, 2006; Filichkin et al., 2010; Zhang et al., 2010). While there are on average around seven AS events per alternatively spliced human gene (Pan et al., 2008), the majority of alternatively spliced genes from Arabidopsis and rice produce only two or three isoforms (Campbell et al., 2006). Perhaps the most spectacular case of alternatively spliced genes known to date is the *Dscam* gene from Drosophila. Through combining exons from four different clusters of mutually exclusive exons, a total of 38,016 different transcripts can be produced from this gene (Schmucker et al., 2000).

Several lines of evidence suggest that AS has been present very early in the evolution eukaryotes (Irimia et al., 2007) and its prevalence increases from unicellular eukaryotes to mammals (Artamonova and Gelfand, 2007). The common types of AS events have also changed along this evolutionary line, from failure to remove introns to frequent exclusion/inclusion of entire exons (Keren et al., 2010). Taken all together, these observations collectively suggest that AS is a key mechanism that contributes to the discrepancy between organismal complexity and the number of protein coding genes.

Basic splicing intron composition and splicing regulation

In order for introns to be recognized by the cell they need to have certain properties or signals that distinguish them from exons. The core splicing-signals which are typically found in introns are the 5’ (donor) splice site, the 3’ (acceptor) splice site, the branch-point sequence and the poly-pyrimidine tract (Figure 2). Recognition of splice-site pairs, which form the boundaries of introns, can both occur across exons (exon definition) and across introns (intron-definition) (Berget, 1995). That is, either exons or introns are the main units recognized by the spliceosome during the initial steps of the splicing process. Given that introns and not exons are the segments which are finally removed from the molecule, complexes that are assembled through exon definition need to be converted to interaction across introns at a later stage. It has been suggested that intron definition is common when the introns are short and that exon definition is more common when exons are short and introns are long (Fox-Walsh et al., 2005). Failure to perform accurate splicing leads to exclusion of entire exon(s) under the exon-definition model and to the retention of intron(s) under the intron-definition model (Barbazuk et al., 2008). Indeed, the results of a recent study on AS in a wide range of species hint towards a relationship between intron size and the prevalence of exon skipping and intron retention (McGuire et al., 2008).

Although the core splicing signals are necessary, they are not always sufficient for recognizing intron boundaries (Lim and Burge, 2001; Cartegni et al., 2002). It has for instance been demonstrated that mammalian transcripts have many sequences that resemble true splice sites but that are not used as such (Sun and Chasin, 2000). There are many individual biological parameters that have an influence on the splicing process. For instance, in plants, AU-rich sequences are important for the definition of introns (Morello and Breviario, 2008). Differential usage of closely spaced splice sites can simply be the result of a stochastic process. In these cases, the resulting ratio between the abundances of the transcript isoforms is governed by the strength of the competing splice sites (Hiller and Platzer, 2008) (Figure 3a).

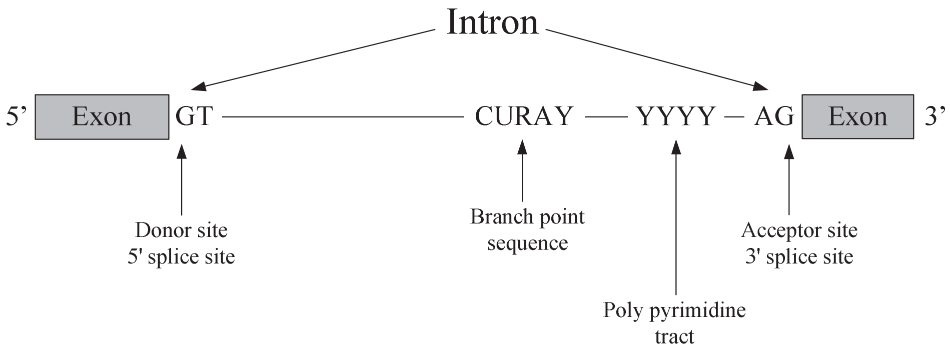

Figure 2. The basic intron anatomy. A schematic overview of the four most common sequence elements of eukaryotic introns. Although the GT-donor and AG-acceptor di-nucleotides are the canonical splice-site boundaries, a small subset of introns uses different donor and acceptor site sequences.

The usage of a particular splice site can be promoted by exonic- (ESE) or intronic- (ISE) splicing enhancers and suppressed by exonic- (ESS) or intronic- (ISS) splicing silencers (Wang and Burge, 2008). Binding of these *cis*-acting enhancer and silencing elements by *trans*-acting factors can for instance promote or inhibit interactions between core splicing elements and spliceosomal components (Matlin et al., 2005) (Figure 3b).
A number of other biological factors which have also been shown to influence the splicing process include secondary structures of RNA-molecules (Graveley, 2005; Shepard and Hertel, 2008), rate of transcription (Kornblihtt, 2006) and chromatin-structure (Keren et al., 2010). The combinatorial effect of the different biological features that determine the final splicing pattern of a pre-mRNA molecule is referred to as the "splicing code" (Blencowe, 2006; Hertel, 2008). It has been suggested that splicing patterns can largely be predicted using the DNA sequence alone (Shepard and Hertel, 2010). In fact, significant progress has been made in formulating an *in silico* "splicing code" that is capable of predicting tissue-specific splicing patterns (Barash et al., 2010).

Development of experimental methods for alternative splicing profiling

The first generation of genome-wide AS profiling studies (e.g. Mironov et al., 1999; Modrek et al., 2001; Campbell et al., 2006; Wang and Brendel, 2006) were performed with cDNA and expressed sequence tags (ESTs) that were obtained using the Sanger method (Sanger et al., 1977). This sequencing method was not only expensive (Wang et al., 2009) but also suffered from limited gene coverage (Modrek et al., 2001), bias towards the 3'- end of transcripts (Lee and Wang, 2005) and other drawbacks (Modrek and Lee, 2002). Nevertheless, this technology has provided the data from which a great part of our current understanding of splicing has been derived.

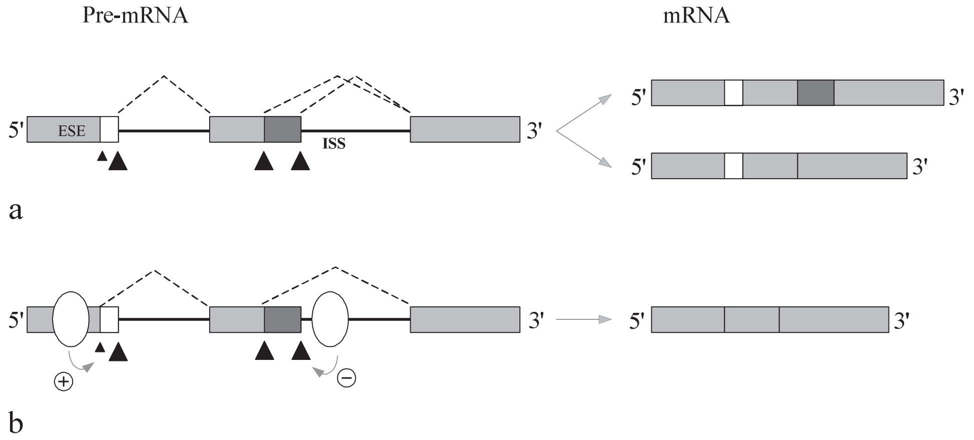

Figure 3. Splice-site regulation. (**a**) a pre-mRNA containing two pairs of competing splice-sites (black triangles). There are two *cis*-acting elements on this molecule: an exonic splicing enhancer (**ESE**) and an intronic splicing silencer (**ISS**). When these elements are unbound, the splice pattern (dashed lines) is governed by the strength of the competing splice sites (size of triangles). As a result of this competition, two different mature mRNA isoforms will be observed. (**b**) When the ESE is bound by a splicing regulator (white ellipse) the usage of the adjacent weak splice site is promoted. When the ISS is bound by a splicing regulator, usage of the adjacent splice site is repressed. As a result of this splice site usage regulation, a single mRNA is predominantly produced.

The AS-profiling studies belonging to the second generation were performed using microarray platforms. These platforms proved to be well suited for confirming known and detecting novel and tissue specific AS events (e.g. Johnson et al., 2003; Pan et al., 2004; Ner-Gaon and Fluhr, 2006; Clark et al., 2007). Microarrays were not *per se* biased towards transcript ends and against low abundant transcripts (Johnson et al., 2003). In addition, this technology has been demonstrated to be feasible for quantitative analysis of AS (Lee and Wang, 2005). Another appealing feature of microarrays was that they enabled large numbers of transcripts to be monitored simultaneously in a single experiment. However, despite their success, microarrays also suffered from a number of limitations such as their reliance on *a priori* knowledge of the genome sequence (including gene structures) and their susceptibility to cross hybridization (Hallegger et al., 2010).

Currently, there is an increasing number of third generation AS profiling studies appearing (e.g. Pan et al., 2008; Wang et al., 2008; Filichkin et al., 2010; Nilsson et al., 2010; Siegel et al., 2010; Zhang et al., 2010) that use data from next generation sequencing (NGS) technologies such as 454 (Roche), Illumina (formerly Solexa sequencing) and Solid (Applied Biosystems). Because RNA-seq (Nagalakshmi et al., 2008) (sequencing of RNA using NGS technologies) is not a hybridization-based method, it can be used for profiling of the transcriptomes of species for which the genome sequence is not known (Wang et al., 2009). RNA-seq is a very sensitive method and is able to detect RNA with very low expression levels (Nagalakshmi et al., 2008)

Although the progressing sequence based technologies are providing an ever more comprehensive picture of the transcriptome, the effectors of biological functions are in most cases proteins (Cox and Mann, 2007). Proteomes are increasingly being explored using high throughput mass spectrometry technologies, which are well suited for characterizing complex protein mixture (Aebersold and Mann, 2003). Indeed, this advancing technology has already successfully been used for confirming and refining existing gene models, identifying novel genes and detecting AS isoforms that are expressed at the proteome level (Fermin et al., 2006; Tanner et al., 2007; Baerenfaller et al., 2008; Castellana et al., 2008; Mo et al., 2008; Tress et al., 2008; Blakeley et al., 2010; Chang et al., 2010).

In silico **analysis of alternative splicing**

The exon/intron structure of a pre-mRNA can be determined by aligning the mRNA sequence to the corresponding genomic sequence (Figure 4a). This task is performed using specialized programs that build spliced alignments (Gelfand et al., 1996) such as sim4 (Florea et al., 1998), GeneSeqer (Usuka et al., 2000) or GMAP (Wu and Watanabe, 2005). Premature mRNA molecules from a constitutively spliced gene are always spliced in the same way and are therefore represented by a single exon/intron structure. In contrast, different exon/intron structures are obtained for each unique mature transcript isoform from an alternatively spliced gene (Figure 4b).

The transcript diversity from alternatively spliced genes is often described in a pairwise fashion. It is common to choose a single transcript as the reference isoform against which all other isoforms (alternative) are compared. Different names and selection criteria have been used for the reference isoform in literature. For instance, in (Resch et al., 2004) the reference isoform is referred to as the “major transcript” and corresponds to the transcript isoform that is supported by the highest number of transcript evidence. Tress and co-workers (Tress et al., 2008) used several rules, mostly based on evolutionary conservation, for identifying the reference isoform which they called the "principle isoform". The differences between the reference and alternative isoforms are described in terms of AS events (Figure 5a). The "classical" nomenclature used for describing AS events is not well suited for naming "complex" events that can be combinations of classical events (Sammeth et al., 2008). Therefore different nomenclatures have been proposed which can capture both "classical" and "complex" events (e.g. Nagasaki et al., 2006; Sammeth et al., 2008).

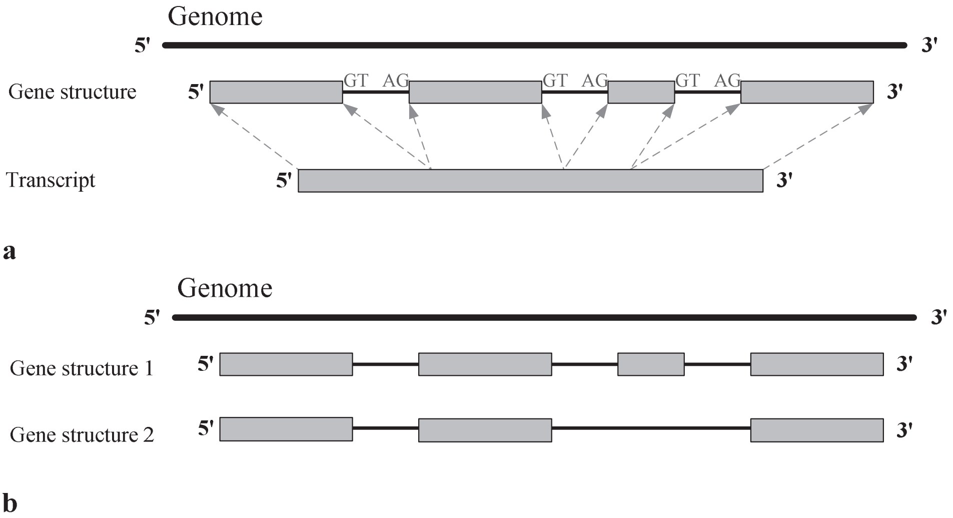

Figure 4. Gene-structure identification. (a) Searching of mRNA sequences against its corresponding genomic location results in the identification of exons (grey) that are separated by introns (horizontal lines). The 'correct' boundary of introns can be identified by for instance searching for the canonical donor (GT) and acceptor (AG) consensus di-nucleotides of introns. (**b**) Mapping of AS transcript isoforms against the same genomic locus results in distinct gene structures.

The full transcript isoform diversity produced by a gene can also be captured in an acyclic directed graph called the "splicing graph" (Heber et al., 2002) (Figure 5b and c). Each path through a splicing graph corresponds to a possible full-length transcript isoform (Lee and Wang, 2005). Different types of splicing graphs have been published and are either built from exon/introns mapped on the genome (e.g. Chacko and Ranganathan, 2009) or from transcript data directly (e.g. Heber et al., 2002). These types also differ in the information that can visually be extracted from them. For instance, splicing patterns can be captured in a graph in which the vertices correspond to introns and edges are “undefined” in the sense that they do not strictly correspond to an intron or an exon (Figure 5b; see also Sugnet et al., 2004). In a different type of splicing graph (see figure 5c), such as used in (Chacko and Ranganathan, 2009) nodes correspond to exons and vertices to the introns that connect them.

Usage of ESTs can result in the prediction of fragmented gene structures because they are often fragments of cDNA molecules. Splicing graphs are useful for reconstructing possible full-length transcripts from these fragmented gene structures. However, it is difficult to determine whether the reconstructed full-length transcripts were indeed present in the biological sample. In order to ensure that only a minimum number of possible full-length transcripts are missed, one can choose to generate all possible transcripts from a splicing graph. Alternatively and in order to prevent the prediction of false full-length transcripts, one can choose to use methods that only reconstruct a minimum set of full-length transcripts that is capable of explaining the observed experimental evidence (reviewed in Lee and Wang, 2005).

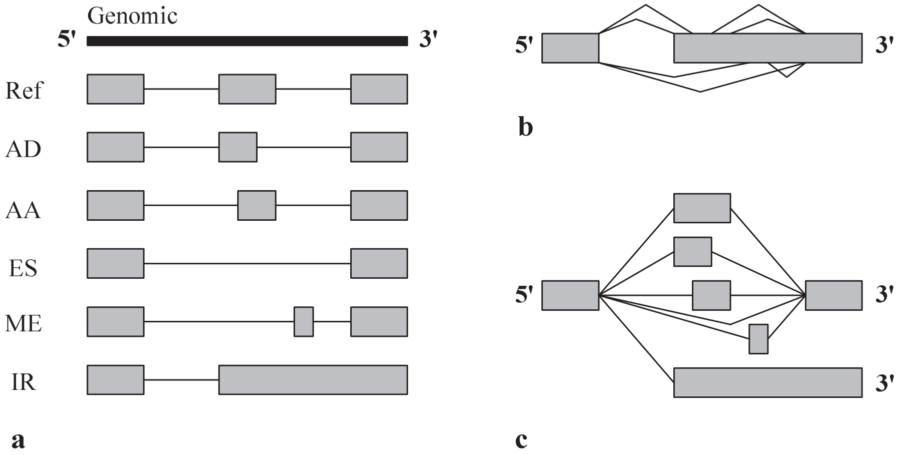

Figure 5. Alternative splicing diversity. (**a**) An example of the different exon (grey boxes) / intron (horizontal lines) structures belonging to pre-mRNA isoforms from an alternatively spliced gene. AS events can be classified by selecting one reference transcript (Ref) against which the remaining transcripts are compared. The following AS events are shown: Alternative Donor (AD), Alternative Acceptor (AA), Exon Skipping (ES), Mutually Exclusive exons (ME) and Intron Retention (IR). The full AS-transcript diversity can also be captured in so-called splicing graphs. (**b**) A splicing graph in which vertices correspond to pairs of splice junctions and edges are either exons or introns. (**c**) A splicing graph in which exons are nodes that are connected by vertices which represent identified introns . Each path from the 5’ to 3’ end corresponds to a potential full length transcript.

As mentioned earlier, microarrays are based on probe hybridization and therefore their design determines what they can detect. One common type of array, tiling arrays, often cover large regions including entire chromosomes and are not specifically focused on specific genomic features such as exons and introns (Figure 6a). These arrays have successfully been used for analyzing exon-skipping (Kampa et al., 2004) and intron retention events (Ner-Gaon and Fluhr, 2006). Other commonly used designs (Figure 6b and Figure 6c) are exon arrays (e.g. Clark et al., 2007) and junction arrays (e.g. Johnson et al., 2003) which, unlike tiling arrays, are based on *a priori* knowledge of gene structures or splice patterns. Microarrays can be custom made for answering specific questions as demonstrated for instance by Clark 2002 et al. (Clark et al., 2002) who used a combination of junction-, exon- and intron probes to compare the splicing patterns between different yeast cells. Specialized methods are required for the deconvolution of raw microarray data into for instance gene-expression differences and alternative splicing (reviewed Cuperlovic-Culf et al., 2006).

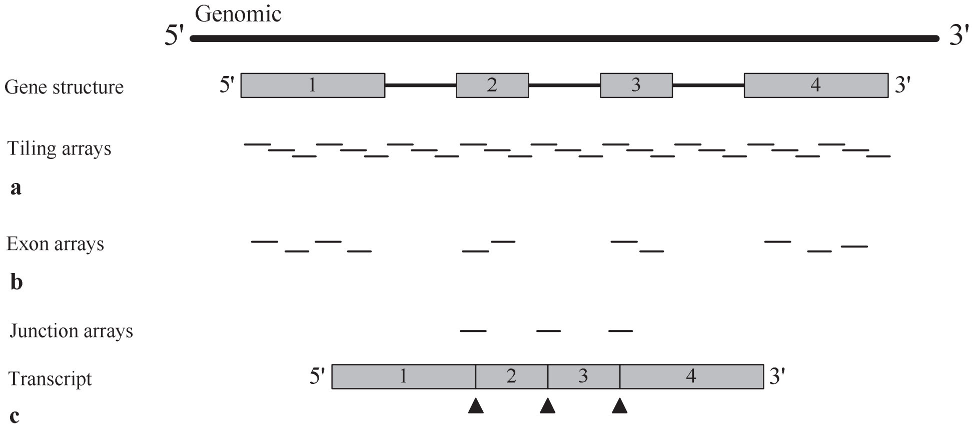

Figure 6. Microarray probe design. A schematic overview is given for microarray designs with probes that overlap with a known gene. (a) Tiling arrays span entire regions or complete chromosomes and both the exon and introns are represented by probes. (b) The probes on exon arrays are limited to predicted exonic regions of the gene. (c) Junction arrays contain probes that span exon/exon junctions as present on the transcript.

RNA-seq experiments yield large amounts of reads with lengths ranging from 20 to 30 and lately up to 75 bp (Haas and Zody, 2010). Reconstruction of full-length transcripts using this data can be done by first mapping the reads to the genome with software packages specialized for mapping short reads, such as e.g. TopHat (Trapnell et al., 2009). The full-length transcripts can afterwards be obtained using graph based approaches (e.g. Guttman et al., 2010). Alternatively, there are also software packages available which are capable to reconstruct transcript isoforms from the RNA-seq reads alone (e.g. Birol et al., 2009). An additional frequently used technique is paired-end tag sequencing which yields short read pairs that represent both ends of a DNA or RNA fragment (Fullwood et al., 2009). Paired-end data can be used to connect graphs from the same transcript that were initially not connected due to limited overlap (e.g. Guttman et al., 2010).

Not all AS transcripts are translated into proteins

Many AS events result in transcripts containing in frame premature termination codons (PTCs) and encode truncated proteins that can be toxic. Eukaryotes have a surveillance mechanism that can recognize these PTC-containing transcripts and initiate their decay through the Nonsense Mediated Decay (NMD) pathway (Lewis et al., 2003). It has been suggested that inefficient splicing resulting in large amounts of unspliced transcripts has been a driving force behind the evolution of the NMD (Roy and Irimia, 2008).

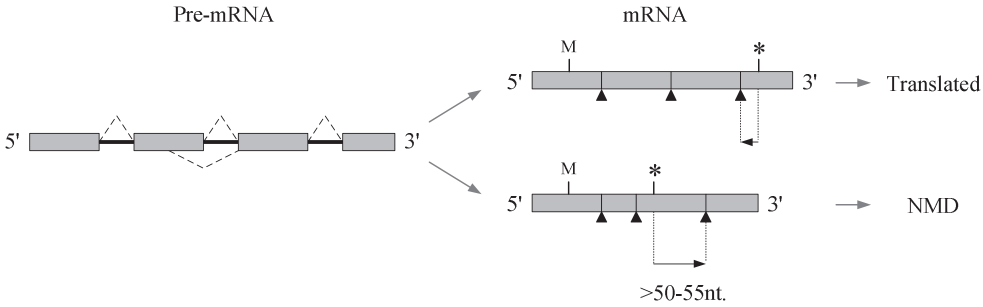

Figure 7. NMD candidature. An example of a pre-mRNA molecule that produces two AS isoforms that differ in their protein coding regions (from M to *). The termination codon (*) of the first mRNA (upper model) is located downstream of the last (most 3’) exon/exon boundary (black triangles). This mRNA molecule is predicted to serve as a template for the production of proteins. The termination codon of the second mRNA (lower model) is located more than 50-55 nt upstream of the last exon/exon boundary. As a result, the mRNA molecule is considered to be a candidate for degradation through the NMD pathway.

There are two versions of the same bioinformatics rule that is commonly used for automated detection of NMD targets. One version of the rule states that a transcript is a likely NMD target when it contains a PTC that is located more than 50-55 nt upstream of an exon/exon junction (e.g. Maquat, 2005; Reddy, 2007; McGlincy and Smith, 2008). In the second version of the rule the relevant distance is measured between the location of the PTC and the 3'- most exon/exon junction (e.g. Nagy and Maquat, 1998; Hori and Watanabe, 2007) (Figure 7).
Around ~35% of human AS events (Lewis et al., 2003) and at least ~36%-78% of plant AS events (Wang and Brendel, 2006; Filichkin et al., 2010; Zhang et al., 2010) are predicted to result in transcripts that are targets for the NMD pathway. The frequent occurrence of NMD-inducing AS events has lead to the suggestion that AS might be widely employed as a means for regulating gene expression at the post-transcriptional level. This type of gene expression regulation has been termed "Regulated Unproductive Splicing and Translation" (RUST) (Lewis et al., 2003; Lareau et al., 2004). In fact, a number of cases of NMD-mediated regulation of gene expression have been described in both plants and animal species (reviewed McGlincy and Smith, 2008). In contrast to humans where the widespread occurrence of RUST has been debated (Pan et al., 2006), it has been suggested that the process might be important in Arabidopsis (Filichkin et al., 2010)

To what extent is AS functional?

There is no doubt, based on the prevalence of AS in many organisms that the process substantially contributes to the transcriptome complexity of (higher) eukaryotes. However, the extent to which the AS-induced transcriptome complexity is indeed functional remains unresolved. The functional importance of the splicing process in humans is illustrated by the substantial fraction (~15%) of genetic-disease causing point mutations that result in defective splicing (Faustino and Cooper, 2003). However, splicing defects or "mis-splicing" can also contribute to phenotypic variation between individuals that is not related to disease. For instance, a naturally occurring mutation that affects "normal" splicing of an *Arabidopsis FLOWERING LOCUS C* homeolog in *Capsella bursa-pastoris* has been shown to be associated with an early flowering phenotype (Slotte et al., 2009). Focusing on the splicing process itself, the question becomes: what fraction of AS events is functionally relevant and does not merely reflect splicing noise (Melamud and Moult, 2009).
One of the first indications that a gene has an alternative splicing event that is functional involves evidence for tissue- or condition-specific production of the different splice variants. Indeed, microarray and NGS studies have uncovered many genes that show such a pattern of AS events (e.g. Clark et al., 2007; Pleiss et al., 2007; Pan et al., 2008; Wang et al., 2008; Filichkin et al., 2010; Zhang et al., 2010). However, this type of evidence does not provide proof that an AS event is functional. What matters is the molecular impact or

the role of adjusting the relative expression levels of the different transcript isoforms. For instances, it has been demonstrated that amino acid starvation in *Saccharomyces cerevisiae* results in a reduction of the splicing efficiency for many transcripts from genes encoding ribosomal proteins (Pleiss et al., 2007). This reduced splicing efficiency likely results in a reduced production of ribosomal proteins and an overall down-regulation of translation.

For only a few genes have the functions of multiple protein isoforms been determined experimentally (e.g. Reddy, 2007; Nilsen and Graveley). It is impractical or impossible to determine the function of multiple protein isoforms from alternatively spliced genes on a large scale using wet lab experiments. Therefore, various *in silico* methods such as protein structure analysis (Tress et al., 2007; Melamud and Moult, 2009) and sequence conservation have been used to determine the extent to which AS contributes to the functional diversity of the proteome. The main similarity between the different methods is the application of predefined criteria for identifying functional AS events at a genome-wide scale.

Scope of this thesis

In this thesis, the contribution of AS to the diversity of the transcriptome and proteome in plants is investigated using different criteria for function. In chapter 2, conservation across species was used as the main criterion for determining whether an AS event is functional. The focus lies on the conservation of AS events as mechanisms for producing multiple specific protein products from a single gene. In addition, it was analyzed whether conserved patterns can be found that link AS to specific gene- or predicted protein domain functions. In chapter 3, manifestation of an AS events at the proteome level was considered to be an indication for its functionality. The main objective in this chapter was to determine whether the observed contribution of AS to the proteome diversity corresponded to the predicted contribution of AS to the proteome diversity in Arabidopsis. In chapter 4 we investigated whether polymorphisms that are typically observed between different AS isoforms have contributed to evolution of protein diversity within the plant kingdom at large. In chapter 5, the focus of analysis was shifted from AS at the genome-wide level to AS at the gene-family level. The chapter presents a detailed study of AS in the MIKC-subgroup of the MADS-box transcription factor family in plants. Various criteria for function were used in order to analyze the potential impact of AS on the protein-protein interaction capabilities of individual MIKC proteins and the evolution of the network topology formed between these proteins.

Chapter 2

Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome

Edouard I Severing, Aalt DJ van Dijk, Willem J Stiekema
and
Roeland CHJ van Ham

Published in BMC Genomics (2009) 10: 154.

Abstract

Background
Alternative splicing (AS) is a widespread phenomenon in higher eukaryotes but the extent to which it leads to functional protein isoforms and to proteome expansion at large is still a matter of debate. In contrast to animal species, for which AS has been studied extensively at the protein and functional level, protein-centered studies of AS in plant species are scarce. Here we investigate the functional impact of AS in dicot and monocot plant species using a comparative approach.

Results
Detailed comparison of AS events in alternative spliced orthologs from the dicot *Arabidopsis thaliana* and the monocot *Oryza sativa* (rice) revealed that the vast majority of AS events in both species do not result from functional conservation. Transcript isoforms that are putative targets for the nonsense-mediated decay (NMD) pathway are as likely to contain conserved AS events as isoforms that are translated into proteins. Similar results were obtained when the same comparison was performed between the two more closely related monocot species rice and *Zea mays* (maize).

Genome-wide computational analysis of functional protein domains encoded in alternatively and constitutively spliced genes revealed that only the RNA recognition motif (RRM) is overrepresented in alternatively spliced genes in all species analyzed. In contrast, three domain types were overrepresented in constitutively spliced genes. AS events were found to be less frequent within than outside predicted protein domains and no domain type was found to be enriched with AS introns. Analysis of AS events that result in the removal of complete protein domains revealed that only a small number of domain types is spliced-out in all species analyzed. Finally, in a substantial fraction of cases where a domain is completely removed, this domain appeared to be a unit of a tandem repeat.

Conclusions
The results from the ortholog comparisons suggest that the ability of a gene to produce more than one functional protein through AS does not persist during evolution. Cross-species comparison of the results of the protein-domain oriented analyses indicates little correspondence between the analyzed species. Based on the premise that functional genetic features are most likely to be conserved during evolution, we conclude that AS has only a limited role in functional expansion of the proteome in plants.

Background

Eukarotyes can produce different mRNAs from a single gene transcript through the process of alternative splicing (AS). Large-scale EST sequencing efforts have revealed that AS is widespread among higher eukaryotes and that it greatly affects their transcriptome complexity, with, for instance, more than 60% of human genes and around 20-30% of plant genes undergoing AS (Campbell et al., 2006; Wang and Brendel, 2006; Kim et al., 2007). The observation that most AS events occur within the coding region of genes suggests that AS has an important role in the proliferation of an organism's proteome diversity (Modrek and Lee, 2002; Reddy, 2007). Next to proteome expansion, AS can also serve as a post-transcriptional mechanism for regulating gene expression by producing transcript isoforms that contain a premature termination codon (PTC) which can trigger the decay of the transcript through the nonsense-mediated mRNA decay (NMD) pathway (Lewis et al., 2003).

Although most AS events have the potential to produce different protein isoforms, the extent to which these isoforms are functional is currently unknown. This issue has been addressed using different *in silico* approaches. A number of studies, in which AS isoforms were structurally modeled, have yielded insight into the impact of AS on the stability and function of proteins (see e.g.Yura et al., 2006; Tress et al., 2007; Birzele et al., 2008).

Structural information, however, is available for only a small fraction of known proteins. An alternative approach has therefore been the comparative analysis of AS between two or more species. The basic assumption behind this type of analysis is that functional genetic features are more likely to be conserved across species than non-functional ones. Those features can be gene-specific, such as the intron/exon structure of a particular gene, or mechanistic and associated with particular cellular processes. In the context of conservation of AS features, the former involve searching orthologous genes for similar AS events that were likely present in the common ancestor of the species under study and that were retained due to an enduring selective advantage of the event. Mechanistic features of a process can often be identified within a single species but if they are also found in a second organism, then their functionality has likely been conserved during evolution.

At the gene-structural level, conserved AS events are identified as similar or distinct event types on orthologous introns and/or exons of at least two species (see for instance Iida and Go, 2006; Kalyna et al., 2006; Li et al., 2006). However, conserved AS events do not necessarily have similar effects on the encoded proteins while non-conserved AS events on the other hand, may have similar effects on the encoded protein. If the functionalities of more than one isoform are simultaneously favored, then it is the polymorphic character of the gene that provides the selective advantage. Conservation of such polymorphic characters can be identified by searching for homologous alternatively spliced genes in two species for which protein isoforms are modulated in similar ways through AS (see for instance Valenzuela et al., 2004).

A number of general mechanistic features of AS have been identified in diverse species over the past few years. For instance, not only are the fractions of all genes that undergo AS similar in *Arabidopsis thaliana* and rice but also the frequencies of the individual AS events are similar in both species (Campbell et al., 2006; Wang and Brendel, 2006). Another conserved mechanistic feature in dicot plant species is the inverse correlation between the UA content of introns and the incidence of AS in these introns (Baek et al., 2008). The presence of UA-rich sequences seems to enhance splicing in rice, indicating that such sequences in introns are important for the splicing process (Baek et al., 2008). A recent comparative study in which alternatively spliced genes of a wide range of eukaryotes were compared has revealed that genes encoding proteins with particular activities, for instance DNA-, RNA- and calcium-binding, had elevated levels of AS (Irimia et al., 2007). In line with this study, various protein domains have been shown to be disproportionately distributed towards alternatively spliced genes in humans (Liu and Altman, 2003).

Other biases involving the mechanisms by which AS modulates the function of proteins have also been identified in animal species. For instance, AS within the protein coding region of animal genes has been shown to occur more frequently outside than inside the boundaries of predicted functional domains (Kriventseva et al., 2003). The authors also reported that evolutionary selection tends to favor AS events that remove entire functional domains. In fact, it has been shown that tissue-specific AS events in mouse transcription factor-encoding genes frequently involve the removal/addition of entire DNA-binding domains (Taneri et al., 2004). Finally, genome-wide functional profiling of spliced-out domains has revealed that a number of domain types, e.g. protein-binding domains, are more frequently spliced-out by AS than expected by chance (Resch et al., 2004). Although it has been suggested that AS in plants has a less prominent role in the expansion of proteome diversity than in animals (Kim et al., 2008), protein-centered studies that focus on the functional impact of AS in plants are very scarce. In this study, we evaluate the hypothesis that AS in plants contributes to expansion of the functional proteome by analyzing the species *Arabidopsis thaliana*, *Oryza sativa* (rice), *Zea mays* (maize) and *Populus thrichocarpa*. To investigate whether the production of different functional isoforms from a single gene through AS persists during evolution, we analyzed the extent to which polymorphism induced by AS is conserved in orthologous genes. The comparative analysis was performed between the monocot rice and dicot *Arabidopsis* and between the two monocots rice and maize. Following this ortholog-based comparison, we analyzed conservation of more general mechanistic aspects of AS. These analyses were performed separately on each species and the results were compared afterwards. Finally, we address the role of AS in proteome diversification and discuss the evolutionary events that might have led to the patterns that we observed in the plant species.

Results

Gene structure and AS predictions

An overview of the gene structure and AS prediction results prior to the construction of hypothetical isoforms is given in Table 1. A surprisingly large number of putative loci were predicted in the maize genome by the methods used here. Unlike the genomic data for *Arabidopsis*, rice and *Populus*, for which a significant part of the genome is represented by few large contigs corresponding to the chromosomes, the genomic data for maize was represented by over 16,000 BAC contigs. For this reason, additional steps were taken for processing the raw maize predictions (see Material and Methods). From all the initially predicted maize loci, only 9360 met our criteria and were considered further.

AS was detected in 24% and 28% of the tested loci from *Arabidopsis* and rice, respectively. The frequencies of the individual AS event types in these two species are comparable to those that have been reported in a previous genome-wide study (Wang and Brendel, 2006). Compared to *Arabidopsis* and rice, the fraction of maize genes that are predicted to undergo AS is relatively high (~40%). However, the distribution of the individual AS events within this species is similar to those found in *Arabidopsis* and rice (Table 1). The lowest number of genes and also the lowest frequency of AS (~12%) were found in poplar (Table 1). These results are likely to be the consequence of the relatively low amount of publicly available transcript data. For this reason, *Populus* was excluded from further analyses.

Table 1: Gene structure and AS predictions

	Arabidopsis		rice		poplar		maize	
Gene predictions								
Gene structures	36,373		60,431		17,573		173,277	
Loci	25,383		39,252		15,676		140,903	
Anchored								
Gene structures	19,949		18,398		4,686		11,147	
Loci	17,754		17,000		4,501		9,360	
Loci with AS	4,338	24.4%	4,703	27.7%	542	12.0%	3,724	39.8%
AS event type *								
AA	1,736	23.0%	1,577	16.5%	176	23.3%	1,589	19.1%
AD	932	12.4%	1,017	10.6%	92	12.2%	1,089	13.1%
AP	209	2.8%	590	6.2%	32	4.2%	391	4.7%
IR-CI	3,918	51.9%	4,972	51.9%	302	40.1%	4,072	48.9%
ES-CE	751	10.0%	1,428	14.9%	152	20.2%	1,178	14.2%
Total events	7,546		9,584		754		8,319	

*AA: Alternative acceptor, AD: Alternative donor, AP: Alternative position, IR: Intron retention, CI: Cryptic introns, ES: Exon skipping, CE: Cryptic exons. The frequencies of the individual AS events are derived from a non-redundant set of AS relationships in which the reciprocal event types CI and IR, and ES and CE are grouped into the IR-CI and ES-CE classes, respectively.

After constructing hypothetical isoforms for the remaining three species, final sets of gene structures were obtained and comprised 36,950, 40,543 and 25,064 isoform models for *Arabidopsis*, rice and maize, respectively. BlastX searches with the virtual transcripts of the isoforms within this set retrieved an additional 467 *Arabidopsis*, 1,219 rice and 38 maize annotated proteins.

Alternative splicing in orthologous genes

Out of 10,859 orthologous groups predicted by Inparanoid between *Arabidopsis* and rice, 6,131 main-orthologs had their CDS fully supported by EST evidence. AS was detected in both species for 713 of these ortholog pairs (39.2% of alternatively spliced *Arabidopsis* loci within the ortholog set). Of the 15,951 initial orthologous groups that were predicted between maize and rice, 3,053 main ortholog pairs were kept. From these, 481 pairs (~53% of alternatively spliced rice-genes within this ortholog set) were predicted to undergo AS in both species.

Two sets of alternatively spliced ortholog pairs were constructed: (*i*) ortholog pairs with AS variants that were predicted to be likely targets for the NMD-pathway in both species (NMD-set); and (*ii*) ortholog pairs with AS variants that were predicted to be translated into proteins in both species (translated-set). These two sets were stepwise dissected into sets containing ortholog pairs with (Figure 1): (*i*) AS events on different positions; (*ii*) AS events on the same position; (*iii*) different type of events on the same position; (*iv*) the same type of event on the same position; (*v*) homologous modification sites; and (*vi*) similar modifications (see Material and Methods).

The number of cases with the same type of AS event in orthologous introns between *Arabidopsis* and rice was low in both the translated and the NMD sets (Figure 1a), in agreement with results from previous studies (Wang and Brendel, 2006; Baek et al., 2008). A total of 16 ortholog pairs (~2.2%) from the translated set, including a pair of previously studied MYB transcription factors (Li et al., 2006), were found to have similar modifications to the principle protein isoform as the result of AS. In contrast, only two such cases were found in the NMD set. The fraction of ortholog pairs with AS in the CDS region that had AS events on different positions were roughly the same in both the NMD- and translated sets. The same was true for the fraction of ortholog pairs with the same and different type of events on orthologous introns.

In total 24 (~4.9%) and 17 (~3.5%) maize-rice ortholog pairs had similar modifications in the translated and NMD set, respectively (Figure 1B). Although these numbers are higher than those observed for the *Arabidopsis*-rice orthologs, they are still low. Within the translated set of the maize-rice orthologs, the distributions of AS events on different positions and same- and different type of events on orthologous intron positions were not too different from the same distributions observed in both the NMD- and translated set of the *Arabidopsis*-rice orthologs. Within the NMD-set of the maize-rice orthologs, the number of different AS event-types on orthologous introns was roughly the same as the number of similar AS events on orthologous introns. In all subsets, similar AS events on orthologous positions were the greatest contributors to the subset of similar modifications. The GO terms corresponding to the ortholog pairs within the different subsets were analyzed but no clear overrepresentation of any particular term could be identified (data not shown).

Disproportionately distributed protein domains

We investigated whether particular protein domains were disproportionately distributed towards either constitutively (CS) or alternatively (AS) spliced genes in all three plants. Disproportionately distributed domains are indicative for the under- or overrepresentation of AS in genes with a particular function. To this end, the distributions of 2,626 domain types in *Arabidopsis*, 2,487 in rice and 1,791 in maize were analyzed.

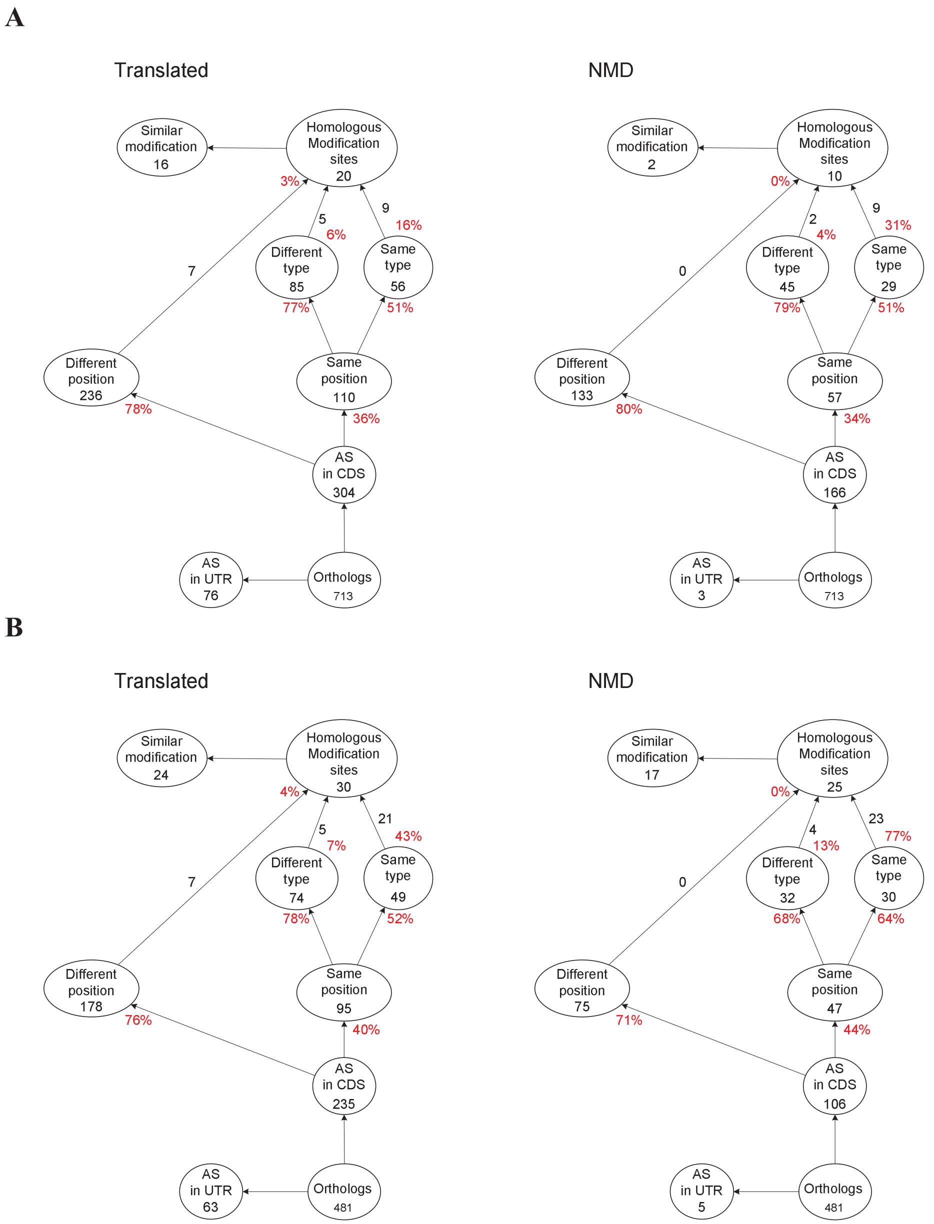

Figure 1 - Dissection of AS events in orthologous genes. A schematic view is provided of the dissection of AS events in Arabidopsis-rice orthologs (A) and maize-rice orthologos (B).The dissection is given for both the translated- and NMD set. Number of orthologous genes within each category is

indicated below the name of the category. Numbers next to arrows represent the number of pairs within the category from which the arrow originates that contribute to the number of pairs within the category pointed to by the arrow. For a number of classes the fraction of ortholog pairs from its preceding class that is contained within the child class is provided (in red). Note: several ortholog pairs can be represented in more than one child class as a number of genes have more than one AS variant.

Table 2: Disproportionately distributed domains

Species	Pfam Id	Domain name	CS genes w. domain	AS genes w. domain	p-value	Bias
Arabidopsis						
	PF00560	LRR_1	153	14	3.55E-08	CS
	PF08263	LRRNT_2	123	9	5.65E-08	CS
	PF02519	Auxin_inducible	55	0	1.44E-07	CS
	PF01535	PPR	146	15	4.03E-07	CS*
	PF00076	RRM_1	114	77	1.61E-06	AS*
	PF00201	UDPGT	69	3	2.73E-06	CS*
	PF00067	p450	143	18	1.03E-05	CS*
	PF00847	AP2	115	15	0.0001159	CS
rice						
	PF01535	PPR	179	17	1.53E-11	CS*
	PF00076	RRM_1	99	89	1.04E-08	AS*
	PF00646	F-box	202	34	4.69E-07	CS
	PF00067	p450	137	19	1.52E-06	CS*
	PF00201	UDPGT	70	5	3.59E-06	CS*
	PF07859	Abhydrolase_3	28	0	0.000101	CS
maize						
	PF01609	Transposase_11	52	0	9.62E-13	CS
	PF01535	PPR	64	3	5.70E-12	CS*
	PF00201	UDPGT	46	7	9.53E-06	CS*
	PF00067	p450	46	10	0.0001739	CS*
	PF00076	RRM_1	47	65	0.0002	AS*

[*] Domains that have the same distribution bias in all three species.

A domain was considered to be disproportionately distributed when the ratio AS- to CS-genes encoding the domain was significantly different from the same ratio for all genes that lack that particular domain. The fraction of domains that are disproportionately distributed in the studied species (Table 2) ranges from 0.2% to 0.3%. Four domain types were found to be disproportionately distributed in all three species. Three of these, the pentatricopeptide (PF01535), the UPD-gluccuronosyl/UDP-glucotransferase (PF00201) and the Cytochrome P450 (PF00067) domains were overrepresented in CS genes. The fourth domain, the RRM_1 domain (PF00076), which is an RNA recognition motif, was the only domain that was significantly overrepresented in AS genes of all the species studied.

Intron associated with protein domains

To determine whether introns located within predicted Pfam domains are less or more frequently subjected to AS compared to other introns within the CDS regions of genes, the ratio between AS intron clusters and CS intron clusters within the predicted domains was compared to the same ratio for intron clusters located outside predicted domains. In *Arabidopsis* 2,486 (~6%) of the 41,107 intron clusters located within the boundaries of predicted Pfam domains and 1,941 (~7.6%) of the 25,657 clusters outside Pfam domains were found to be AS clusters. Similarly, in rice, 2,487 (~7%) of the 35,595 intron clusters within domains and 2,088 (~9%) of the 23,093 clusters outside predicted domains were AS clusters. The same trend was also observed in maize where 2,292 (~14%) of the 16,466 intron-clusters within predicted domains were AS-clusters, while 1,634 (~19%) of the 8,677 intron clusters outside predicted domains were AS clusters. The difference between the AS-cluster frequencies within and outside predicted domains is significant in all three species (p-value << 0.01).

Next, we investigated whether particular protein-domain types have retained a higher propensity for AS during evolution by having particular sequence or structural features. To this end, the ratio between AS and CS intron clusters located within a particular domain type was compared to the same ratio between intron clusters located within all other domains types. A total of 304, 309 and 316 domain types in *Arabidopsis*, rice and maize, respectively, were excluded from the test as no introns were found within their boundaries. The number of domain types that were enriched with either AS or CS introns ranged from six in maize to twelve in *Arabidopsis* (Table 3). The Kinesin domain (PF00225) in *Arabidopsis* was the only domain that was significantly enriched with constitutively spliced introns; no domain types enriched with CS intron clusters were detected in rice and maize. Not a single domain type was found that was enriched for AS intron clusters in all species.

Spliced-out protein domains

The function of a protein can be modulated by AS through the addition or deletion of complete functional domains but also of small stretches of amino acids. However, because it is impossible to assess the impact of small amino acid sequence changes on protein function without additional structural or experimental data, only AS events were considered that removed whole protein domains while leaving at least one domain intact. A total of 936, 976 and 1,211 alternatively spliced loci with at least one isoform encoding a multi-domain protein were analyzed in *Arabidopsis*, rice and maize, respectively.

Table 3: Intron/Domain associations

Species	Pfam Id	Domain name	CS Introns	AS introns	p-value	Bias
Arabidopsis						
	PF00303	Thymidylat_synt	0	14	2.20E-16	AS
	PF00724	Oxidored_FMN	8	9	1.76E-07	AS
	PF06487	SAP18	0	5	8.32E-07	AS
	PF04628	Sedlin_N	2	6	1.27E-06	AS
	PF00697	PRAI	1	5	4.74E-06	AS
	PF00225	Kinesin	228	1	8.68E-06	CS
	PF07847	DUF1637	12	8	1.21E-05	AS
	PF03226	Yippee	13	8	1.85E-05	AS
	PF00582	Usp	55	14	5.99E-05	AS
	PF01716	MSP	1	4	6.52E-05	AS
	PF01789	PsbP	12	7	8.07E-05	AS
	PF07172	GRP	4	5	8.53E-05	AS
rice						
	PF00234	Tryp_alpha_amyl	0	11	2.33E-13	AS
	PF00504	Chloroa_b-bind	14	12	6.19E-08	AS
	PF00230	MIP	32	15	4.94E-07	AS
	PF02913	FAD-oxidase_C	1	6	8.50E-07	AS
	PF04172	LrgB	0	5	1.82E-06	AS
	PF05564	Auxin_repressed	2	5	3.38E-05	AS
	PF00011	HSP20	3	5	8.49E-05	AS
maize						
	PF01559	Zein	0	7	1.07E-06	AS
	PF01040	UbiA	18	14	4.36E-05	AS
	PF01596	Methyltransf_3	0	5	5.45E-05	AS
	PF03763	Remorin_C	7	9	9.41E-05	AS
	PF00297	Ribosomal_L3	5	8	9.96E-05	AS
	PF00722	Glyco_hydro_16	10	10	0.0001406	AS

An overview is given for all domains in *Arabidopsis*, rice and maize with a significant excess of either constitutive (CS) or alternative (AS) intron clusters.

Table 4: Number of shared spliced-out domains

Pfam Id	Domain name	*A.thaliana*	rice	maize
PF00400	WD40	15	10	7
PF00023	Ank	9	5	-
PF00076	RRM_1	10	11	4
PF02798	GST_N	3	3	3
PF00153	Mito_carr	3	6	3
PF01357	Pollen_allerg_1	-	3	3
PF03765	CRAL_TRIO_N	2	3	2
PF00270	DEAD	2	3	-
PF00226	DnaJ	3	-	2
PF00043	GST_C	3	2	3
PF07646	Kelch_2	2	-	3
PF00560	LRR_1	6	9	2
PF00072	Response_reg	2	2	-
PF00627	UBA	-	2	2
PF08240	ADH_N	1	3	4
PF00892	DUF6	6	1	3
PF08263	LRRNT_2	3	7	1
PF00249	Myb_DNA-binding	10	1	5
PF00614	PLDc	2	1	2
PF01535	PPR	7	3	1
PF00085	Thioredoxin	1	2	2
PF00035	dsrm	2	1	2
PF00036	efhand	11	10	1
PF00643	zf-B_box	2	2	1
PF00096	zf-C2H2	1	2	4
PF00642	zf-CCCH	1	6	4
PF00098	zf-CCHC	6	1	6
PF00641	zf-RanBP	2	1	2

In total, 218 (117 types), 244 (131 types) and 148 (88 types) spliced-out domains were identified in *Arabidopsis*, rice and maize, respectively. Twenty-nine domain types were found to be spliced out in all three species, 28 of which (Table 4) were located on at least two different genomic regions in at least two of the three species. We tested whether particular domains are spliced-out at significantly elevated rates within the subset of isoforms that were analyzed for spliced out domains. Only the PF00098 (zf-CCHC) domain was spliced out at significantly elevated rates in both maize and *Arabidopsis* compared to all other spliced out domains (Fisher exact test p-values: 8.4e-05 in *Arabidopsis* and 0.001 in maize). No domain type was found that was spliced out with significantly elevated rates compared to other domains in all three species.

Table 5: Number of splicing patterns of domain types spliced-out from a tandem repeat in all three species

Pfam Id	Domain name	*A.thaliana*	rice	maize
PF00076	RRM_1	4	5	4
PF00400	WD40	9	6	4
PF00153	Mito_carr	2	3	2
PF00035	dsrm	2	1	1
PF00036	efhand	5	5	1
PF00096	zf-C2H2	1	1	1
PF00249	Myb_DNA-binding	2	1	2
PF00412	LIM	1	1	2
PF00614	PLDc	1	1	1
PF00641	zf-RanBP	1	1	2
PF00642	zf-CCCH	1	2	2
PF00643	zf-B_box	1	1	1
PF00892	DUF6	2	1	1
PF01535	PPR	5	1	1
PF06943	zf-LSD1	1	1	1

The top five assignments of molecular function-related GO terms (Ashburner et al., 2000), accounting for around half of all assignments to the spliced-out domains in *Arabidopsis* and rice, were: nucleic acid binding (GO:0003676), zinc ion-binding (GO:0008270), calcium ion-binding (GO:0005509), protein-binding (GO:0005515) and ATP binding (GO:0005524). In maize, around 45% of GO term assignments were nucleic acid binding and zinc ion-binding

Next, the splicing patterns of the spliced-out domains and their resulting domain architectures (products) were analyzed and compared between the species. A total of 152 unique splicing patterns resulting in 129 unique products were identified in *Arabidopsis.* In rice these numbers were 174 and 151, respectively. In maize, a total of 103 patterns leading to 92 products were identified. Cross-species comparison resulted in the identification of 14 splicing patterns and 20 products that were shared by all three species.

Closer inspection of the splicing patterns revealed that 65 (~42.8%) of the *Arabidopsis*, 66 (~37.9%) of the rice and 43 (41.7%) of the maize splicing patterns involved the removal of a domain which was a unit of a tandem-repeat. Fifteen domain types were identified that were at least once removed from a tandem repeat in all three species (Table 5). Eight out of the 14 (57%) splicing patterns that were shared between the three species involved the removal of a domain from a tandem repeat.

Discussion

In this study we addressed the functional impact of alternative splicing in plants using a comparative analysis approach. According to the ascribed role of AS as a mechanism for expanding proteome diversity, a gene can become polymorph in its expression through an additional splice variant that encodes a different, yet functional protein. When the principle and alternative protein isoforms each confer a selective advantage, both are likely to be retained during evolution. This can be achieved through either retention of the AS-induced polymorphism or through gene-duplication followed by sub-functionalization of the duplicates (Su et al., 2006). Not only different protein isoforms but also AS-mediated regulation of gene expression through the production of PTC containing isoforms that are destined to be degraded by the NMD-pathway can provide a selective advantage. The exact AS event is not necessarily required to remain the same for both the production of protein isoforms and the production of PTC containing transcripts isoforms.

In this study, we investigated to what extent AS induced polymorphism as well as AS-mediated gene-regulation are conserved between orthologous genes of the dicot *Arabidopsis* and the monocot rice and orthologous genes of the two monocots maize and rice. The comparison was focused on the functional outcome of the AS events rather than on the AS events themselves. The number of cases in which orthologs in *Arabidopsis* and rice contained AS variants that are likely targets for NMD was roughly half that of ortholog pairs with variants that can be translated into proteins. Only a very small number of cases were found in which AS events in both orthologs resulted in similar modifications to the principle protein product. Interestingly, the fractions of ortholog pairs that have AS events on different positions and those that have similar or different event types on the same position were quite similar in both the NMD and translated subsets of the *Arabidopsis*-rice orthologs. Because the isoforms in the NMD subset would never function as protein, this similarity suggests that a considerable fraction of the putatively conserved AS events are not the result of functional conservation. Similar values for the distributions mentioned above were observed in the translated subset of the maize-rice orthologs. This similarity further strengthens the notion that conservation of AS events is not the result of the preservation of functional AS induced protein polymorphism. Although the fraction of orthologs that have similar AS-event types on orthologous intron-positions leading to modifications at similar sites is higher in the maize-rice comparison than that observed for the *Arabidopsis*-rice orthologs in the translated as well as the NMD set, these numbers are still remarkably low. In the NMD subset of the maize-rice orthologs the fractions of different- and similar event types on orthologous intron-positions are very similar. In summary, the results suggest that conservation of AS events as the result of the selective advantage of retaining the ability to produce multiple protein isoforms through AS is not frequent, even at short evolutionary distances.

Previous studies on animal data (see introduction) have revealed similarities between genes undergoing AS in various species. The observed patterns are not the result of ortholog-level conservation, which is evident from the low numbers of conserved AS events between, for instance, human and mouse (Nurtdinov et al., 2003). However, the patterns do point to the existence of mechanistic features that are associated to- or reflected in the AS process. One such mechanistic feature in animals is the high frequency of exon skipping events which is the consequence of the exon definition mechanism through which splice-sites are recognized (Berget, 1995). In contrast, the high frequency of intron retention events in plants is the consequence of the intron definition mechanism through which many plant introns are recognized (Barbazuk et al., 2008).

In contrast to animals, protein function-oriented studies of genome-wide AS patterns in plants are scarce. Motivated by the results obtained from computational analysis of the impact of AS on protein function in animals, we have applied a number of those analyses on plant data. The first analysis involved the distribution of Pfam domains over constitutively and alternatively spliced genes using a similar method as previously performed on human data (Liu and Altman, 2003). Three domains were significantly underrepresented in alternatively spliced genes in all three species. Genes encoding these domains share at least two properties that have previously been shown to be correlated with AS frequencies. First, the genes belong to highly expanded gene families in plants (Paquette et al., 2000; Paquette et al., 2003; O'Toole et al., 2008). Recent studies in animal species have shown that the incidence of AS is inversely correlated to gene family size (Kopelman et al., 2005; Su et al., 2006; Hughes and Friedman, 2007). However, such an inverse correlation has not been found in *Arabidopsis* and rice (Lin et al., 2008). These conflicting results may be the consequence of the different methods that have been used for delineating gene families. Second, genes encoding these domains have relatively low average numbers of introns in all three species (data not shown). It has previously been shown that the incidence of AS is positively correlated with the number of introns (Irimia et al., 2007).

Only one domain, the RRM domain, was significantly overrepresented in alternatively spliced genes in all species. This RNA recognition domain, like the domains mentioned above, is found in a large number of genes, many of which are involved in the regulation of the splicing process (Schindler et al., 2008). However, in contrast to the three domains that were underrepresented, the average number of introns harbored by genes encoding the

RRM domain is higher than the average number of introns in all genes in all three species and this might in part explain the observed data (data not shown).

Zooming in on the sequence level, it has been shown in four animal species that AS occurs less frequently within than outside the boundaries of predicted protein domains (Kriventseva et al., 2003). We observed a similar trend in all three plant species. This result, together with those presented by Kriventseva and co-workers (Kriventseva et al., 2003), suggests that in general AS events within the boundaries of protein domains are less favored in evolution.

It has been reported that domains with particular functions are more frequently targeted by AS than expected by chance (reviewed in Artamonova and Gelfand, 2007). Not only the location of introns but also the location of AS events within a gene are to a certain degree related to sequence- and structural features of the encoded protein (Wang et al., 2005; De Kee et al., 2007) and it was hypothesized that particular domains are more prone to undergo AS within their boundaries as the result of such features. This possibility was tested by searching for domains that were significantly enriched with alternatively spliced rather than with constitutively spliced introns. Although a few domains were identified that were significantly enriched with alternatively spliced introns in individual species, none of such domains were found across all species. This result suggests that none of the studied domains has a detectable evolutionary propensity for undergoing AS within its boundaries.

The final analysis involved AS events that remove entire protein domains. The specific AS events that lead to the removal of the domains are not required to be located within the boundaries of the spliced-out domain. In human genes, various domains have been shown to be more often spliced-out than average (Resch et al., 2004). Here, we identified a number of cases in which AS events resulted in the removal of complete protein domains in all three plant species, but only one domain was found to be spliced out at a significantly elevated rate in both *Arabidopsis* and maize. The top five molecular function-related GO terms of the spliced-out domains, comprising around 50% of all GO-term assignments, were the same in *Arabidopsis* and rice while in maize only two of these five GO-terms comprised 45% of GO term assignments. Genes having these GO terms in *Arabidopsis* have been shown to have elevated levels of AS (Irimia et al., 2008). However, in that study only the AS levels of genes with nucleotide binding activities were found to be significantly elevated.

Interestingly, in a large fraction of the unique splicing patterns in all species, the spliced-out domain was a unit of a tandem repeat. A similar observation has been reported in a study that was conducted on data derived from the Swiss-Prot and PDB- databases (Birzele et al., 2008). The authors suggested that AS within repeated regions of a protein is more likely to be tolerated. Their conclusion is supported by the observation that variation in repetitive regions has frequently been used in evolution for modifying the properties of proteins without invoking loss of their fold (Andrade et al., 2001). It is also possible that the large fraction of domains that are spliced-out from tandem repeats is the consequence of the duplication events themselves as is seen for the frequent appearance of AS through tandem exon duplications (Kondrashov and Koonin, 2001). However, further research is needed to clarify the observed frequent removal of domains from a tandem repeat.

Conclusions

Although a considerable number of functional AS events in plants have been described in literature (Reddy, 2007), we have shown here that only a small number of AS events have conserved effects at the protein sequence level in *Arabidopsis* and rice. The same has been shown for ortholog genes of the more closely related species maize and rice. The persistence of gene-specific function expansion has thus been very limited over the divergence time between not only the monocot rice and the dicot *Arabidopsis* but also between the two monocots rice and maize. Functional diversification through AS within protein domains appeared to be independent on the nature of the domain itself. The identification of domains that are frequently spliced out in both species is likely to be the result of their abundance rather than their specific function. Taken together, these results, in the context of comparative analysis, suggest that the role of AS as a mechanism for expansion of function proteome diversity in plants is very limited. It has previously been suggested that the majority of AS variants are unlikely to be functional (Tress et al., 2007; McGuire et al., 2008). However, as long as such non-functional isoforms are selectively neutral, they can still serve as templates for further diversification and eventually lead to the acquisition of new functions. Their propagation might be enhanced by the NMD-pathway which protects them from being selected against (McGlincy and Smith, 2008). In conclusion, AS seems to be more a mechanism that leads to functional diversification over evolutionary time, comparable with e.g. gene duplication, than a mechanism for providing an organism during its life span with more functional proteins than the number of genes in its genome.

Material and Methods

Data sources and gene structure predictions

Over 1,000,000 transcript sequences for *Arabidopsis*, rice and maize and around 90,000 transcript sequences for poplar were downloaded from the PlantGDB database (Dong et al., 2004) and inspected using the SeqClean package downloaded from (DFCI Gene Indices Software Tools from http://compbio.dfci.harvard.edu/tgi/software/.). The (pseudo-)chromosome sequences and annotation data versions TIGR 5.0 and TAIR 7.0 were used for rice and *Arabidopsis*, respectively. For maize, BAC contigs and annotation data were download from the maize sequence ftp site . The chromosome contigs and annotation data for *Populus* version 1.1 were downloaded from the JGI poplar download-site . Annotations sets of maize and *Populus* were expanded by adding proteins from these organisms that were extracted from NCBI Nr database .The most likely genomic origin of each transcript was determined by performing blastN searches (Altschul et al., 1997) (word size=32; e-value threshold=1e^{-10}). Only those transcripts for which ≥90% of their bases could be matched to the genome with a mean identity ≥97% were considered further. Clusters of transcripts that mapped to overlapping genomic regions were realigned to the genome using the program GeneSeqer (Brendel et al., 2004) which has specific splice-site models for each species. Only exons that were predicted with an alignment score ≥0.98 were considered further. As the maize genome, unlike the genomes of the three other species, is represented by more than 16,000 BAC contigs, only one copy was considered from the initially predicted loci that had exactly the same set of gene structures. Gene structures encoding annotated proteins were identified through blastX searches with their virtual transcripts against the corresponding predicted proteome of the species. Only those loci containing at least one gene structure with an exact match to an annotated protein were considered further.

Prediction of AS and construction of isoforms

The majority of the available transcript data for *Arabidopsis* and rice is represented by ESTs. To overcome the shortage of full-length cDNA sequences, additional hypothetical full-length isoforms were constructed using a previously described method (Kan et al., 2002) with the following modifications. In our method, only those isoforms that fully encoded an annotated protein were used as reference isoforms. AS events were identified by comparing these reference isoforms to all other isoforms (second isoform) from the same locus for which either of the following conditions were met: (*i*) the two isoforms shared at least one intron or exon; or (*ii*) fifty percent of the exonic nucleotides were shared between both isoforms. The following types of AS events were classified (using terms different from those in Kan et al., 2002): alternative donor (AD), alternative acceptor (AA) and alternative position (AP) events were identified as overlapping introns differing in their donor-, acceptor- or both splice sites, respectively (Figure 2A). Intron retention (IR) events were identified as introns in the reference isoform that were fully contained within an exon of the second isoform (Figure 2B) and the reciprocal event, cryptic intron (CI), involved an intron in the second isoform that was fully contained within an exon of the reference isoform. Exon skipping (ES) events were identified as internal exons in the reference isoform that were fully contained within an intron in the second isoform (Additional file 1, C) and the reciprocal event, cryptic exons (CE), involved exons in the second isoform that were fully located within an intron of the reference isoform. The outermost 3'- and 5'- genomic boundaries of a hypothetical isoform only differed from those of its reference isoform when the first or last exon of either isoforms was fully located within the first or last intron of the other isoform.

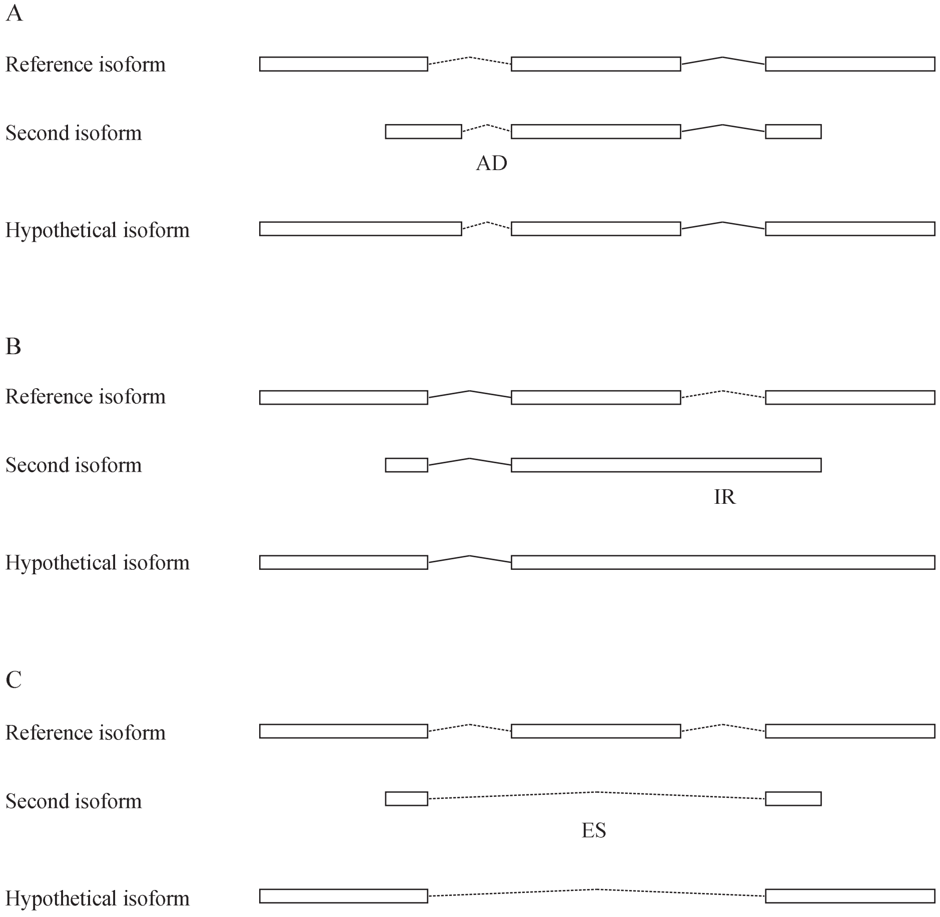

Figure 2. Detection of AS events and construction of full-length isoforms. A. Alternative donor (AD) events are identified as two overlapping introns (dashed), one on each isoform, that differ in their donor sites. In the hypothetical isoform, the intron on the reference isoform is substituted by the overlapping intron on the second isoform. **B**. Intron retention (IR) events are characterized as an intron on the reference isoform (dashed) which is fully contained within an exon on the second isoform. The hypothetical isoform, when compared to the reference isoform lacks the retained intron. **C**. Exon skipping (ES) events are identified as two consecutive introns (dashed) on the reference isoform that overlap with a single intron on the second isoform (dashed). In the hypothetical isoform the consecutive introns of the reference isoform are substituted by a single intron of the second isoform.

Two open reading frames (ORFs) were assigned to those isoforms that did not fully match to an annotated protein: (*i*) the longest ORF from start to stop that overlaps with the ORF of the reference isoforms and; (*ii*) if possible the ORF starting at the same genomic position as the ORF of the reference isoform. All isoforms containing a termination codon more than 55 nt upstream of the last exon/exon junction were considered as putative targets for the NMD-pathway (Hori and Watanabe, 2007). The impact of AS on the primary protein sequence encoded by the reference isoform was analyzed by comparing the ORFs of each reference/hypothetical isoform pair at the gene level.

Comparison of AS in orthologous genes

Orthologs were identified using the program Inparanoid version 1.35 (Remm et al., 2001). Needleman-Wunsch alignments (Needleman and Wunsch, 1970) were constructed for all possible ortholog pairs using the program Needle from the EMBOSS package version 4.0 (Rice et al., 2000). In a number of orthologous groups in which multiple proteins in one of the species were classified as orthologs, the protein pair with the highest global identity was kept. In those cases were multiple orthologs were predicted in both species, all pairs with the bidirectional highest global alignment identity were kept. The final ortholog set only contained protein pairs that were fully supported by EST/cDNA evidence in both species.

As explained above, the classification of some AS events depends on which isoform is chosen as the reference. One commonly used method is to choose the isoform with highest amount of transcript evidence as the reference isoform (Xing et al., 2004). The choice can also be based on the evolutionary conservation of either the exon/intron structure of transcript isoforms or the protein isoforms (Tress et al., 2008). In the current comparative analysis, the transcript isoforms corresponding to the orthologous proteins were chosen as the reference isoforms.

Comparative analysis of AS patterns at the gene structural level (Figure 3) involved searching for conserved introns within the coding regions (CDS) that were predicted to be alternatively spliced in both species. To achieve this, the introns located within the CDS underlying the aligned orthologous proteins were projected onto their global alignment. Conserved introns were subsequently identified as “aligned” introns with the same phase or as “unaligned” introns with the same phase within a distance of six amino acids that could unambiguously be matched.

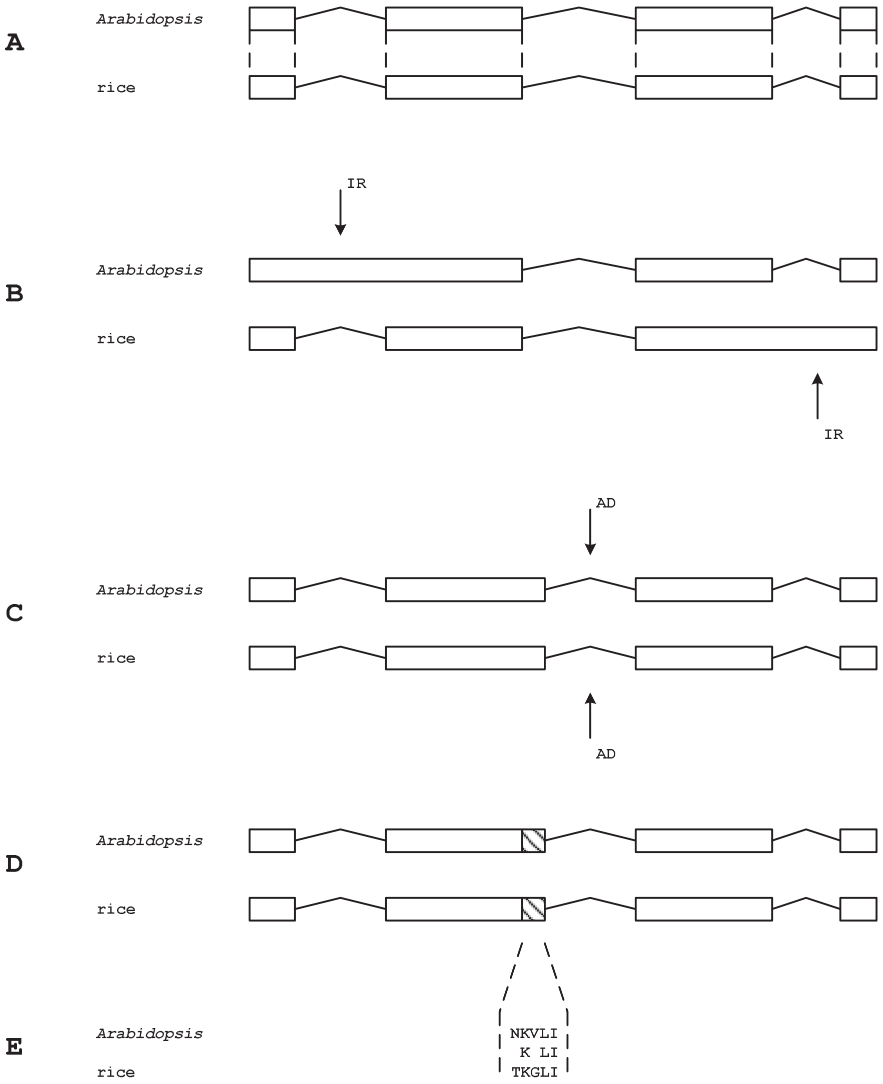

Figure 3. Dissection of AS events in orthologous genes. A) Orthologous exon and introns within the CDS region that have been identified through projection of introns onto the global alignment of the encoded protein sequences. B) Both species have an intron retention event (IR) but on different positions, which is designated “different position”. C) Both species have an alternative donor event (AD) on the same position (“same type, same position”). D) These events result in the addition of a stretch of amino acids at homologous sites (“homologous modification sites”). E) These events result in similar changes in the protein sequence (”similar modification”).

The second level of comparison involved the identification of similar changes to the orthologous protein sequences resulting from AS events. Three types of similar changes were investigated: (*i*) similar additions were identified by searching for identical insertion sites and comparing the stretches of amino acids; (*ii*) similar substitutions caused by frame shifts were identified by searching for aligned amino acids that were substituted in AS isoforms in both species and comparing the stretches of amino acids resulting from the frame shift(s); (*iii*) similar deletions were detected by searching for stretches of aligned amino acids that were deleted as the results of AS in both species. These changes at identical positions were designated "homologous modification sites". Both (*i*) and (*ii*) were designated "similar modification" if the similarity of the aligned stretches was over 40%.

Comparison of mechanistic patterns

In contrast to cross species comparison of AS induced polymorphisms, the comparison of mechanistic patterns was not restricted to orthologous gene pairs. The InterProScan program (Zdobnov and Apweiler, 2001) version 4.3.1 was used for the identification of Pfam domains version (21.0) (Finn et al., 2006) in the deduced protein sequence of each isoform. Only domains that were predicted with an e-value < 0.01 were considered.

The identification of domains that were disproportionately distributed towards either constitutively (CS) or alternatively spliced (AS) genes was performed as described elsewhere (Liu and Altman, 2003), with the exception that the *Holm's* correction method (Holm, 1979) was used in the present study for controlling false discovery rates.

Prior to testing whether particular domains were significantly enriched with either constitutively or alternatively spliced introns, clusters of overlapping introns were constructed. All clusters containing more than one unique intron and singleton introns that were retained in one or more isoforms were labeled as alternative splicing (AS) clusters. All other singleton introns were labeled as constitutive (CS) clusters. This clustering step prevents a bias in the results caused by over-counting of individual AS introns (with the exception of CI and RI events, all other events involve at least two AS introns). Fisher exact tests (with Holm's correction method) were performed for each domain type to test whether the ratio between its number of associated AS and CS introns deviated significantly from the ratio between all AS and CS introns located within the boundaries of all domains other than the one being tested.

A spliced-out domain (Resch et al., 2004) was identified as a domain that was present in only one of the isoforms of a reference/hypothetical isoform pair. The detection of spliced-out domains was only performed on pairs of isoforms that were not predicted to be likely targets for the NMD-pathway.

Authors’ contributions

ES designed the study, performed the analysis and drafted the manuscript. AvD participated in the analysis and preparation of the manuscript. WS supervised the study. RvH initiated and supervised the study and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the BioRange programme (SP 3.2.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI).

Chapter 3

Assessing the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana* using proteomics data

Edouard I Severing, Aalt DJ van Dijk
and
Roeland CHJ van Ham

Published in BMC Plant Biology (2011) 11(1): 8

Abstract

Background

Large-scale analyses of genomics and transcriptomics data have revealed that alternative splicing (AS) substantially increases the complexity of the transcriptome in higher eukaryotes. However, the extent to which this complexity is reflected at the level of the proteome remains unclear. On the basis of a lack of conservation of AS between species, we previously concluded that AS does not frequently serve as a mechanism that enables the production of multiple functional proteins from a single gene. Following this conclusion, we hypothesized that the extent to which AS events contribute to the proteome diversity in *Arabidopsis thaliana* would be lower than expected on the basis of transcriptomics data. Here, we test this hypothesis by analyzing two large-scale proteomics datasets from *Arabidopsis thaliana*.

Results

A total of only 60 AS events could be confirmed using the proteomics data. However, for about 60% of the loci that, based on transcriptomics data, were predicted to produce multiple protein isoforms through AS, no isoform-specific peptides were found. We therefore performed *in silico* AS detection experiments to assess how well AS events were represented in the experimental datasets. The results of these *in silico* experiments indicated that the low number of confirmed AS events was the consequence of a limited sampling depth rather than *in vivo* under-representation of AS events in these datasets.

Conclusion

Although the impact of AS on the functional properties of the proteome remains to be uncovered, the results of this study indicate that AS-induced diversity at the transcriptome level is also expressed at the proteome level.

Background

Alternative splicing (AS) is a common phenomenon in higher eukaryotes that involves the production of multiple distinct mRNA molecules from a single gene. RNA-Seq surveys have shown that more than 90% of human and over 40% of *Arabidopsis thaliana* and rice genes are capable of producing multiple diverse mRNA molecules through AS (Wang et al., 2008; Filichkin et al., 2010; Lu et al., 2010). A large fraction of AS events are predicted to result in transcripts that encode premature termination codons (see for instance Filichkin et al., 2010; Zhang et al., 2010) and that are likely to be degraded through the nonsense mediated decay (NMD) pathway (Lewis et al., 2003). Although it has been the subject of several genome-wide studies (e.g. Tress et al., 2007; Melamud and Moult, 2009; Severing et al., 2009), the extent to which the remaining fraction of AS events contribute to the functional protein repertoires of eukaryotes remains relatively unknown.

We concluded in a previous genome-wide comparative analysis of AS in three plant species that AS does not substantially contribute to functional diversity of the proteome (Severing et al., 2009). Our conclusions were based on the limited conservation of AS events that can contribute to proteome diversity and the lack of conserved patterns that relate AS to gene function. Following this conclusion, it is conceivable that most AS events, in particular those that are not targeted towards NMD, result from noise in the splicing process (Melamud and Moult, 2009) and are not strongly manifested at the protein level. However, lack of conservation can also mean that many protein isoforms have a confined, species-specific function rather than no function at all. In this scenario, it might be expected that most AS events are also expressed at the protein level. Determining which of these two scenarios is the most likely has been a difficult task because the majority of genome-wide studies of AS have been performed using protein isoforms deduced from transcriptomics data. For most of these isoforms no evidence for their expression at the protein level was available.

The gap between the availability of transcriptomics and proteomics data is steadily being bridged by the advancing field of mass spectrometry-based proteomics. This technology, which can be used to characterize complex protein mixtures (Aebersold and Mann, 2003), is of great value for studying the impact of AS at the proteome level. Indeed, a number of studies have appeared that describe the use of proteomics data for the identification of protein polymorphisms that are the result of AS (Tanner et al., 2007; Mo et al., 2008; Tress et al., 2008).

In this study we address the impact of AS on proteome diversity in the model species *Arabidopsis thaliana* by reanalyzing the data from two independent large-scale proteomics studies (Baerenfaller et al., 2008; Castellana et al., 2008). Although AS was briefly addressed in these studies, their primary focus was on the confirmation and revision of existing gene structures and on the identification of new protein coding genes. The main objective of our study is to assess whether the predicted contribution of AS to the proteome diversity in *A. thaliana*, as based on transcriptomics data, is indeed observed at the proteome level.

We limited our study to those AS events that could be deduced from the annotated gene structures in the genome annotation database of *A.thaliana* version TAIR 10.0 (www.arabidopsis.org) and that are predicted to contribute to proteome diversity in this species. The absolute numbers of AS events that could be confirmed using the experimental peptide sets were by themselves not very indicative for the contribution of AS to the proteome diversity in *A. thaliana*. This is because these numbers depend on the depth of sampling in the experiments. We therefore performed *in silico* AS detection experiments using randomly generated peptide sets to assess the representativeness of the experimental sampling. This type of *in silico* experiments has previously been described and applied to *Drosophila* data (Tress et al., 2008).

We show that the outcome of the *in silico* experiments can lead to conflicting conclusions about the impact of AS on the proteome diversity, depending on the assumption that is used for generating the random peptide sets. We evaluate two of such assumptions and according to the biologically most realistic one, we show that AS events were not under-represented in the analyzed proteomics sets. This implies that variation due to splicing is to a large extent expressed at the proteome level.

Results

Throughout this study we used three experimental datasets, the first two of which, hereafter referred to as the Castellana and Baerenfaller sets, contain peptides from two large-scale proteomics experiments on *A. thaliana* (Baerenfaller et al., 2008; Castellana et al., 2008). The third set, hereafter called the Merged set, was created by merging the Castellana and Baerenfaller sets into a non-redundant set. As it was essential for our study that each experimentally identified peptide could be reproduced by an *in silico* digestion of its parent protein, we only considered those peptides that met the following criteria: first, only one missed cleavage site (internal lysine or argine residues that were not used as cleavage sites by the trypsine enzyme) was allowed per peptide. Second, only those peptides that could be mapped to their parent proteins according to a strict set of rules were considered (see Material and Methods).

The initial set of annotated *A. thaliana* proteins (TAIR10.0) was also filtered by removing all proteins for which the exon/intron structure underlying its CDS region was not sufficiently supported by transcript data (see Material and Methods). The filtered protein set contained a total of 25,039 unique protein sequences derived from 21,136 nuclear-encoded, protein-coding TAIR 10.0 loci. Around 14.2% of the loci within the filtered protein set were predicted to produce distinct proteins through AS (hereafter called AS loci).

Peptide mapping

The number of peptides that could be mapped back to TAIR 10 proteins (excluding chloroplast and mitochondrial encoded proteins) and the number of TAIR loci with at least one uniquely mapped peptide are summarized in Table 1. Although the number of mapped peptides from the Castellana set was slightly smaller than that of the Baerenfaller set, more loci were identified with the peptides from the Castellana set. However, the Castellana set was ~1.5 times larger than the initial Baerenfaller set (Table 1) and thus already represented more loci prior to the filtering step.

Table 1. Identification of nuclear encoded TAIR 10 loci.

Set	Total number of peptides[a]	Nr. of mapped peptides	% of peptides mapped	Nr. of TAIR loci identified	% of TAIR loci identified
Castellana	131,077	71,243	54.4	12,067	57.1
Baerenfaller	86,078	72,264	84.0	11,282	53.4
Merged	179,174	109,293	61.0	14,190	67.1

[a] Totals refer to peptides containing at most one missed cleavage site

We note that a large fraction of the peptides from both the Baerenfaller (~16%) and Castellana (~45%) sets could not be mapped to any protein using our stringent criteria. These were kept stringent to ensure reproducibility of mapping results in the *in silico* experiments.

AS detection results

AS events correspond to specific differences between the intron/exon architectures of two transcripts. If the AS event is located in the coding region of these transcripts, the resulting protein isoforms will in many cases differ by an indel (only these type of sequence variations were considered in this study). In order to confirm the contribution of a particular AS event to proteome diversity, peptides have to be identified that uniquely map to the variable protein regions that are associated with the AS event (Figure 1A and 1B). In addition, these peptides have to map according to a specific set of rules that differs per AS event type (see Material and Methods). Due to the preference of trypsin to cleave after K- and R-residues (Olsen et al., 2004), only a fixed number of peptides can, at least in theory, be obtained from a particular protein upon complete digestion. However, certain AS events may not be detectable because the peptides needed to confirm the events are not produced during digestion. Taken all together, the number of AS events that can be confirmed using proteomics data not only depends on the sampling depth and the number of co-expressed protein isoforms in a given sample, but also on the sequences of these proteins. For each of the experimental sets it was therefore determined what number of AS events could theoretically be confirmed (identifiable AS events). This was done by first performing an *in silico* digestion of all TAIR 10.0 proteins encoded by the loci that were expressed (represented by proteins) in the biological samples. The resulting *in silico* generated peptides were then mapped to their parent proteins and subsequently used for confirming AS events in the same way as was done for the experimentally identified peptides (Table 2).

Table 2. Experimentally confirmed AS events. For each experimental set, the number of TAIR 10 loci with identifiable AS events (AS loci) are given together with the number of identifiable AS events. Both the number of identifiable AS events that were confirmed using experimentally identified peptides and the number of AS loci with at least one confirmed AS events are provided. The percentages are fractions of AS loci and identifiable AS events.

Set	AS loci	Identifiable AS events	AS loci w. confirmed AS events	(%)	Number of confirmed AS events	(%)
Castellana	1,434	1,789	38	2.6	38	2.1
Baerenfaller	1,318	1,641	21	1.6	21	1.3
Merged	1,644	2,059	59	3,6	60	2.9

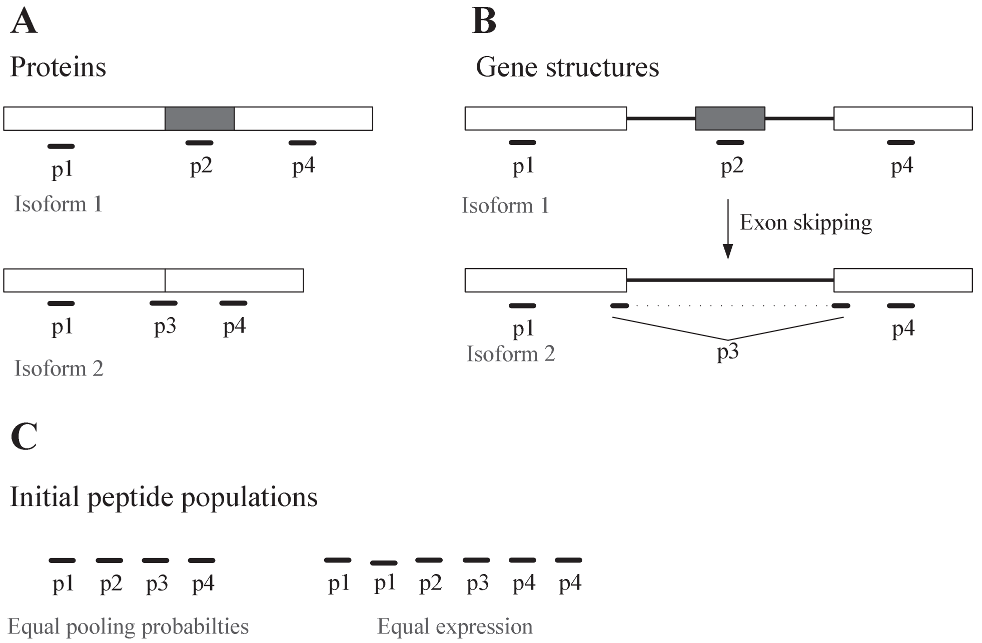

Figure 1. Isoform and non-isoform specific peptides. (A) Two protein isoforms (1 and 2) from an alternatively spliced gene that differ by a local polymorphism (inclusion/exclusion grey rectangle) yield two different peptide sets (Isoform 1: p1, p2, p4; Isoform 2: p1, p3, p4) when digested. While peptides p1 and p4 are non-specific because they map to both isoforms, peptides p2 and p3 are specific for isoform 1 and isoform 2, respectively. **(B)** The gene structures (exons correspond to the rectangles and the lines connecting them represent the introns) underlying these protein isoforms show that the AS event that is associated with the variable protein region is an exon-skipping event. In order to confirm the contribution of this specific exon-skipping event to the proteome diversity, both peptides p2 and p3 need to be identified. The dotted line indicates that p3 spans an exon/exon junction. (**C**) The initial peptide populations that are constructed for the *in silico* AS detection experiments differ under the two probability assumptions that are used in this study. Under the "equal pooling probability" assumptions, the initial peptide population consists of only unique peptides. Therefore the population contains only four different peptides. Under the "equal expression" assumption, the isoforms are represented by equal numbers of molecules prior to digestion. As a result, non-specific peptides are more abundant than isoform specific peptides.

A total of 38 AS events, corresponding to 38 AS loci were confirmed using the experimentally identified peptides from the Castellana set. Usage of the peptides from the Baerenfaller set resulted in the confirmation of 21 AS events from 21 AS loci (Table 2). Although more peptides from the Baerenfaller set could be mapped to their parent proteins than from the Castellana set, more AS events were confirmed using the latter set (Table 2). Comparison of the AS loci revealed that seven AS loci had confirmed AS events in both the Castellana- and Baerenfaller sets. In total, 60 AS events corresponding to 59 AS loci were confirmed using the experimental peptide set. These AS events represent ~2.9% of all AS events that could theoretically be confirmed using the merged peptide set. We note that for the Merged set the number of confirmed AS events was higher than the number of AS loci with confirmed AS events. This was due to a single AS locus that had more than one confirmed AS event.

Sampling of AS regions

Next, we analyzed how well protein regions that corresponded to the location of AS events were sampled in each of the experimental sets. Here, sampling refers to the identification of peptides that map to either one of the two protein variants that are associated with an AS event. This is illustrated by the example shown in Figure 1A and B, in which either peptide p2 or p3 is identified, but not necessarily both. The analysis revealed that around 29% to 36 % of AS events corresponding to ~31-38% of AS loci were sampled (Table 3).

Table 3. Sampling of AS events. The percentage of identifiable events that have been sampled (i.e. at least one peptide is present that covers the region where AS induces local variation) and the percentage of AS loci with at least one sampled AS event are provided. The percentages in this table are relative to the number of identifiable AS events and AS loci for the corresponding sets as provided in Table 2.

Set	Nr. of sampled AS events	% of identifiable AS events	AS loci w. sampled events	% of AS loci
Castellana	525	29.3	446	31.1
Baerenfaller	537	32.7	452	34.3
Merged	748	36.3	626	38.1

In silico AS detection experiments

In silico AS detection experiments (Figure 2) were performed to assess how well AS events were represented in the experimental peptide sets. In brief, because of our strict mapping rules, all the experimental peptides that were considered in this study could be reproduced by performing an *in silico* digestion of the parent protein. As a result, each experimental peptide set was in fact a subset of an initial population that was generated by performing an *in silico* digestion of all annotated proteins encoded by the loci that were expressed in the biological samples. It was therefore possible for each of the experimental sets to test whether the number of confirmed AS events significantly differed from the number of events that can be expected to be confirmed using an equally sized, random subset of the same initial peptide population. The expected number of events corresponded to the average number of AS events that could be confirmed using 1000 randomly pooled peptide sets.

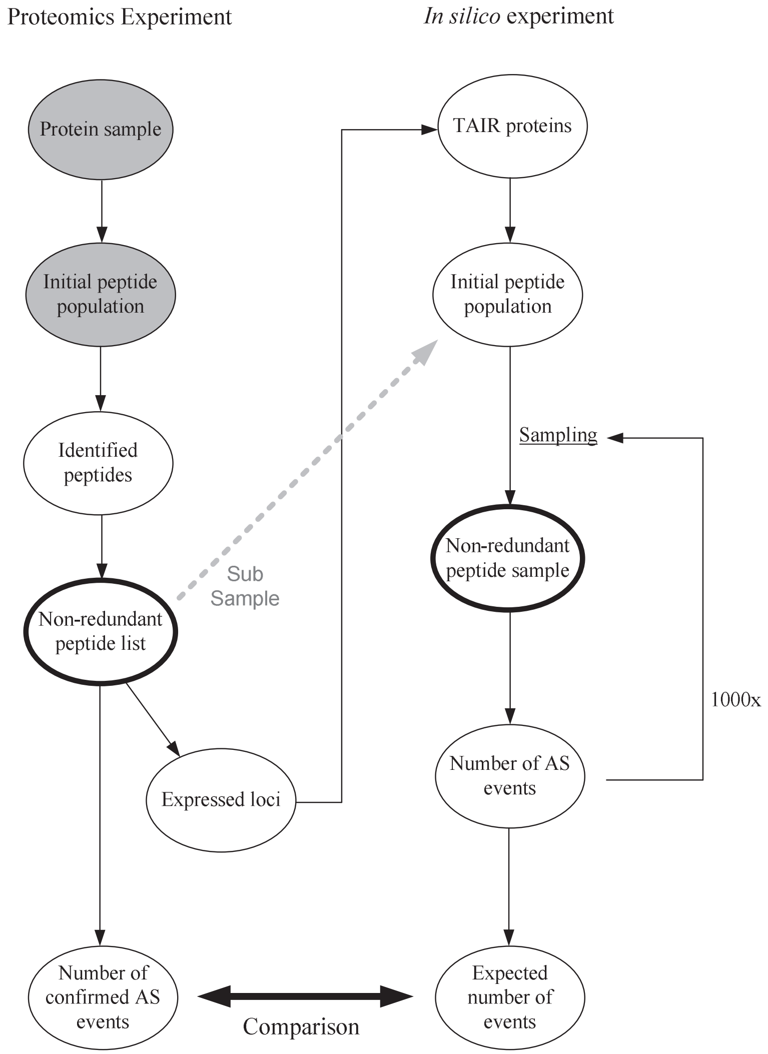
Proteomics Experiment
In silico experiment
Protein sample
Initial peptide population
Identified peptides
Non-redundant peptide list
Sub Sample
Expressed loci
Number of confirmed AS events
Comparison
TAIR proteins
Initial peptide population
Sampling
Non-redundant peptide sample
1000x
Number of AS events
Expected number of events

Figure 3. Workflow for *in silico* AS detection experiments. In the experimental proteomics study (left workflow), the (unknown) protein sample was digested using a protease enzyme. For a subset of the (unknown) initial peptide population the amino acid sequence was determined. This non-redundant peptide list was used for determining which loci were expressed (represented by a protein product) in the protein sample. The starting point for the simulations (right workflow) is a set of all annotated (TAIR) proteins encoded by the loci that were expressed in the biological sample. An initial peptide population was created by performing an *in silico* digestion of the set of annotated proteins. Note that the non redundant list of experimentally identified peptides is a subset of the *in silico* generated initial peptide population (grey dashed arrow). One thousand non-redundant peptide samples equal in size to the non-redundant list of experimentally identified peptides (thick lined ellipses in both workflows) were pooled from the initial peptide population. For each of the pooled peptide samples the number of AS events that could be confirmed with that sample was determined. Finally, the number of experimentally confirmed AS events was compared to the expected number of AS events which corresponds to the average number of AS events confirmed using the randomly generated peptide samples.

The composition of the random peptide sets and therefore also the AS detection outcome depends on the pooling probabilities that are assigned to the individual peptides in the initial *in silico* peptide populations. These pooling probabilities simply reflect the relative abundances of the peptides within the initial populations (see Material and Methods). We used two different assumptions for assigning pooling probabilities to the individual peptides (Figure 1C). The first assumption, to which we refer as the "*equal pooling probability*" assumption, has previously been described by Tress and co-workers (Tress et al., 2008). Under this assumption, all peptides in the initial population are unique and therefore have the same probability of being pooled. Under the second assumption, hereafter referred to as the "*equal expression*" assumption, it was assumed that all genes were represented by equal numbers of protein molecules and that all isoforms of an AS locus were equally abundant in the protein sample. A consequence of this assumption was that the peptides within the initial populations were not equally abundant (Figure 1C).

Under the "*equal pooling probability*" assumption, the number of experimentally confirmed AS events in the Castellana set was 2.2 times smaller than the expected number of events as determined by the *in silico* experiments (Table 4; Simulations A). For the Baerenfaller and Merged sets, this same ratio was 4.8 and 2.7, respectively. Hence, when equal pooling probabilities are assumed, the *in silico* experiments indicate that AS events were underrepresented in all experimental peptide sets.

Table 4. *In silico* AS detection experiments. The means and standard deviations are provided for the AS events that were confirmed in the *in silico* AS detection experiments. The experiments were performed under both the "*equal pooling probability*" (A) and "*equal expression*" (B) assumptions.

Set	Number of experimentally confirmed AS events	Simulations A		Simulations B	
		Mean nr. of AS events	SD	Mean nr. of AS events	SD
Castellana	38	85.4	9.3	20.5	4.7
Baerenfaller	21	100.4	10.1	26.0	5.2
Merged	60	160.6	12.9	39.7	6.4

A different picture emerged from the simulations performed using the "*equal expression*" assumption. In this case, the number of experimentally confirmed AS events in the Castellana set was around 1.9 times larger than the expected number of events (Table 4; Simulations B). In contrast, the number of experimentally confirmed AS events for the Baerenfaller set fell within just 1 SD of the mean number of events as determined by the *in silico* experiments. Finally, the number of experimentally confirmed events for the Merged set was one and a half times larger than the expected number of events. In summary, under the "*equal expression*" assumption the *in silico* experiments indicate that; (*i*) AS events were not under-represented in the Baerenfaller set, and; (*ii*) AS events were over-represented in both the Castellana- and the Merged set.

Disordered regions

The peptides in both the Castellana and Baerenfaller set were extracted from different organs and cell cultures. However, the Castellana set also contained peptides that were derived from a phosphopeptide-enriched sample. It has previously been shown that phosphopeptide enrichment can result in an enhanced detection of AS events that are typically located within disordered regions of proteins (Tress et al., 2008). Analysis of the protein regions to which the peptides from each experimental set were mapped indeed revealed a higher fraction of peptides mapping to disordered regions in the Castellana set than in the Baerenfaller set (Figure 3). For the Merged set this fraction fell, as expected, in between those for the Baerenfaller and Castellana sets. Comparison to the same fraction calculated for the TAIR set (peptides generated from all nuclear encoded TAIR proteins and mapping to disordered regions) revealed not much difference with the Castellana. However, the fractions for the Baerenfaller and Merged sets were smaller than the fraction for the TAIR set. Hence, compared to the TAIR - and Castellana sets, disordered regions were under-represented in the Bearenfaller and Merged sets.

Next, it was investigated whether the experimentally confirmed AS events were biased towards or against disordered regions, relative to expectation. To this end, the fraction of experimentally confirmed AS events from disordered regions was compared to a theoretical fraction. This theoretical fraction corresponded to the average fraction of AS events from 1000 randomly generated AS event sets (containing the same number of events as the corresponding experimental set) that overlapped with disordered regions. The AS events within these randomly generated sets were pooled from all identifiable AS events. Note that the number of identifiable AS events differs per experimental set. The results indicated that experimentally confirmed AS events were biased towards disordered regions in the Castellana set (Figure 4). Removal of all peptides containing phosphorylated residues (8,128 peptides) from the Castellana set did not affect this result (data not shown). In contrast, the fraction of confirmed AS events from the Baerenfaller set that were located in disordered regions was lower than its theoretical fraction. Finally, the fraction of AS events from the

Merged set that were located in disordered regions was, similar as for the Castellana set, higher than its theoretical fraction. In summary, while the AS events in the Bearenfaller set were biased against disordered regions, the opposite was true for the AS events in the Castellana and Merged sets.

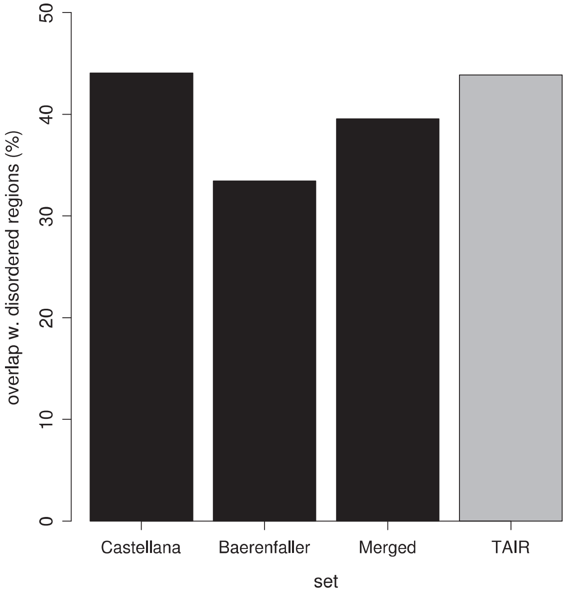

Figure 3. Fraction of peptides that overlap with predicted disordered regions. The fraction of peptides that overlap with disordered regions for all experimental sets (black) are shown together with the fraction of peptides generated through an *in silico* digestion of all nuclear encoded TAIR proteins that overlap with disordered regions (grey).

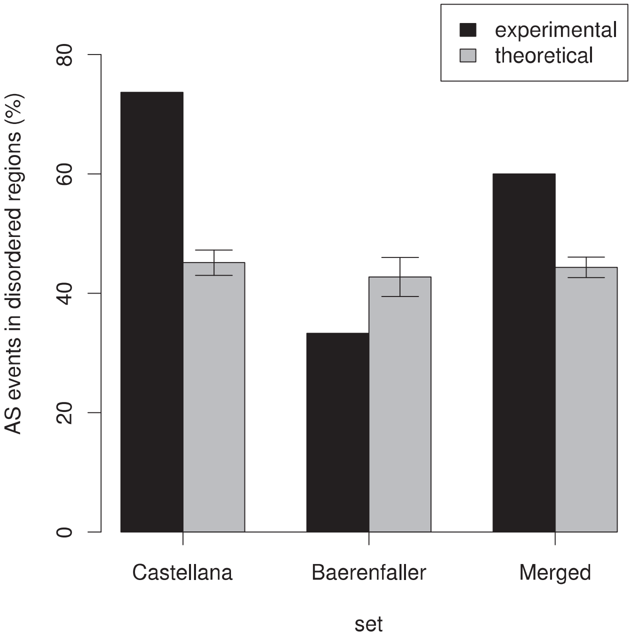

Figure 4. AS events overlapping with disordered regions. For all sets, the fraction of experimentally confirmed AS events that overlap with disordered regions (black) is shown next to the mean fraction of simulated events that overlap with disordered regions (grey). Error bars correspond to 1 SD from the mean.

Discussion

Genome-wide studies that address the impact of AS on proteome diversity have thus far mainly been performed using indirect evidence from transcriptomics data. Data that can be used to directly assess this impact is increasingly being provided by high-throughput proteomics experiments. Here we studied the impact of AS on proteome diversity in the model species *Arabidopsis thaliana* by reanalyzing data from two previous, large-scale proteomics studies (Baerenfaller et al., 2008; Castellana et al., 2008). The main goal of our study was to determine whether the contribution of AS events to proteome diversity as predicted using transcriptomics data, is indeed observed at the proteome level.

The absolute numbers of AS events that could be confirmed using the experimentally identified peptides were not particularly high and only represented around 2 to 3% of identifiable AS events. Analysis of the representation of protein regions corresponding to the location of AS events that were sampled in the experiments showed that for roughly two thirds of AS loci no peptides were detected that could discriminate between the different protein isoforms. The absolute numbers of confirmed AS *per se* are therefore not very indicative for the extent to which AS contributes to proteome diversity in *A. thaliana.*

We performed *in silico* AS detection experiments to determine how well AS events were represented in the biological samples, given the sampling depth achieved in the proteomics experiments. The *in silico* experiments should thus reveal whether the number of AS events identified using the experimental peptide sets significantly deviated from the expected number of AS events. The latter was calculated using an equally-sized random subset of *in silico* peptides pooled from the an initial peptide population. This initial peptide population consisted of all peptides that theoretically could be obtained through digestion of the proteins (including isoforms resulting from AS) that were encoded by the loci expressed in the experimental samples.

One factor that critically influenced the outcome of these *in silico* experiments involved the pooling probabilities that were assigned to the individual peptides in the initial population. We performed the *in silico* experiments using two different pooling probability assumptions. The first, "*equal pooling probability*" assumption, indicated that AS events were under-represented in all experimental peptide sets. In a previous proteomics study performed on *Drosophila* data, the same "*equal pooling probability*" assumption was used for generating peptide samples and determining the number of expected AS events (Tress et al., 2008). The results in our study are comparable to those obtained for the Brunner set in that study.

The results of the *in silico* experiments were very different for the "*equal expression*" assumption. In this case, AS events were found to be over-represented in the Castellana and Merged sets, while for the Baerenfaller set, the number of experimentally identified AS events fell within 1 SD of the expected number of events. The observation that AS events were not under-represented in the experimental samples corresponds to the results of a recent study in which many AS transcript isoforms were shown to be actively translated (Jiao and Meyerowitz, 2010).

The inconsistency between the conclusions obtained under the two pooling probabilities assumptions is the result of the fact that isoform-specific peptides associated with AS events have higher pooling probabilities under the "*equal pooling probability"* assumption than under the "*equal expression*" assumption. Under the first assumption, isoform-specific peptides and non isoform-specific peptides are equally abundant. In contrast, under the "*equal expression*" assumption, non isoform-specific peptides are more abundant than isoform-specific peptides (Figure 1C). This difference results in different pooling probabilities, in which the "*equal pooling probability*" assumption provides an upper bound to the expected number of AS events. The "*equal expression*" assumption, however, does not provide a corresponding lower bound, because it does not consider the relative expression levels between two or more AS isoforms. Indeed, the effect of lowering of the expected number of events would only further increase if unequal expression of isoforms would be taken into account and would therefore strengthen the conclusion that AS events were not under-represented in the experimental peptide sets.

Although neither of the two pooling probability assumptions is truly realistic in a biological sense, the "*equal expression*" assumption arguably provides the better approximation. This follows from the fact that isoform-specific peptides are necessarily less abundant than non-isoform specific peptides. Using Figure 1 as illustration, this can be understood by considering the total amount of peptides produced from a single locus, whatever the relative expression level of the two underlying isoforms is: the amounts of the constitutive peptides p1 and p4 will be the same and will always equal the sum of p2+p3. Given this reasoning, the conclusion derived under the "*equal expression*" assumption, namely that AS is over-represented, or at least not under-represented in the experimental proteomics datasets, is the most plausible.

A key factor that might explain the over-representation of AS events in the Castellana set compared to the Baerenfaller set, involves the bias of AS events towards disordered regions of proteins in the former set. AS events located within disordered regions can introduce variations that have a limited impact on protein folding (Romero et al., 2006). Because cells have evolved mechanisms that can recognize and remove incorrectly folded proteins (Goldberg, 2003), AS events that have a limited impact on the protein structure are more likely to be viable and manifested at the protein level. In fact, it has recently been shown that pairs of AS isoforms, for which evidence was available that they were expressed, differed by polymorphisms that were more often located within disordered regions than expected (Hegyi et al., 2011).

One property of disordered regions is that they allow proteins to bind with multiple partners with high specificity and low affinity (Dunker et al., 2008). AS within such regions are interesting because they might play an important role in regulating protein-protein interactions.

Conclusions

We conclude that the low numbers of AS events that could be confirmed using the proteomics datasets for *A. thaliana* are the result of a relatively low depth of sampling in the proteomics experiments. *In silico* AS detection experiments, performed under the assumption of equal expression of isoforms, indicate that AS events were not under-represented in the experimental peptide sets. An important implication of this is that much or all of the AS variation in *A. thaliana* that is expressed at the transcriptome level and not degraded through the NMD pathway, is also manifested at the proteome level. The true extent, however, to which AS variants are functional remains to be uncovered. Given that AS variation is not well conserved in plants (Severing et al., 2009), genome-wide expression of AS variation at the proteome level could point to the possibility that many of the AS events are associated with protein isoforms that either have a species-specific function or that are stable enough to escape rapid protein turnover.

Material and Methods

Initial data

Peptide sequences from the study performed by Baerenfaller and co-workers (Baerenfaller et al., 2008) were obtained by querying the Pride database (Jones et al., 2008) using the available BioMart interface. Peptide sequences from the study of Castellana and co-workers (Castellana et al., 2008) were downloaded from the webpage of the authors (site referenced in their publication). An additional peptide set was constructed by merging the Baerenfaller and Castellana peptide sets into a non-redundant set. Because trypsin was used for digesting proteins in both proteomics studies, peptides containing internal lysine (K) or arginine (R) residues that were not immediately followed by a proline (P) residue, were considered to be the result of missed cleavage sites. All peptides that contained two or more missed cleavage sites were discarded.

The predicted proteome of *Arabidopsis thaliana* version TAIR 10 was downloaded from www.arabidopsis.org. The information within the “confidenceranking_exon”-file (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/confidenceranking_exon) was used for filtering the proteome using the following criteria: (*i*) a protein encoded by a multi exon gene was only kept if all splice junctions located within the corresponding CDS region were supported by transcript data (mRNA) data, and; (*ii*) a protein encoded by a single exon gene was kept if at least 80% of the gene was supported by transcript data.

Mapping peptides against their parent proteins

Vmatch (http://www.vmatch.de/) was used for performing exact searches with the peptides against the filtered proteome of *A. thaliana*. All matches were subsequently filtered using the following criteria: (*i*) peptides that did not map to the C-terminus of their parent protein were required to have a K- or R- residue at their C-terminus; (*ii*) peptide matches were discarded if the corresponding region of the parent protein was not immediately preceded by a K- or R-residue, unless the peptide mapped to the N-terminus of the parent protein; (*iii*) peptide matches were discarded if the corresponding region of the parent protein was immediately followed by a P-residue. Finally, only those proteins were considered that had at least one mapped peptide which was unique for the locus from which the protein originated.

Identification of AS events at the proteome level

AS events were deduced from the annotated gene structures using a previously described method (Severing et al., 2009). The identification of AS events at the proteome level was only performed with peptides that were unique for one or more, but not all of the protein isoforms of a locus. A schematic overview of the rules that were used for the identification of AS events at the proteome level is provided in Figure 5.

In silico generation of peptide fragments

Peptides were generated by performing an *in silico* trypsin digestion involving cleavage after K- and R- residues that were not followed by a P-residue. Only one missed cleavage site was allowed per peptide. All peptides with a mass outside the observed mass-range of the experimentally identified peptides (~523-5,399 Da and ~725-4,962 Da for the Castellana set and Baerenfaller set, respectively) were discarded.

In silico AS detection experiments

The *in silico* AS detection experiments involved randomly pooling non-redundant peptide samples, equal in size to the experimental peptide samples, from an initial peptide population. This initial population only contained peptides that mapped to the protein products encoded by the loci which were expressed in the experimental samples. The probability of pooling a particular peptide depends on its abundance within the initial peptide population.

The *in silico* detection experiments were performed using either one of the following two assumptions on the abundance of individual peptides within the initial peptide populations.

Under the first assumption to which we refer as the "equal pooling probability" assumption, all *in silico* generated peptides are equally abundant and therefore have the same probability (*1/N*) of being pooled, which depends on the size of initial peptide population (*N*). This pooling strategy, which has previously been described in (Tress et al., 2008), reflects a biological scenario in which individual proteins within an experimental sample are present in such numbers that subsequent digestion of the sample results in a population of equally abundant peptides.

Under the second assumption, to which we refer as the "*equal expression*" assumption, two basic rules are applied: (*i*) all genes are represented by equal amounts of protein molecules, and; (*ii*) all protein isoforms from an AS locus are present in equal numbers. The abundance of each protein within the sample is therefore determined as follows: Let *M* be the number of protein isoforms produced by the alternatively spliced gene with the highest number of unique protein isoforms. In order for rule (*i*) to be fulfilled, each gene has to produce *M* protein molecules. The protein product from a constitutively spliced gene is

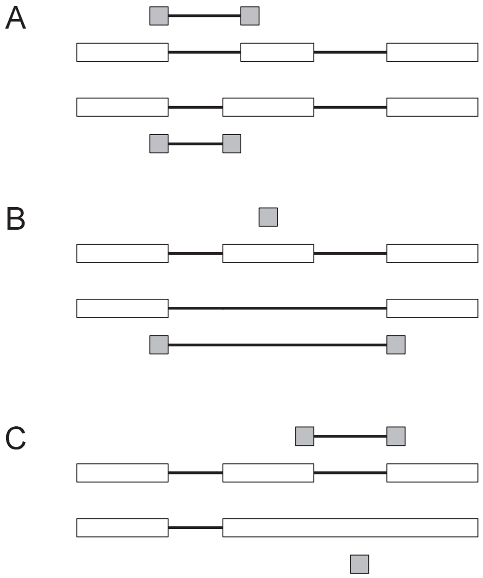

Figure 5. Detection of protein polymorphisms resulting from AS events. A. Detection of alternativeacceptor events: Peptides (grey) are required that span both overlapping introns that differ in their acceptor sites. (The same rule is applied for alternative donor and position events). **B.** For the confirmation of exon skippingevents it is required that at least one peptide (grey) is mapped to the skipped exon and one peptide spans the intron on the second isoform in which the skipped exon is fully engulfed. **C.** Intron retention events are confirmed when at least one peptide (grey) spans the retained intron and another peptide spans the spliced intron.

therefore present *M* times within the entire protein sample. To fulfill rule (*ii*), the number of molecules that correspond to a particular protein isoform of an AS locus that produces *X* different protein isoforms equals *M / X*. As a consequence, each peptide originating from this specific protein isoform is also represented by *M / X* molecules in the total peptide mixture after digestion. When for simplicity each peptide within the final sample is considered to be unique (even when multiple exact sequence copies exists), its pooling probability equals its abundance divided by the total number of peptides within the initial peptide population.

Prediction of disordered regions

Putative disordered regions were predicted using the FoldIndex method (Prilusky et al., 2005) which is based on an algorithm developed by Uversky and co-workers (Uversky et al., 2000). In brief, the method uses hydrophobicity and net charge of protein sequence segments in order to distinguish disordered from ordered regions. By sliding over the protein sequences using a window of 51 AA and a step size of 1, disordered regions were identified as regions of at least five consecutive amino acid residues located in the centre of a window with a negative FoldIndex value.

Authors' contributions

EIS conceived the experiments, carried out the study and drafted the manuscript. ADJvD participated in the design of the study and in drafting the manuscript. RCHJvH conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the BioRange programme (SP 3.2.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI).

Chapter 4

Analysis of conservation of alternative splicing indicates limited contribution to protein diversity in the plant kingdom

Edouard I Severing, Aalt DJ van Dijk
and
Roeland CHJ van Ham

Abstract

Background
Although it is well established that alternative splicing (AS) substantially increases the transcriptome complexity of eukaryotes, the extent to which the process also contributes to functional diversity of the proteome is not clear. It has previously been suggested, based on the observation of limited conservation of AS, that the majority of AS events and associated protein variants are unlikely to be functional. However, an alternative hypothesis is that many AS isoforms are functional but preferentially retained by either becoming the constitutive isoform of a paralogous gene or by replacing existing isoforms. In this study we test this hypothesis by analyzing a large collection of EST/cDNA sequences from hundreds of different plant species.

Results
In order to determine whether the alternative ways of conservation of novel AS isoforms are more prevalent than conservation of AS events *per se*, we first analyzed the extent to which AS events were conserved in our dataset. Of all AS events that were detected, 16% were found to be conserved. For around 4% the corresponding AS isoforms were conserved as paralogs in another species. We also detected several hundred cases of putative isoform replacement. Rather than being evenly distributed, all types of AS isoform conservation

tended to cluster within a limited number of homologous sequence clusters. An additional analysis revealed that homologs with conserved AS events tend to have nucleotide distances that fall within the same range as those observed between paralogs.

Conclusions
Based on our results, we reject the hypothesis that AS isoforms are preferentially conserved by means other than retention of the corresponding AS event. We conclude that novel functions evolve only rarely through alternative splicing. We predict that the vast majority of novel AS isoforms are lost within a similar time period as redundant gene copies.

Introduction

It is by now well established that Alternative Splicing (AS) substantially increases the transcriptome complexity of higher eukaryotes by enabling the production of multiple distinct transcript isoforms from a single gene. When for a given transcriptome those transcripts that are either non-coding or degraded through the nonsense mediated decay pathway (Lewis et al., 2003) are discarded, a large number of transcripts that are likely to be translated into proteins remains. The extent to which the encoded proteins contribute to functional diversity of the proteome is however not well understood.

From an evolutionary perspective, AS events are comparable to gene duplication events in the sense that they provide genes with "internal" paralogs (AS isoforms) that can serve as the raw building blocks for the evolution of novel functions (Modrek and Lee, 2003). However, while recent gene duplicates are likely to produce the same set of transcripts as their parent genes, a newly evolved AS isoform will always be different from pre-existing transcript isoforms. AS can therefore be classified as a process that similarly as for instance point mutations and domain shuffling events (Patthy, 1999) increases genetic diversity.

There are three ways in which novel AS isoforms can be conserved (Figure 1). The first involves retention of the AS event through which the novel isoform emerged. Accumulation and selection of mutations can result in the acquisition of the sequence elements that are needed to control the expression levels of the different AS isoforms in a tissue- or condition specific matter (Keren et al., 2010). This way of conserving AS isoforms can be expected when the novel AS isoform does not functionally replace pre-existing isoforms but provides novel functionality. The second way of conserving multiple AS isoforms involves a gene duplication event that is followed by each of the duplicate genes adopting one of the AS isoforms of the parent gene. The conversion of "internal" to "external" paralogs, a process called externalization (Irimia et al., 2010), can be viewed as a subfunctionalization (Zhang, 2003) process because each of the daughter genes have a part of the functional diversity of the parent gene. Although several externalization cases have been described in literature (Yu et al., 2003; Rosti and Denyer, 2007; Irimia et al., 2010), the importance of this process, particularly in the plant kingdom, is unclear. Finally, novel AS isoforms can also be conserved by replacement of the pre-existing isoforms. This way of conserving AS isoforms can be expected when the function of the novel isoform is equivalent to, or an improvement of the function of a pre-existing isoform.

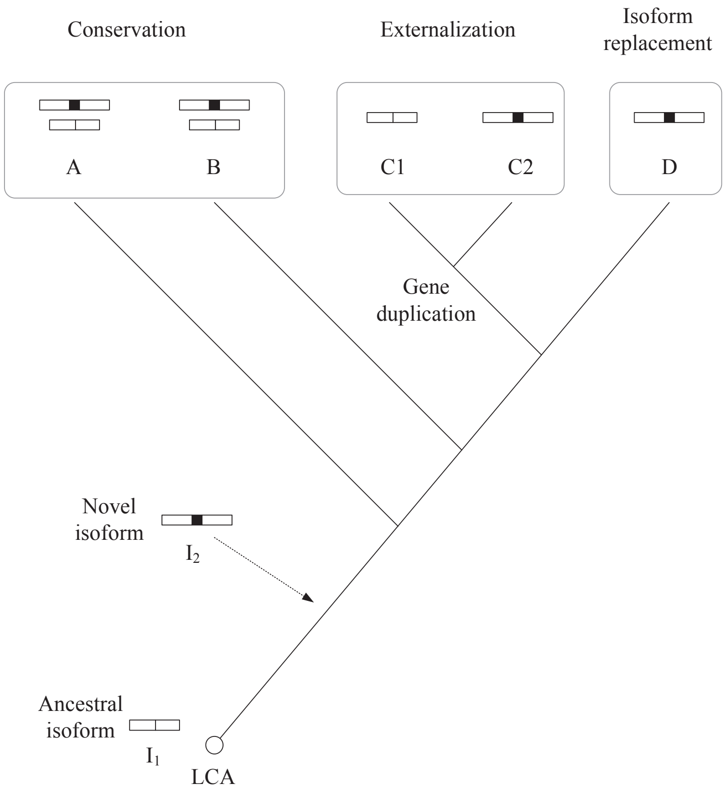

Figure 1. AS Isoform conservation. The last common ancestor (LCA) of species A, B, C and D possessed a gene which produced only one protein isoform (I_1). In time a novel isoform I_2 was gained through AS. The underlying AS event has been conserved in both species A and B and therefore both isoforms are produced by a single gene in these species. However, an externalization event occurred in species C because a duplication of the ancestral gene was followed by loss of one of the AS isoforms by each of the duplicates (C1 and C2) in a complementary way. In species D, the most ancestral isoform I_1 was replaced by I_2.

The results of several comparative analyses indicate that only a small fraction of AS induced variation within a species is likely to be conserved in another species (in the sense of the first type of conservation mentioned above, retention of the AS event itself) (Modrek and Lee, 2003; Nurtdinov et al., 2003; Yeo et al., 2005; Wang and Brendel, 2006; Baek et al., 2008; Severing et al., 2009). It has even been demonstrated that AS patterns can substantially differ between members of the same species (Nembaware et al., 2004; Kwan et al., 2008; Wang et al., 2008). Under the premise that functional genetic elements are more likely to be conserved than non-functional ones, the lack of conserved AS events could be an indication that the majority of AS events and isoforms are not functional. However, an alternative hypothesis that can explain the lack of conserved AS events, while still compatible with functional innovation through AS, is that novel AS isoforms are preferentially retained by either replacing existing AS isoforms or by being constitutively expressed by a duplicate gene (i.e. the second and third type of conservation discussed above).

In this study we test this hypothesis within the plant kingdom by analyzing a large collection of EST/cDNA data from hundreds of different species. By using the predicted genes from the well-annotated species *Arabidopsis thaliana* and *Oryza sativa* as informants, we were able to detect protein sequence variations that coincide with predicted intron boundaries using EST/cDNA sequences alone. In order to test whether AS isoforms are preferentially conserved through means other than retention of the AS event itself, we first analyzed the conservation of AS events within our dataset. We subsequently used our set of identified AS events in order to detect putative cases of AS isoform externalization. We identified instances of isoform replacement by searching for indel variations between homologous sequences that coincide with intron boundaries but which were not associated with AS conservation or externalization events. We further analyzed the nucleotide distances between homologous genes with conserved AS events in order to obtain an indication as to how long novel AS isoforms are likely to persist during evolution. We combined the results of these different analyses to determine whether AS isoforms are indeed preferentially retained in ways other than through conservation of the underlying AS event and to determine to what extent AS contributes to functional diversity of the proteome.

Results

We distinguish three different types of conservation of newly induced AS isoforms: (*i*) conservation of the AS event; (*ii*) externalization; and (*iii*) isoform replacement. Here, we first present the results for our identification of AS events and AS isoforms, and then analyze these three different kinds of conservation. Predicted genes from *Arabidopsis thaliana* and *Oryza sativa* were used as informants, and homologous sequence clusters (HSCs) were constructed by grouping all EST/cDNA sequences that had one of the protein products from the same informant gene as their best blast hit. Although sequences from different species within an HSC are homologs they are not by definition orthologs. Paralogs within an HSC could have originated from duplication events that either happened before or after the split of the subject and informant species. In the former case, one of the duplicate genes has been lost in the informant species, otherwise the paralogs would have been placed in separate HSCs. Hence, the outcome of this clustering method and the subsequent analysis of alternative splicing could potentially depend on the informant species. In order to compensate for this dependency, we performed all analyses twice using either *Arabidopsis* or rice as the informant species. Although these species, which diverged from a common ancestor around 150 MYA ago (Rizzon et al., 2006), provide very distant anchor points within the evolutionary history of plants, the main results and conclusions from the parallel analyses were very similar.

HSCs were constructed for a total of 8,008 *Arabidopsis* and 6,781 rice genes. Intron boundary polymorphisms were detected within 2,332 (29%) and 1,946 (29%) of the *Arabidopsis* and rice HSCs, respectively. These HSCs formed the core sets on which all subsequent analyses were performed. A total of 598 and 596 different plant species were at least once represented in the *Arabidopsis* - and rice HSCs, respectively. Around 83% of the *Arabidopsis* and 81% of the rice HSCs contained sequences from at least ten different plant species. Note that a species in this study refers to a unique taxid in the NCBI taxonomy database (Sayers et al., 2009).

The *Arabidopsis* genes within the core set encoded 11,483 introns and 8,954 exons that were fully located within coding sequence (CDS) regions. The rice genes within the core set encoded 9,916 CDS introns and 7,930 CDS exons. Polymorphisms of intron boundaries were identified between the projected boundaries of 3,978 (34,6%) *Arabidopsis* and 3,394 (34,2%) rice CDS introns. We also searched for exon polymorphisms which involve protein regions that are homologous to at least one entire *Arabidopsis* or rice exon. In total, 1,892 *Arabidopsis* (21,1%) and 1,766 rice CDS exons (22,3%) were found to be fully homologous to at least one polymorphic protein region.

Intron boundary- and exon polymorphisms that were detected between two sequences derived from the same gene of the same species were considered to be the result of extant AS events. As we only used EST/cDNA sequences we did not classify the type of the AS events that induced the observed polymorphisms. AS-induced polymorphisms (AIPs) were detected for 215 and 214 different species within the HSCs of 1,818 *Arabidopsis* and 1,529 rice genes, respectively. The number of AIPs that were fully contained within the boundaries of projected introns was roughly three and a half times larger than the number of AIPs that involved protein regions that were homologous to at least one entire *Arabidopsis* or rice exon (Table 1). This ratio is consistent with previous studies that showed that in plants exon skipping events are less common than either intron retention events or events that only modify the boundaries of introns (Campbell et al., 2006; Wang and Brendel, 2006; Kim et al., 2007). It has been suggested that the low level of exon-skipping events is the result of the prevalence of intron rather than exon-definition based recognition of intron boundaries in plants (Barbazuk et al., 2008).

Table 1. Detected AIPs. The table provides an overview of the number of AS-induced polymorphisms (AIPs) that were detected using either *Arabidopsis* or rice as the informant species. Separate numbers are provided for AIPs that were fully located within the projected boundaries of informant introns and AIPs for which the corresponding protein region was homologous to an entire informant exon**.**

Informant	All	Within projected intron boundaries	(%)	Homologous to entire informant exon	(%)
Arabidopsis	3521	2725	77.4	796	22.6
Rice	3049	2359	77.4	690	22.6

Conservation of AIPs

Next we investigated the conservation of AIPs (the first of the three ways of conserving AS isoforms) by searching for pairs of similar AIPs (see material and methods) from different species within the same HSCs. Our search resulted in the identification of 648 and 522 conserved AIP pairs within the HSCs of 198 *Arabidopsis* and 144 rice genes respectively (Table 2). In total 579 (16%) and 449 (14%) of the AIPs detected within *Arabidopsis* and rice HSCs, respectively, were conserved. Interestingly, around 50% (324) of the conserved AIP pairs detected within *Arabidopsis* HSCs were found in only 12 (6%) HSCs. A similar pattern was observed for rice HSCs, where 50% (263) of the conserved AIP pairs were detected within 9 (6%) HSCs.

Table 2. AS isoform conservation. The numbers of identified cases are provided for each of three ways in which AS isoforms can be conserved.

Informant	Conserved AS events [a]		Externalization		Isoform replacement	
	Cases	HSCs	Cases	HSCs	Cases	HSCs
Arabidopsis	648	198	102	68	867	396
Rice	522	144	103	59	807	308

[a] Conserved AS events correspond to conserved AIPs

Conservation of AIPs between *Arabidospis EUV1D* homologs

An example of a gene with more than ten conserved AIP pairs in its corresponding HSC is the *Arabidopsis UBIQUITIN E2 ENZYME VARIANT 1D* (*UEV1D*) gene, which we use here to illustrate a typical pattern of AIP conservation. This gene, which is involved in the DNA damage response pathway is known to produce at least two different transcript isoforms (Wen et al., 2008). The single valine AIP that distinguishes the *Arabidopsis* AS isoforms was found to be conserved in several other members of the Brassicaceae (Figure 2A). Interestingly, a second conserved AIP was found in another group of five species. This group contained species from both the Rosidae- and the Asteraceae clades. The polymorphic protein region that is associated with this conserved AIP is homologous to the entire exon at the 3'-side of the alternatively spliced intron from the Brassicaceae species (Figure 2B).

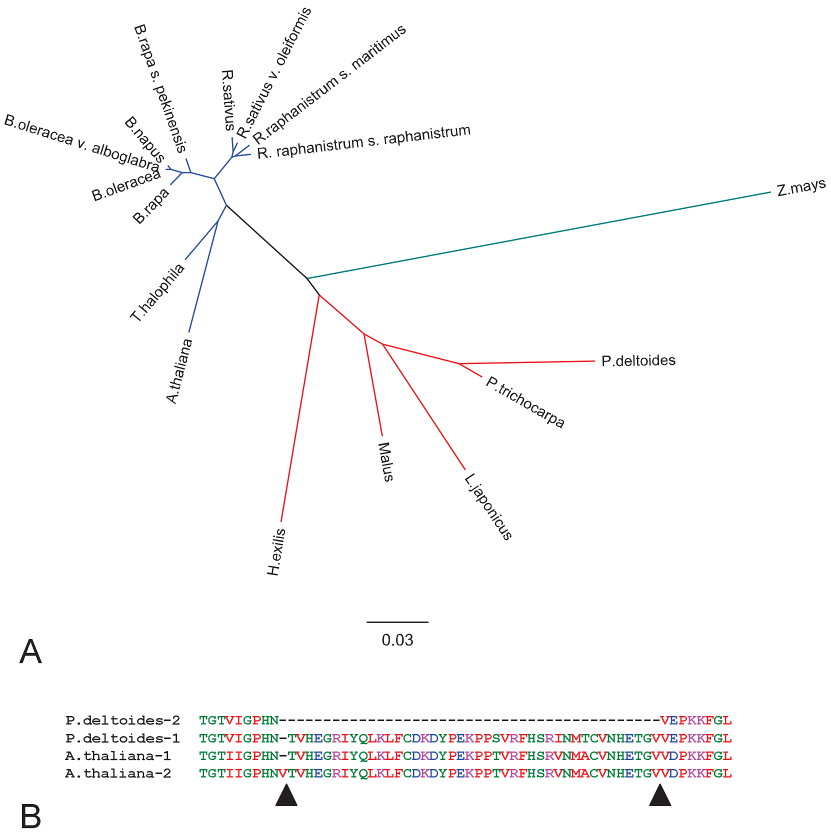
B.rapa s. pekinensis
R.sativus
R.sativus v. oleiformis
R.raphanistrum s. maritimus
R. raphanistrum s. raphanistrum
B.oleracea v. alboglabra
B.napus
B.oleracea
B.rapa
Z.mays
T.halophila
A.thaliana
P.deltoides
P.trichocarpa
Malus
L.japonicus
H.exilis
0.03
A
P.deltoides-2 TGTVIGPHN---VEPKKFGL
P.deltoides-1 TGTVIGPHN-TVHEGRIYQLKLFCDKDYPEKPPSVRFHSRINMTCVNHETGVVEPKKFGL
A.thaliana-1 TGTIIGPHN-TVHEGRIYQLKLFCDKDYPEKPPTVRFHSRVNMACVNHETGVVDPKKFGL
A.thaliana-2 TGTIIGPHNVTVHEGRIYQLKLFCDKDYPEKPPTVRFHSRVNMACVNHETGVVDPKKFGL
B

Figure 2. Conserved AIPs between *Arabidopsis UEV1D* homologs. (A) Phylogenetic relationships between homologs of the *Arabidopsis UEV1D* gene are illustrated using an unrooted Neighbour joining tree. The tree was calculated from an F84-distance matrix using the neighbor program of the phylip package (Felsenstein, 1989). *Z. mays* was included for distance comparison purposes. The blue and red branches correspond to the two different conserved AIPs (see main text). **(B)** Partial multiple sequence alignment of the polymorphic region of the *Arabidopsis* and *P. deltoides* isoforms. The positions of the *Arabidopsis* introns are indicated by the black triangles.

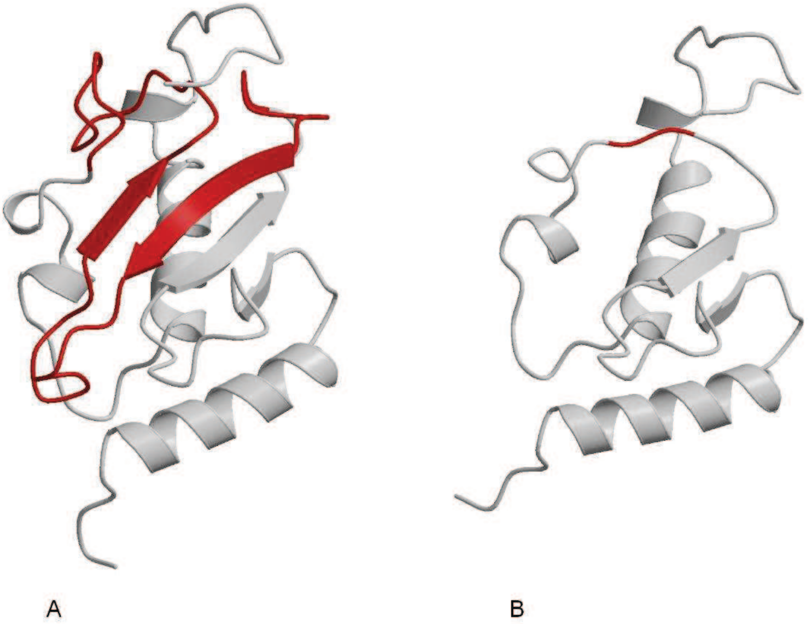

Figure 3. Predicted structures of *Populus deltoides UEV1D* isoforms. The region of the long isoform (**A**) marked in red is deleted in the short isoform (**B**). For illustration purposes, the residues immediately flanking the deleted segment on the short isoform are also highlighted in red.

It has previously been shown that both AS isoforms of the *Arabidopsis UEV1D* gene are able to bind to *UBIQUITIN-CONJUGATING ENZYME UBC13* albeit with different specificities (Wen et al., 2008) . The results of that study suggest that both *Arabidopsis* isoforms are capable of adopting a stable 3D-confirmation. We analyzed whether exclusion of the protein region corresponding to the AIP in the non-Brassicaceae species affects the three dimensional structure of the protein. To this end we modelled the 3D structures of the AS isoforms from *Populus deltoides*. Comparison of the predicted structures (Figure 3) revealed that removal of the polymorphic protein region does not alter the structure of the rest of protein. A possible explanation for this is that this polymorphic protein region encodes two beta strands which are attached to the rest of the protein by flexible loop structures. Moreover, the two connection points (N-terminal and C-terminal to the polymorphic region) are, although far apart in the protein sequence, relatively close to each other in the model for the 3D structure. Although the functional difference between the UEV1D AS isoforms needs to be determined, their conservation and predicted ability to adopt stable 3D confirmations suggests that both isoforms are functional.

Externalization

We identified externalization cases (the second way of conserving newly generated AS isoforms) by searching for intron-boundary or exon polymorphisms between paralogous sequences in one species that corresponded to an AIP in another species. We detected a total of 102 and 103 putative externalization cases in the HSCs of 68 *Arabidopsis* and 59 rice genes (Table 2). The externalization cases in the HSCs of *Arabidopsis* involved 59 different species and those detected in the rice HSCs involved 61 different species. In total 152 (4%) and 116 (4%) of AIPs detected within the HSCs of *Arabidopsis* and rice, respectively, corresponded to an externalization event.

There are several factors that can influence the detection of externalization cases. Given that AS events do not persist long in evolution and that AS patterns quickly diverge after gene-duplication events (Su et al., 2006; Zhang et al., 2010), it is likely that externalization cases will most often only be detectable between very recent paralogous genes. As for most species no genomic sequence was available, it was difficult to determine whether transcripts were derived from very recent paralogs or from allelic variants of the same gene. In addition, it is likely that many AS variants have been missed due to the limited sampling of the transcriptomes of the majority of species. Therefore some externalization cases might actually represent conserved AS events. Future RNA seq studies can either confirm or reject a number of the conversion cases identified in this study.

Despite these limiting factors, a similar albeit weaker clustering pattern over HSCs as the one observed for conserved AIPs was found for the externalization cases. A total of 52 (51%) externalization cases were identified within 18 (26%) of the *Arabidopsis* HSCs. Similarly, 53 (51%) of the externalization cases within rice HSCs were identified within 12 (20%) of the HSCs. Further analysis revealed that 33 (30%) of the externalization cases within *Arabidopsis* HSCs corresponded to AIPs that were also conserved at the event level in two or more species. The same was found to be true for 24 (22%) of the conversion cases detected within the HSCs of rice genes.

Isoform replacement

The third and final way of conserving newly generated AS isoforms, isoform replacement events, were identified as intron boundary- or exon polymorphisms within a homologous sequence cluster that were not associated with an AIPs or externalization. However, isoform replacement events can correspond to an externalization event of which either the corresponding AIP or one of the paralog pairs was not detected. Isoform replacement events can also correspond to AIPs for which one of the corresponding variants was not detected. Finally, an isoform replacement case can correspond to intron-boundary polymorphisms that are the result of mutations which alter the splice sites. In these cases the splice-pattern change is not the result of selection for a beneficial AS isoform. Due to these uncertainties our isoform replacement estimates are likely to represent an upper bound. We identified a total of 867 and 807 putative isoform replacement cases in the HSC of 396 *Arabidopsis* - and 308 rice genes, respectively (Table 2). In total, 434 (50%) of the isoform replacement cases were identified in the HSCs of 66 (17%) *Arabidopsis* genes. For rice 408 (50%) of the isoform replacement cases were identified in 14% of the HSCs.

Persistence of AS events in evolution

Above, we analyzed three different ways by which novel AS isoforms could in principle be conserved and found only a limited number of cases for each of those. Subsequently, as a final analysis, we investigated how long AS events are likely to persist during evolution by comparing the F84 (Felsenstein and Churchill, 1996) nucleotide distances between sequences from homologous genes with conserved AIPs to three reference distance sets. These reference sets contained F84 distances between one of the following type of homologous sequences: (*i*) allelic variants of the same gene; (*ii*) paralogous sequences; (*iii*) all homologs from different species (with or without conserved AIPs). Note that the “all homologs” set was constructed by only keeping the smallest distance between the sequences of each pair of species within an HSC. This analysis gives insight into the evolutionary distance over which AS events are conserved, in comparison with the evolutionary distance between allelic variants, between duplicated genes, and between homologous genes in general. The distances between these reference sets are clearly different (Figure 4) with the shortest distances in the allelic variant sets and the largest in the “all homologs” set.

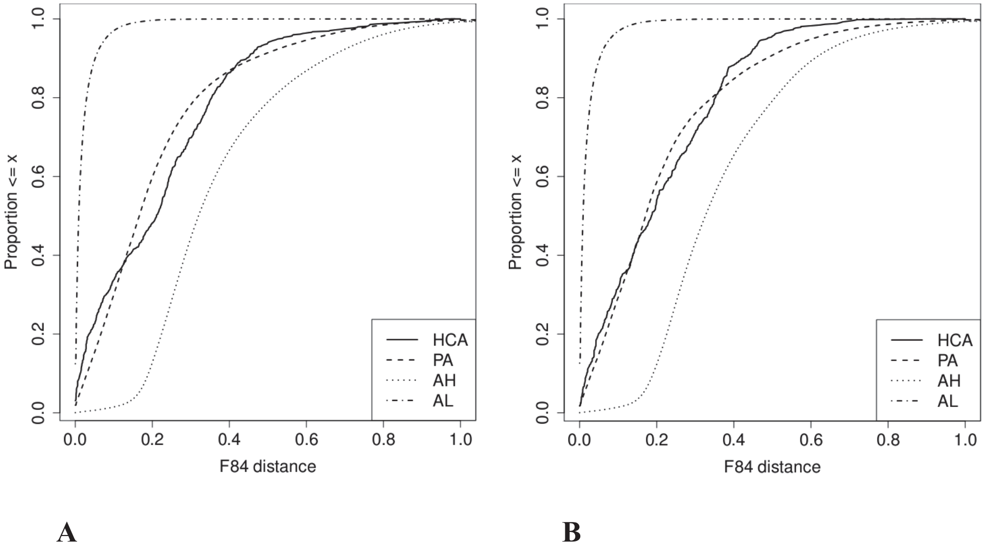

Figure 4. F84 nucleotide distances. Cumulative distributions of F84 distances between different types of homologous sequences within the HCSs of *Arabidopsis* (A) and rice (B). The following homologous sequences were considered: Homologs with conserved AIPs (HCA); Paralogs (PA); all homologs (AH) and Allelic variants (AL).

The distances between homologs with conserved AIPs are smaller than the distances between "all homologs" and larger than the distances between allelic variants (Figure 4). The distribution of distances between homologs with conserved AIPs is significantly different from the distribution of distances between "all homologs" and between allelic variants (Kolmogorov-Smirnov test-test; alternative two sided; all p-values: $2e^{-16}$). Around 30% and 26% of the distances between homologs with conserved AIPs within the HSCs of *Arabidopsis* and rice, respectively, fell within the same range as ~95% of the corresponding distances between allelic variants. This overlap corresponds well with the previous findings that AS events can substantially vary between individuals of the same species (Nembaware et al., 2004; Kwan et al., 2008; Wang et al., 2008).

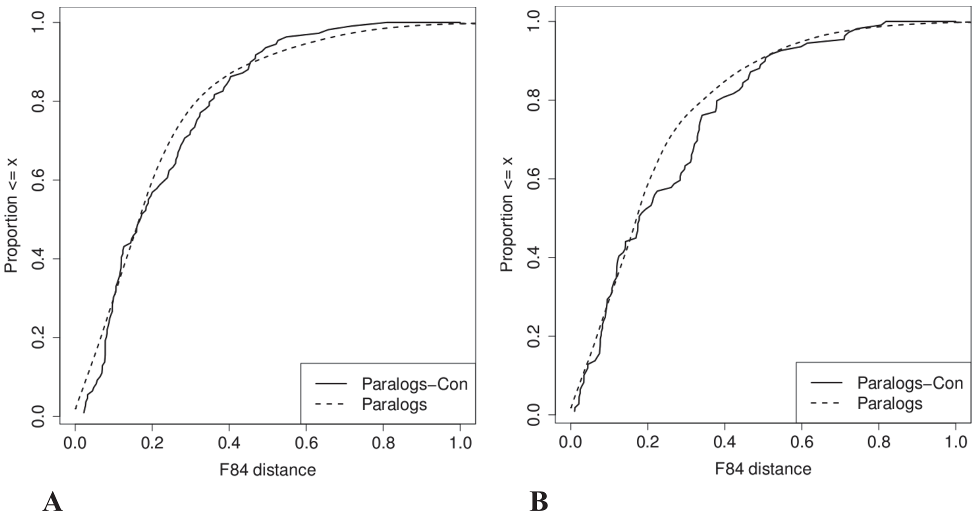

Figure 5. **F84-distances between paralogs.** Cumulative distribution of F84 distances between the different types of homologous sequences in the *Arabidopsis* (A) and rice (B) HSCs. The distributions are provided for the distances between all paralogs ("Paralogs") and between paralogs corresponding to an "internal" to "external" paralog conversion event ("Paralogs-Con").

Importantly, the distances between homologs with conserved AIPs in the HSCs of both *Arabidopsis* and rice genes tend to fall within the same range as the corresponding distances between paralogs (Figure 4). Although the distance distributions for the homologs with conserved AIPs and for the paralogs obtained from the *Arabidopsis* HSCs were significantly different (p-value: 8e^{-10}), the difference between these distributions obtained from the rice HSCs was less pronounced (p-value: 0.01), and for both the difference is clearly much less pronounced than the differences between the homologs with conserved AIPs and the all homologs or allelic variant distributions. The distribution of the distances between pairs of paralogs resulting from putative externalization was for the *Arabidopsis* HSCs not significantly different from the corresponding distribution of all paralogs (Figure 5) (p-value: ~0.16). However, the difference between the same distributions was just significant for rice (p-value 0.01). Hence, our analysis of nucleotide distances demonstrates that homologs with conserved AS events or with AS isoforms conserved via externalization are at similar evolutionary distances from each other as paralog sequences, and the distribution of distances between all homologs or between allelic variants is quite different from that distribution.

Discussion

It has previously been shown that AS events, even between closely related species, are not well-conserved. One explanation for this lack of conservation is that the majority of AS events and isoforms are not functional. However, an alternative hypothesis is that AS isoforms are preferentially retained through ways that do not involve conservation of AS events. In this study we explored this hypothesis by analyzing a large collection of EST/cDNA data.

We specifically consider three different ways in which AS isoforms can be conserved (Figure 1). The first way for conserving AS events involves conservation of the corresponding AS event. More specifically we identified the polymorphism induced by the AS event (AIPs). Around 16% of all identified AIPs were found to be conserved in at least two species. It was also found that rather than being evenly distributed over many HSCs, around half of the conserved AIP pairs were detected within only six percent of the analyzed HSCs. This distribution suggests that in only a few gene families natural selection has shaped AS as a mechanism for the production of multiple functional proteins. The genes involved have retained the AS events as a mechanism for producing multiple distinct proteins.

The second way in which AS isoforms can be conserved involves externalization of AS isoforms over different paralogs. For around 4% of all AIPs the corresponding AS isoforms have been partitioned over two paralogous genes in another species. The limited number of externalization events that we identified is consistent with a previous study in animal species in which it was shown that the differences between paralogous genes are not the same as the differences observed between AS isoforms (Talavera et al., 2007). Also the much weaker clustering of externalization cases compared to conserved AIPs cases indicates that the process has been less common than AS event conservation. Around 20 to 30% of externalization cases involved conserved AIPs. In these cases the corresponding isoforms might be truly important because they have been conserved through different ways.

It has previously been shown that AS frequencies and gene family size are inversely correlated (Kopelman et al., 2005; Su et al., 2006). It has been suggested that subfunctionalization (Zhang, 2003) involving the externalization of AS isoforms might be the cause of this negative correlation (Su et al., 2006). The limited number of externalization cases that we have detected does not support such a scenario in plants. We conclude that AS isoform conservation through externalization is not preferred over retention of AS events.

The final way of conserving AS isoforms involves the replacement of existing AS isoforms by novel ones. The number of HSCs with isoform replacement cases was two times larger than the number of HSCs with conserved AIP pairs. However, clustering was not as strong as observed for conserved AIP pairs. We also found that in the vast majority of HSCs with

isoform replacement cases, a single isoform dominated while the remaing isoforms were only present in a few, often one, species (data not shown). Therefore AS isoform replacement can not have had a strong impact on the evolution of protein diversity in plants.
Given the results of our analyses we conclude that the lack of conserved AS event/AS isoforms is not the result of AS isoforms being preferentially conserved through ways other than retention of the corresponding AS event. Novel functions thus appear to only rarely arise through AS-induced modifications to existing proteins.
It has previously been suggested that the most likely outcome of gene duplication events is the rapid non-functionalization of one of the redundant gene copies (Lynch and Conery, 2000). Although the shapes are different, the distribution of nucleotide distances between homologous sequences with conserved AIPs and between paralogous sequences are very similar. We therefore conclude that the vast majority of AS isoforms, which are not functional, are lost within a time period that is similar to the time in which most redundant gene copies are lost after the gene duplication event.

Material and Methods

Initial datasets
EST/cDNA sequences from the informant species *Arabidopsis thaliana* and *Oryza sativa* (rice) were downloaded from the PlantGDB website (Dong et al., 2004). The genome and predicted proteome of *Arabidopsis thaliana* version TAIR 8.0 were downloaded from www.*Arabidopsis*.org. The genome and predicted proteome of *Oryza sativa* version 5.0 were downloaded from ftp.plantbiology.msu.edu. Gene structures and AS-events were for both *Arabidopsis* and rice predicted using previously described methods (Severing et al., 2009).
The est_others file as of October 2009 was downloaded from ftp.ncbi.nim.nih.gov and filtered by only keeping sequences from plant species other than *Arabidopsis* and rice. A translated sequence set was constructed by determining the longest open reading frame (≥30AA) for each EST/cDNA sequence using the getorf program from the EMBOSS package (Rice et al., 2000) version 5.0.

Homologous sequence clusters
Blastp (Altschul et al., 1997) searches (Low complexity filter: off; e-value threshold: $1e^{-10}$) were performed with the translated sequences against the predicted proteomes of the informant species. Blast hits were only considered if the corresponding hsp met the following criteria: (*i*) ≥ 40% of the hsp positions were identical residue pairs and (*ii*) ≥60% of the informant residues were included in the hsp. Homologous sequence clusters (HSCs) were

constructed by grouping all translated sequences that had a protein product from the same informant gene as their best blast hit.

Classification of homology relationship types

All sequences within an HSC from the same species were clustered using the CAP3-program (-p 95, -o 100) (Huang and Madan, 1999). Each of the resulting CAP3 clusters was considered to represent a single gene. The QualitySNP program (Tang et al., 2006) was used to further dissect each CAP3 cluster into sub groups of sequences that were likely to be derived from the same allelic variant of the corresponding gene. The sequences within each allelic variant cluster were assembled into contigs using CAP3. Local alignments (Smith and Waterman, 1981) were constructed between the longest open reading frame encoded on each contig and the corresponding informant protein. A contig was discarded if less than 80% of the informant protein residues were included in the alignment.

Intron position projection

In this study we rely on the assumption that the positions of introns within coding regions of genes are highly conserved (Knowles and McLysaght, 2006; Lin et al., 2006; Roy and Penny, 2007). Each intron within the corresponding CDS region of an informant protein was represented by two boundaries. These boundaries corresponded to the 5' and 3' amino acid residues immediately flanking the position of the intron. In the case that the intron interrupted a codon, the two amino acids immediately flanking the interrupted codon were taken as the boundaries. Each intron was projected onto a translated contig by first creating a global alignment between the translated sequence and the informant protein and then searching for the amino acids that were homologous to the amino acids representing the boundaries of the informant intron. It was required that out of the ten alignment positions immediately flanking a projected intron, at most one was a gapped position and at least four were pairs of identical residues.

Detection of AS events and AS conservation

Intron boundary polymorphisms were detected by searching for pairs of sequences that had a different number of amino acids located between the projected boundaries of the same informant intron. Intron boundary polymorphisms were only considered if one of the corresponding sequences could be aligned to the informant protein without any insertions or deletions around the corresponding intron position. As a result of this requirement, intron boundary polymorphisms could be represented by a single amino acid sequence (i.e. an insertion in the second of the corresponding sequences). Exon polymorphisms were identified as variable protein regions that were homologous to at least one entire informant exon.

Intron boundary- and exon polymorphisms that distinguish two sequences from the same gene within a species were considered to be the result of present day AS events. These polymorphisms are hereafter called AS induced polymorphisms (AIPs). AIPs were considered to be conserved if 40% of the positions from a global alignment between the corresponding proteins sequences were pairs of identical residues. Externalized AS isoforms were identified as paralogous sequences that differed by a similar polymorphism as two AS isoforms in another species (the same rule as for conserved AIPs).

Three dimensional protein structure predictions

The structures for the *P.deltoides* UEV1D isoforms were predicted with Modeller 8 (Sali and Blundell, 1993) using the crystal structure of human UEV1 in a ubiquitin conjugating enzyme complex (Moraes et al., 2001) (hMms2, PDB identifier 1j7d) as the template. The corresponding protein sequence shares 49% identical residues (69% similar residues) with the long *P. deltoides* isoform. For both *P.deltoides* isoforms, the best out of 1000 generated models was selected based on the objective score as calculated by Modeller. Inspection of the selected structures using Procheck (Laskowski et al., 1996) revealed that >90% of the residues were located within the core-region. These percentages indicate that the predicted structures were of sufficient quality.

Nucleotide distances

Multiple protein sequence alignments were constructed for each HSC using Muscle (Edgar, 2004) version 3.7. Each alignment was trimmed to the column containing the leftmost N-terminal residue from one of the informant proteins. All residues corresponding to intron boundary- or exon polymorphisms were excluded from the alignments. Nucleotide alignments were constructed from the protein alignments by replacing each amino-acid with its corresponding codon triplet and by extending each gapped position to three positions. F84-distances (Felsenstein and Churchill, 1996) were calculated for all pairs of aligned sequences using the dnadist program from the phylip (Felsenstein, 1989) package version 3.69. Because many genes in this study were represented by several allelic variants, the distance between two genes corresponded to the smallest of all pairwise distances between their allelic variants. Cumulative distributions of the F84 distances were constructed and compared using the ECDF- and ks.test (Kolmogorov-Smirnov test) functions in R (R Development Core Team, 2009), respectively.

Chapter 5

The Impact of Alternative Splicing on Plant MADS Domain Protein Function

Edouard I. Severing, Aalt D.J. van Dijk, Giuseppa Morabito, Jacqueline Busscher-Lange, Richard G.H. Immink
and
Roeland C.H.J. van Ham

ABSTRACT

Several genome-wide studies have demonstrated that alternative splicing (AS) significantly increases the transcriptome complexity in plants. However, the impact of AS on the functional diversity of proteins is difficult to assess using genome-wide approaches. The availability of detailed sequence annotations for many genes and gene families allows for a more specific assessment of the potential effect of AS on their function. One example is the MADS-box gene family, members of which interact to form protein complexes that function in transcription regulation. Here, we analyze the impact of AS on the protein-protein interaction capabilities of MIKC-type MADS-domain proteins. AS events resulting in putatively translated transcripts were considered functional if one of the following criteria was met: 1) they overlapped with predicted interaction motifs; 2) they were located in regions involved in multimeric complex formation, or 3) they had an effect on the protein sequence that was conserved in different species. Nine out of twelve MIKC genes predicted to produce multiple protein isoforms harbored putative functional AS events according to those criteria. All events with conserved effects were located at the borders of or within the K-box domain. We illustrate how AS can contribute to the evolution of interaction networks through an example of selective inclusion of a recently evolved interaction motif in the *MAF1-3* subclade. We confirmed expression of many of the isoforms at the RNA level and

were able to give hints about the potential impact of many of the events on the interaction capabilities of the encoded MIKC proteins.

INTRODUCTION

Alternative splicing (AS) is a frequent phenomenon in higher eukaryotes and involves the production of multiple distinct transcript isoforms from a single gene. Although it is well established that AS substantially increases transcriptome complexity, the extent to which the process has functional implications at the proteome level remains relatively unknown.

Several genome-wide studies have addressed this issue by determining the prevalence of AS events that are likely to be functional according to predefined criteria such as conservation (Severing et al., 2009) or the predicted effect on protein structure (Taneri et al., 2004; Tress et al., 2007; Melamud and Moult, 2009). Other genome-wide studies have focused on the identification of more general patterns that relate AS to gene or domain functions (Liu and Altman, 2003; Irimia et al., 2007). A number of interesting patterns has been unveiled, but by their design, these studies will identify only those aspects that are general enough to be present in large numbers of proteins. However, each gene and gene family has its own evolutionary history and can be affected by AS in specific ways that are not described by the globally observed patterns. The way in which a gene is affected by AS depends for instance on the size of the family it belongs to (Kopelman et al., 2005) or on specific genomic rearrangements such as tandem exon duplications that have occurred in the gene's evolutionary history (Kondrashov and Koonin, 2001; Letunic et al., 2002). Hence, in order to fully appreciate the functional impact of AS, it is important to study the process at the level of individual genes or gene families.

One of the best studied gene families in plants is the MADS-box transcription factor family. Members of this family are involved in a number of developmental processes (Becker and Theißen, 2003) but they are probably best known for their role in regulating the onset and patterning of flowering (Causier et al., 2010). MADS-box genes can be divided into two main groups: the type I and type II or MIKC genes (Alvarez-Buylla et al., 2000; Parenicova et al., 2003). While little is known about the former group, a wealth of information is available for the latter. MIKC proteins exert their function mainly in the form of di- or multimeric protein complexes (Immink et al., 2010). The availability of a comprehensive yeast two-hybrid interaction map for *Arabidopsis thaliana* MADS-domain proteins (de Folter et al., 2005) as well as an extensive yeast three-hybrid screen (Immink et al., 2009) illustrates the large diversity of complexes that are formed between members of this family.

The sequence of MIKC proteins can be divided into four global regions with specific functions (Riechmann, 1997; Kaufmann et al., 2005). The MADS (M) domain has a DNA-binding function and, together with the intervening (I) domain, is involved in determining the specificity of protein dimerization. The dimeric protein-protein interaction is promoted by the Keratin-like (K) domain. The C-terminal part of the K-domain is also involved in the formation of higher-order complexes. The C-terminal (C) domain is involved in transcriptional activation and the formation of higher-order complexes. Recently, we have developed computational methods that aid in the identification and understanding of sequence features that are important determinants of the interaction specificity of individual MIKC proteins (van Dijk et al., 2008; van Dijk et al., 2010).

One of the ways through which AS can influence the formation of di- or multimeric complexes is by regulating the availability of individual MIKC proteins. This can be achieved by the increased production of transcripts that are targets for the Nonsense Mediated Decay (NMD) pathway (Lewis et al., 2003). Alternatively, AS can influence the interaction specificity of the encoded proteins by disrupting or introducing individual interaction sites. This has for instance previously been shown for the two AS variants from the *Arabidopsis B-sister (ABS, AGL32)* gene which encode proteins with different higher-order interaction specificities and function (Nesi et al., 2002; Kaufmann et al., 2005; Folter et al., 2006; Immink et al., 2009).

Although, several individual cases of AS events have been reported for members of the MADS-box gene family (e.g. Montag et al., 1995; Kyozuka et al., 1997; Hartmann et al., 2000; Kitahara and Matsumoto, 2000; Scortecci et al., 2001; Nesi et al., 2002; Ratcliffe et al., 2003; Caicedo et al., 2004; Kim et al., 2005; Balasubramanian et al., 2006; Folter et al., 2006; Lightfoot et al., 2008), a systematic analysis of the functional impact of AS within this family is still lacking. In this study we investigate the potential role of AS in the formation of protein complexes between MIKC proteins. We analyzed the impact of AS on both the MIKC-type protein group as a whole and on a selection of members individually. Three independent criteria were used to postulate a functional implication of an AS event. The first criterion required AS to overlap with predicted interaction motifs that can be considered to correspond to those regions that form the contact surface, using our recently developed interaction motif prediction method (van Dijk et al., 2008; van Dijk et al., 2010). The second criterion required overlap with the region known to be important for higher-order complex formation. In contrast to these criteria, which focus on the nature of the function, the third criterion considered conservation of an AS-induced polymorphism (AIP) at the protein level among different species and only indicates a potential function. By using the functional annotation of sequences together with available experimental data, we provide hints as to which specific aspects of protein complex formation are likely to be affected by the individual AS events. In addition to providing interesting candidates for further experiments, our results illustrate the potential of computational analysis of AS within the functional context of individual genes or gene families.

RESULTS

General data description

In total, 80 of the 111 annotated MADS-box genes (including MADS-box like) in *Arabidopsis* had at least one annotated protein for which the corresponding open reading frame was fully supported by transcript evidence and were considered further. Of the 38 loci for which at least one encoded protein had a clearly identifiable MIKC structure, 17 had evidence for AS. Of all considered loci, 58 had at least one annotated protein product with experimentally identified protein-protein interactions.

Domain function and alternative splicing

We investigated the prevalence of AS within the individual domains of MIKC proteins. The fractions of intron clusters (overlapping introns were clustered) that were involved in AS events were 24%, 11% and 9.1% for the I-, K-box- and C-terminal domains respectively. Fisher's exact tests indicated that none of these fractions was significantly different from their expected values (best P-value: 0.03 for the I-domain). Nevertheless, intron clusters in the I-domain were twice as frequently involved in AS events than those located in the K-box or C-terminus domain. Given that the K-box domain is a PFAM domain and the I-domain is not, the lower fraction of AS within the K-box domain is consistent with our previous observation that intron clusters within predicted PFAM-protein domains are less frequently involved in AS events than intron clusters located outside such predicted domains (Severing et al., 2009). We did not find any AS event that affected the MADS domain. This is not surprising because there is only one *Arabidopsis* MADS box gene (At1g33070) that has introns in its MADS domain (Kofuji et al., 2003). We did not find transcript (EST) support for this gene.

Identification of functional AS events

AS can result in transcripts that encode truncated protein due to the presence of a premature termination codon (PTC). These PTC containing transcripts can be recognized by the cell and degraded via the NMD pathway (Lewis et al., 2003). Because we were only interested in the function of AS events that are manifested at the protein level, all events resulting in putative NMD targets were discarded. Transcripts were regarded to be potential targets for the NMD pathway if they encoded a pre-mature stop codon that was located more than 50-55 nt upstream of the last exon/exon junction (Hori and Watanabe, 2007). In a number of cases it was difficult to determine whether the transcript was likely to be degraded or translated from a downstream ATG codon due to the position of the PTC. After removal of these ambiguous cases, a total of twelve loci, accounting for 13 AS events, remained and were predicted to produce multiple protein isoforms.

We analyzed the functional impact of AS by investigating the effect of AS on predicted interaction motifs and on regions responsible for higher-order complex formation. The interaction motifs were predicted using our recently developed IMSS method (van Dijk et al., 2008; van Dijk et al., 2010). Finally, we also investigated the conservation of the AS-induced polymorphisms across a wide evolutionary range of plant species.

Ten out of the thirteen candidate events (nine of the twelve loci) were considered functional because they met at least one of the above criteria. All of these events were located in the I- or K-region of the proteins (Figure 1).

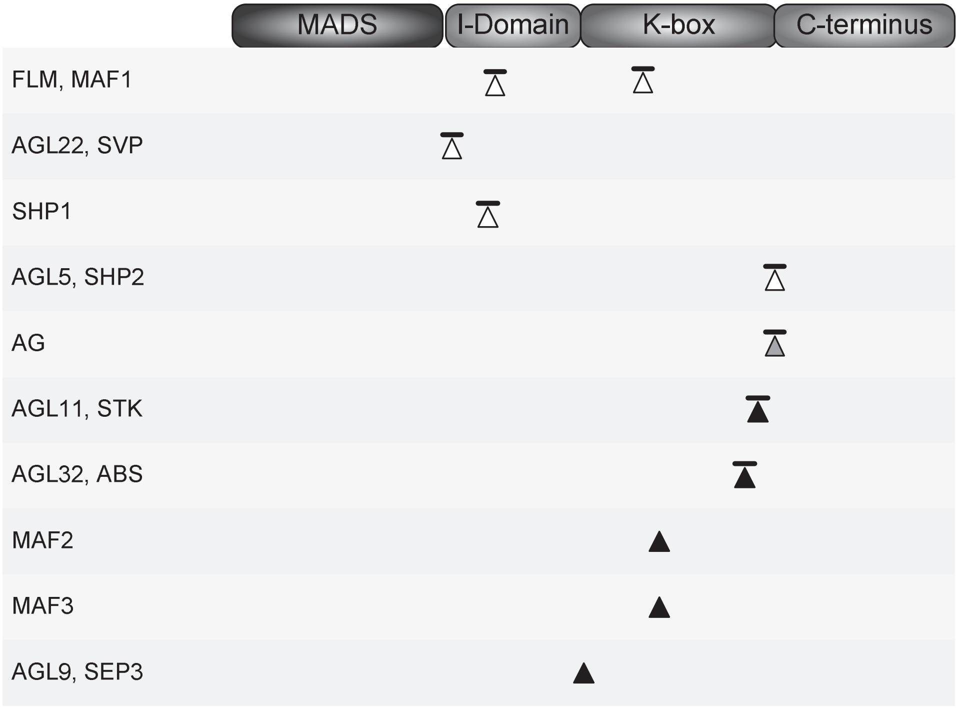

Figure 1. Putatively functional Alternative Splicing (AS) events in the *Arabidopsis* MIKC-type MADS-box family. Triangles indicate the position of AS events in relation to protein-domain architectures of MIKC proteins. Bars on top of the triangles indicate that the event overlapped with at least one computationally predicted interaction motif. The color of a triangle indicates whether the AS induced polymorphism corresponding to the event was only identified in *Arabidopsis* (white), was conserved between different species but was not observed in *Arabidopsis* (grey) or, was conserved between *Arabidopsis* and other species (black).

Seven of the functional AS events (six loci) were found to overlap with predicted interaction motifs (Figure 1: horizontal bars; Supplementary Table S1). Three of these events overlapped with interaction motifs located in the I-domain, which was previously shown to be a hotspot for determination of interaction specificity (van Dijk et al., 2010).

Experimental validation of predicted AS events

We performed two experimental analyses in order to confirm the occurrence of the AS events. First, we performed qRT-PCR on RNA samples extracted from *Arabidopsis* leaves and carpels in order to confirm the expression of multiple transcript isoforms that differ as a result of the AS events. Almost all predicted functional AS-events could be confirmed at mRNA expression level by this analysis and in all these cases sequencing of the amplified fragment revealed amplification of the expected isoform. Second, we analyzed RNAseq data from Jiao and Meyerowitz (Jiao and Meyerowitz, 2010), which represents fragments of RNA molecules that were associated with ribosomes. Hence this data set can provide important clues as to whether the functional AS events are manifested at the proteome level. Using this dataset, we were able to confirm the AS events from *FLM/MAF1*, *SHP2* and *SEP3*. Overall, for two cases (SHP2 and SEP3) we found additional confirmation in both datasets, and for three cases (STK, ABS, FLM/MAF1) we found additional confirmation in one dataset. The lack of confirmation for *MAF2* and *MAF3* in RNAseq data might be explained by the relative low expression level of these two genes in young flowers (Schmid et al., 2005), which is the material sampled for the translatome analysis(Jiao and Meyerowitz, 2010).

Predicted impact of AS on protein interactions

The IMSS predictor was used next to determine whether the isoforms of those loci with AS events that overlapped with protein-protein interaction motifs within the I-domains might have different dimer interaction specificities. Note, that the predictor has been trained to only predict dimer and not higher-order interactions. Only the isoforms of the *SHORT VEGETATIVE PHASE (SVP)* locus, which differ by the presence of a sequence containing a single interaction motif, were predicted to have different interaction specificities. The short SVP isoform (previously named SVP3 or SVP1 EFCSSS56-61D; (van Dijk et al., 2010)) lacking the motif was predicted to have five dimerization partners while the long SVP isoform (previously named SVP1; (van Dijk et al., 2010)) was predicted to have a total of 26 interaction partners. The large loss of predicted interaction partners for the short isoform corresponds well with the experimental yeast-based interaction studies (Figure 2A) (van Dijk et al., 2010). In order to further substantiate putative differences in the biological role of those SVP variants, we performed overexpression studies with *SVP1* and *SVP3*, showing as observed previously that ectopic expression of *SVP1* results in a late flowering phenotype and floral abnormalities (Figure 2B-D) (Masiero et al., 2004). The strength of the floral abnormalities was linked to the strength of the floral repression in the segregating population. In contrast, ectopic expression of *SVP3* from the same *CaMV35S* promoter had no significant effect on flowering time (Figure 2B) and flowers developed without obvious modifications. Nevertheless, RT-PCR experiments confirmed ectopic expression of the *SVP3* transcript in the selected lines. Additional primary transformants were selected, but none showed obvious flowering time effects or floral abnormalities. The observed differences in effects on flowering-time upon ectopic expression of *SVP1* and *SVP3* are in line with the expectations based on protein-protein interactions: only SVP1 is able to interact with the strong repressor of flowering FLOWERING LOCUS C (FLC) (Van Dijk et al, 2010) and previously it was shown that SVP and FLC function is mutually dependent (Li et al., 2008). The same holds for the differences in observed floral defects, which probably are at least partially caused by direct interactions between SVP1 and the ABC-class MADS domain proteins, a capacity that is lacking by SVP3 (van Dijk et al., 2010).

Four AS events (four loci) were found to overlap with interaction motifs located within the K-box domain. Of these, only the AS event of the *MAF1* locus was located within the N-terminal region of the K-box which has been demonstrated to mediate dimer interactions (Yang et al., 2003). This AS event involves the retention of an intron which leads to the introduction of an interaction motif without disruption of the downstream protein sequence. The remaining three events were located in the C-terminal region of the K-box domain which has been shown to be important for higher-order interactions (Immink et al., 2009). The isoforms encoded by the *ABS* locus have experimentally been shown to form different higher-order complexes (Immink et al., 2009). Strikingly, the positions of the AS-induced variation in *SEEDSTICK* (STK) and *SHATTERPROOF2* (SHP2) are quite similar to that in ABS, which leads us to predict that these also impact higher-order complex formation (see further discussion below). In addition, according to our IMSS predictions, ABS should not have different dimer interaction specificities for its two protein isoforms, and this indeed has been shown experimentally (Folter et al., 2006).

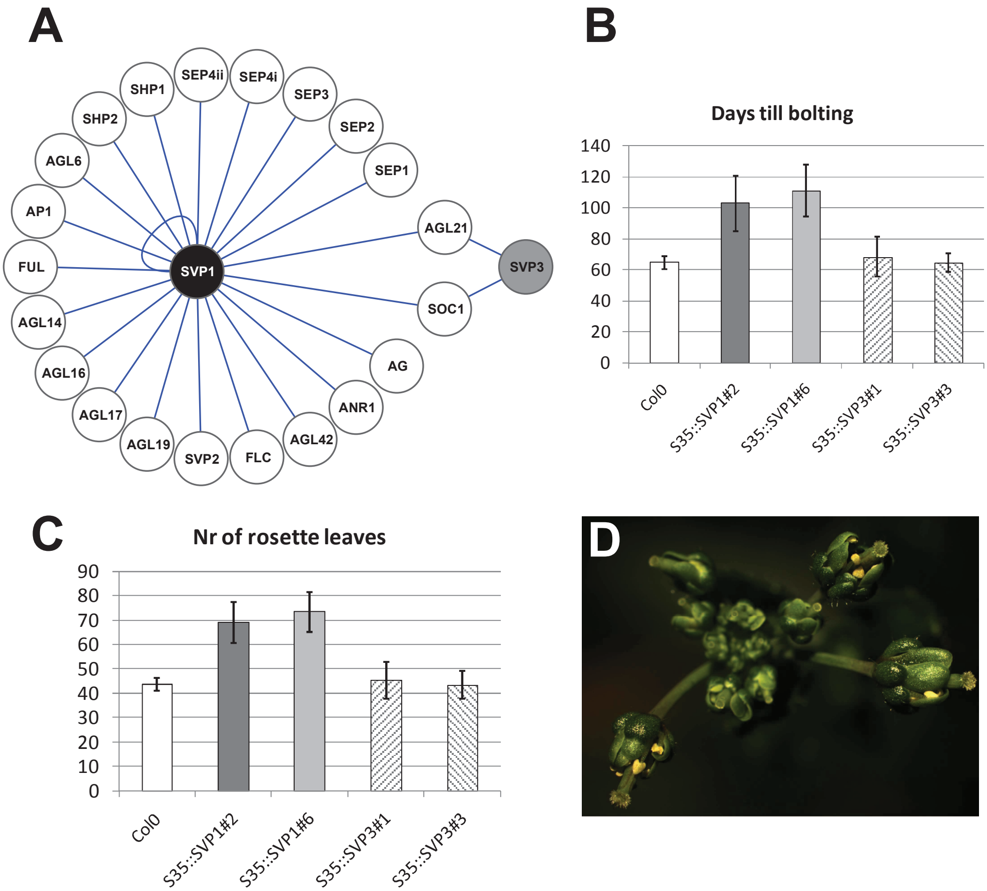

Figure 2. Functional analysis of two *SHORT VEGETATIVE PHASE (SVP)* isoforms. A) Protein-protein interaction capacity of the SVP1 (Black) and SVP3 (Grey) splicing variants as determined by matrix-based yeast two-hybrid studies (Van Dijk et al, 2010). **B) and C)** Effect of ectopic expression of *SVP1* (CZN094) and *SVP3* (CZN756) on flowering-time under short day conditions. For both constructs two segregating lines were analyzed. **D)** Floral phenotypes upon ectopic expression of *SVP1*. First and second whorl organs are greenish and leaf-like and flowers are partially sterile due to reduced anther filament elongation.

Conservation of AS events

Next, we investigated for how many AS events the corresponding AIP was conserved in other plant species. Surprisingly, conserved AIPs were found for a total of six *Arabidopsis* MIKC loci (Figure 1, grey and black triangles), which is a rather high number given the limited numbers of conserved events consistently found in previous genome-wide analyses

(Wang and Brendel, 2006; Baek et al., 2008; Wang et al., 2008; Severing et al., 2009). The single-residue AIP located at the border of the I- and K-box domain of *SEPALLATA3 (SEP3)* was detected in the closely related Brassicaceae species *Brassica napus* and *Raphanus sativus* (Supplementary Figure S1). An additional AIP at this position that involved two amino acid residues was conserved between the closely related Fabaceae species *Glycine max* and *Cyamopsis tetragonoloba*.

The recently duplicated *Arabidopsis MAF2* and *MAF3* paralogs have a conserved alternative acceptor event within the N-terminal region of the K-box domain. The AIPs corresponding to these events were almost exactly conserved in both *B. napus* and *Brassica rapa* (Supplementary Figure S2).

Three additional conserved AIPs were found that, in contrast to those of *SEP3* and *MAF2-3*, both overlapped with interaction motifs and were located within the C-terminal region of the K-box domain (Figure 1). First, the AIP corresponding to the exon-skipping event of the *Arabidopsis STK* locus was also detected in the distantly related Asteraceae species *Taraxacum officinale* (Supplementary Figure S3). Most likely, this represents a case of convergent evolution.

Second, two different conserved AIPs were found between homologs of the *Arabidopsis ABS* locus. An AIP of the exact same size as the AIP of the *Arabidopsis ABS* locus was identified in *B. napus* (Supplementary Figure S4). The second AIP was conserved between the species *Ricinus communis* and *Gossypium hirsutum* that belong to the Malvales and Balanopales orders, respectively. The stretch of residues involved in these AIPs is homologous to the entire exon that is affected by the AS event in the *Arabidopsis ABS* locus.

Third, the *Arabidopsis AGAMOUS (AG), SHATTERPROOF1 (SHP1),* and *SHATTERPROOF2 (SHP2)* are paralogs that originated from a common ancestral gene through two independent duplication events. The first gave rise to the *AG* and *SHP* ancestor lineages and the second led to the *SHP1* and *SHP2* paralogs (Causier et al., 2005). Although no AS events were detected for the *Arabidopsis AG* locus, conserved AIPs were identified in *AG* homologs from seven species (Figure 3). The conserved AIP within the Asteraceae subclass involved a single Q-residue whereas the conserved AIP within the Brassicaceae family involved three residues. The intron position corresponding to the two-residue AIP in the *SHP2* gene is orthologous to the intron position corresponding to the conserved AIP in *AG* in non-*Arabidopsis* species. Both the *SHP2* and *AG* loci encode an interaction motif that spans this intron position. In a recent study it was demonstrated that insertion of a single amino-acid into the predicted motif in the AG protein eliminates its ability to induce female organ development in the first whorl (Airoldi et al., 2010). In addition, it was shown that similar phenotypic differences are induced by the presence or absence of a single glutamine residue within a sequence region in the FARINELLI (FAR) protein of *Antirrhinum majus* that is homologous to the motif in the AG protein. The motif in the AG protein is almost exactly conserved in the homologous sequences from the species with the conserved AIPs (data not shown).

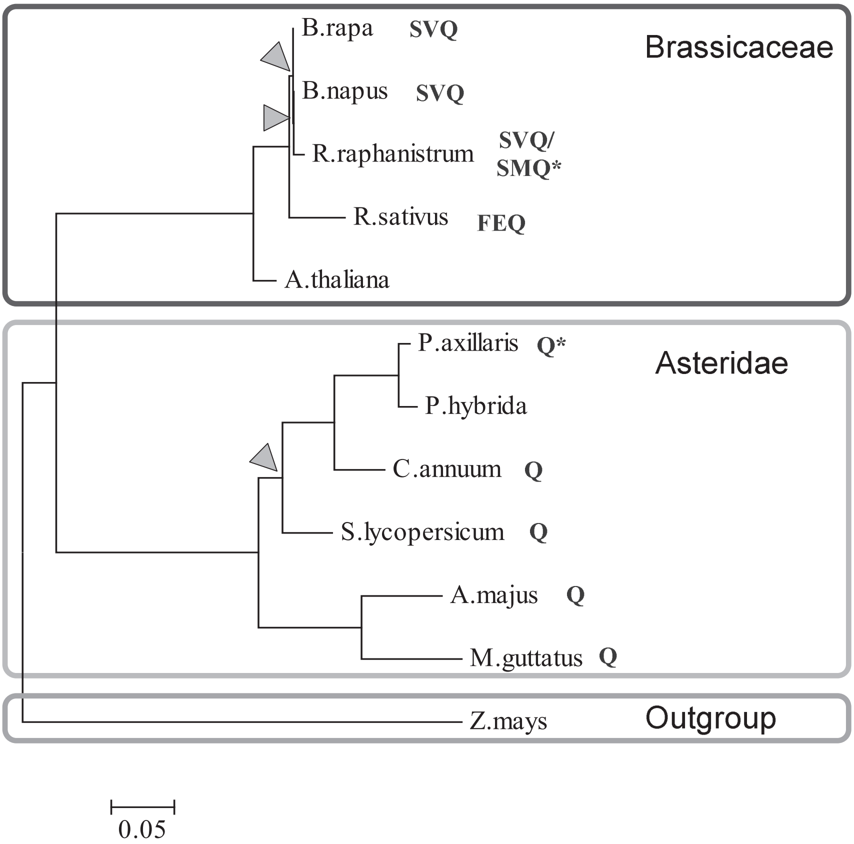

Figure 3. Conservation of AS induced polymorphisms (AIPs) between homologs of the *Arabidopsis AGAMOUS (AG)* protein. A neighbor-joining tree illustrates the phylogenetic relationship between homologs of the *Arabidopsis AG* protein. Nodes with less than 70% bootstrap support are indicated with grey triangles. Except for *Zea mays*, which was included as outgroup species, the taxa within the tree are either members of the *Brassicaceae* subfamily or the *Asteridae* class. Residues behind taxon names correspond to the AIP sequence segment and an asterisk indicates that only the inserted residue(s) were found. No insertions were found in *Arabidopsis* and in *P.hybrida*. The distinct insertions found in *R.raphanistrum* might be the result of different allelic variants or sequencing errors. Note that the sequences from the taxa within the *Asteridae* are homologs of the *FARINELLI (FAR)* gene in *Antirrhinum majus*. The full names of the species used in this tree are provided in the Material and Methods section.

The alternatively spliced exon of the *Arabidopsis STK*-gene encodes three consecutive interaction motifs, two of which span the introns flanking the exon. The motif spanning the 5'- intron is homologous to the motif that is affected by the AS event of the *ABS*-locus. On the other hand, the motif that spans the 3'- intron is homologous to the motif that overlaps with the AIP of *SHP2*. The latter motif is also homologous to the conserved motifs that overlap with the AIPs in the non-*Arabidopsis* species (Figure 4).

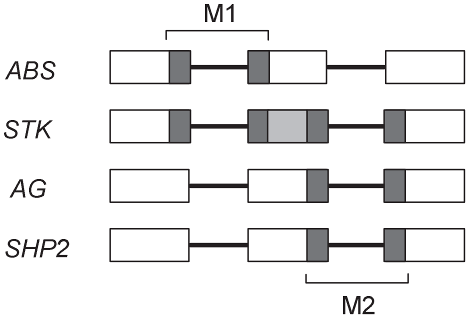

Figure 4. Homologous interaction motifs overlapping with AIPs. Conserved intron positions of the *Arabidopsis ABS*, *STK*, *AG* and *SHP2* genes are represented by horizontal lines and their flanking exons by rectangles. Grey colored regions correspond to predicted interaction motifs that overlap with AIPs. The light grey motif only overlaps the AIP in *STK*. **M1**: Homologous interaction motifs that overlap with AIPs in *ABS* and *STK*. **M2**: Homologous interaction motifs that overlap with AIPs in *STK*, *AG*, *SHP-2*.

AS in the evolution of the FLOWERING LOCUS C (FLC)-clade

As mentioned above, the mutually exclusive exon event of the *MAF1* locus results in the production of protein isoforms with different interaction motif architectures in the I-domain (Figure 5). The proteins encoded by the transcripts containing the 5'-exon have one interaction motif more than those encoded by transcripts that include the 3'-exon. Both exons encode for amino acid residues at their 3'-site that combine with the first residues of the downstream constitutively spliced exon to form interaction motifs. Hence, the mutually exclusive exon event provides a "switch" between two interaction motifs. Only the 5'-exon was detected in transcripts from other *FLC*-like clade members (data not shown). A BlastN search (Altschul et al., 1997) with the 3'-exon (cryptic exon) against the genome of *Arabidopsis* revealed the presence of highly similar sequences within the second introns of the *MAF2* and *MAF3* loci (Figure Supplementary Figure S5A). In fact the nucleotide identity is high over the entire length of the alignment between introns.

The intronic region of the *MAF2* locus that is homologous to the cryptic exon of *MAF1* can fully be translated in the same frame as the *MAF1* exon (Supplementary Figure S5B). However, the nearest acceptor site is located 2 nt upstream of the 5'-site of the region corresponding to the cryptic exon. Usage of this acceptor site would result in a frame shift followed by a premature termination codon. Inspection of the translated intronic region of *MAF3* revealed the presence of an in-frame termination codon. Hence, this region cannot fully be incorporated into a protein sequence.

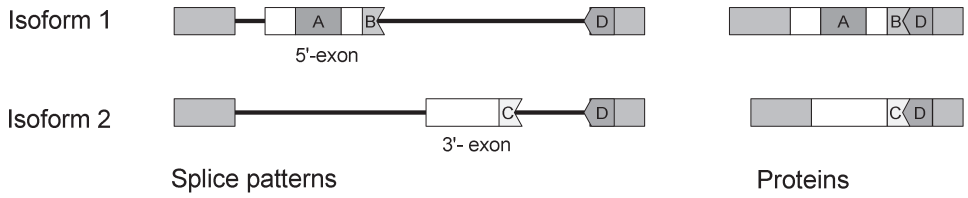

Figure 5. Interaction motif architecture of *MADS AFFECTING FLOWERING1 (MAF1)* isoforms. Two *MADS AFFECTING FLOWERING1* (*MAF1*) isoforms differ as the result of the selective inclusion of either one of the two mutually exclusive exons (named 5'- and 3'- exon). Only the 5'-exon of the mutually exclusive pair that is included in isoform 1 contains motif **A**. Both the 5'- and 3'-exons of the mutually exclusive pair have residues at their 3'-boundary that can form interaction motifs (**BD** or **CD**) together with the first residues of the downstream constitutively spliced exon. Introns are indicated by horizontal lines.

DISCUSSION

In this study we performed a systematic analysis of AS within the MIKC-type MADS-box transcription factor family in *Arabidopsis*. Our study was focused on the functional impact of AS on different aspects of the formation of protein complexes. We restricted our analysis to transcripts likely to be translated, excluding those AS events that are likely to result in transcripts that are degraded by the NMD-pathway. Nevertheless, we cannot exclude potential functions of these transcripts, which would only further add to the importance of AS in the MADS domain family.

Although AS events were not significantly overrepresented in any of the domains, intron clusters within the K-box and C-terminal domain were twice less frequently involved in AS events than intron clusters located in the I-domain. Although, the I-domain is not a predicted PFAM domain, it has a known function in determining dimerization specificity (Immink et al., 2010; van Dijk et al., 2010) .

AS events that overlapped with interaction motifs within the I-domain were shown to potentially affect dimer interaction specificities. Indeed, of the three loci with AS events that overlapped with interaction motifs within the I-domain, the *SVP* locus was predicted to have isoforms with different interactions specificities. The differences between the *SVP* locus and the two other loci with AS events in the I-domain are the locations of the affected interaction motifs. Further experiments are needed to determine whether these other two AS events affect interaction specificities. For SVP however, ectopic expression demonstrated clear differences in impact between the two AS isoforms, where SVP1 resulted in a late flowering phenotype and floral abnormalities whereas SVP3 did not.

It has been demonstrated that AS can be associated with fast evolving regions of proteins (Xing and Lee, 2005). The elevated rate of evolution of the less well conserved I-domain (De Bodt et al., 2003) is not only evident from the amino acid substitutions in this region but also from the repeated changes that have occurred to its underlying gene structure (Kaufmann et al., 2005). The increased rates of AS in this region represent a further mechanism to diversification and the overall high variability of the I-domain may reflect its role as a determinant of dimer interaction specificity.

The potential effect of an AS event in the K-box depends on the location of the event within this domain. In accordance with previously published data (Yang et al., 2003) we hypothesize that the interaction motif that is introduced through an intron retention event located in the sequence encoding for the N-terminal region of the MAF1 K-box domain has a potential effect on dimer formation. AS events that affected the C-terminal region of the K-box were considered to influence multimer formation, because this region was shown to be important for higher-order complex formation (Immink et al., 2009). Based on the rather similar positions of their AS events, we hypothesize that the variants encoded by the *STK*- and *SHP2*- loci form different higher-order complexes, similar to what has been shown experimentally for the ABS isoforms (Immink et al., 2009).

All AIPs that showed conservation across species were located at the boundary of, or within the K-box region and can roughly be divided into two groups. The first group consists of the AIPs from the *Arabidopsis SEP3*-, *MAF2* and *MAF3*- loci that were located near the sequence encoding the N-terminal boundary of the K-box domain. Although, none of these AIPs involved residues corresponding to interaction motifs, the positions of these AIPs suggest that they might influence dimer formation (Immink et al., 2010). The second group contains conserved AIPs that overlapped with predicted interaction motifs located near the C-terminal region of the K-box domain. The motifs involved in these AIPs were all encoded in orthologous exons. The motifs spanning 5'- and 3'- introns of the skipped exon of the *STK* gene are homologous to the motifs that overlap with the AIPs of *ABS* and both *SHP2* and *AG* (in the non-*Arabidopsis* species), respectively.

Previous comparative analyses in plants have shown that AS events are generally not well conserved (Wang and Brendel, 2006; Baek et al., 2008; Wang et al., 2008). Conservation is even less pronounced when the effect of the event on the protein sequence is also taken into account (Severing et al., 2009), and not only the location of the event. It is therefore remarkable that conserved AIPs were found for six members of the same gene family. In contrast to these previous large scale studies, our present conservation analysis was not limited to a few species with large fractions of their genome sequenced. Instead, we searched for conserved AIPs in a large collection of EST/cDNA sequences from a wide variety of species. Most of the conserved AIPs in this study were found between closely related species and are congruent with the phylogenetic relationships of the surrounding sequences. While this is not unexpected, it is less clear whether those AIPs that span entire exons in *Arabidopsis*, as found for the STK and ABS homologs, are true cases of conservation. These AIP types might be easier to detect at larger distances because exons are in general much better conserved than intronic sequences. As these AIPs clearly contain sequences that resemble the interaction motifs detected in *Arabidopsis*, the effect of the AIP might also resemble the potential effect in *Arabidopsis*.

The importance of considering the effect of AS on conserved interaction motifs is highlighted by a very recent study (Airoldi et al., 2010). It was shown in that study that obtained phenotypic differences can be explained by the inclusion or exclusion of a single amino acid into a conserved sequence region from the *Arabidopsis* AG and the *Antirrhinum* FAR proteins. This phenotypic difference is caused at the molecular level by altered interaction specificities resulting from the presence or absence of the single amino-acid. Interestingly, the single amino acid insertion into FAR that is the crucial difference responsible for functional diversification between this protein and its paralog *PLENA* (PLE) is identical to the observed differences between AS variants of AG homologs. From an evolutionary perspective, FAR and PLE could represent two ancient AS isoforms that have been partitioned over separate genes. Although a number of cases of such externalization (Irimia et al., 2010) events have been described in literature (Yu et al., 2003; Rosti and Denyer, 2007; Irimia et al., 2010), the extent to which this process has contributed to the protein diversity in plants is not known.

MAF1 is a member of the *FLC*-like clade in *Arabidopsis* which consists of genes that are subject to extensive alternative splicing and genomic rearrangements (Scortecci et al., 2001; Ratcliffe et al., 2003; Balasubramanian et al., 2006; Caicedo et al., 2009). An interesting aspect of the *MAF1* locus is the complexity of the rearrangement of motif content that results from a mutually exclusive exon event. While one of the mutually exclusive exons could be detected in transcripts of other *FLC*-like genes, remnants of the cryptic exon were only detected in the closely related *MAF2*- and *MAF3* loci. As the *MAF1-3* loci form a monophyletic group within the *FLC*-like clade (e.g. Figure 6 in Parenicova et al., 2003) it is likely that the cryptic exon originated recently within the ancestor of this group. Alternative splicing has previously been associated with the emergence of new exons (Keren et al., 2010). In accordance with the view that AS provides genes with “internal paralogs” (Modrek and Lee, 2003), the recent exon provides the material for these *MAF* proteins to "experiment" with new dimer interactions without the irreversible loss of existing interactions. The *MAF2*- and *MAF3* loci are not likely to express the region corresponding to the cryptic exon at the protein level. Such rapid divergence of AS patterns between recent paralogs has previously been demonstrated to be a common phenomenon (Su et al., 2006; Zhang et al., 2010).

Natural occurring mutations in splice sites that affect splicing patterns have been identified in *FLOWERING LOCUS C (FLC)* related genes in different species. An *FLC* gene has been identified in the *Arabidopsis* Bur-0 accession that produces a transcript encoding a protein with a modified C-domain (both insertion and truncation) due to a mutation affecting the acceptor site of the sixth intron (Werner et al., 2005). In that same study it was reported that the *FLC* gene of the Van-0 accession has a non-sense mutation in the sixth exon. The gene produces two transcripts that differ by the presence of the sixth exon. Inclusion and exclusion of the sixth exon lead to a protein with a truncated C-domain or large deletion in the C-domain, respectively. Although the mutation in the Van-0 accession does not directly affect a splice site, it can affect a potential *cis*-element such as a splicing enhancer element that might be needed for correct recognition of this exon which is very short compared to the average *Arabidopsis* exon (42 nt. versus 217 nt on average (Barbazuk et al., 2008)). Natural variation affecting the donor site of the sixth intron of the *FLC* gene in *B. rapa* has been reported (Yuan et al., 2009). Individuals carrying this mutation produce two transcripts in which either an alternative donor site within intron 6 is used or similarly as in the Van-0 *Arabidopsis* accession, the sixth exon is skipped entirely. It is however, not known whether the early flowering phenotypes associated with these reported *Arabidopsis* and *B. rapa FLC* genes are the result of reduced multimer formation capabilities of FLC or reduced transcription activation capabilities.

Another naturally occurring splice site mutation in *Capsella bursa-pastoris* results in the skipping of the fifth *FLC A* exon and is associated with early flowering (Slotte et al., 2009). This exon corresponds to the fourth exon of the *Arabidopsis ABS* gene which encodes the earlier mentioned motif involved in higher order interaction specificities. We therefore hypothesize that the early flowering phenotype associated with this exon skipping event is the result of altered multimer formation of *FLC A*.

In addition to these cases in which *cis* splice signals were genetically altered, cases have been found in which the relative abundance of the splice variants from *FLC* related genes is regulated in a manner dependent on temperature (Balasubramanian et al., 2006; Balasubramanian and Weigel, 2006; Reeves et al., 2007) or developmental stage (Zhang et al., 2009). These findings suggest a "conserved" usage of the alternative splicing process as mechanism for regulating flowering-time. However, the details of flowering-time regulation can differ between species even if the AS events in homologous genes appear to be conserved. For instance, the temperature-dependent skipping of the fifth exon in the *Beta vulgaris FLC* *like* 1 gene results in transcripts encoding proteins with stronger repressor capabilities than those encoded by transcripts that include the exon (Reeves et al., 2007). The fifth exon that is skipped in *C. bursa-pastoris FLC A* transcripts is homologous to the skipped exon in *B. vulgaris* (data not shown). However, *C. bursa-pastoris* transcripts that contain the exon are stronger repressors and those transcripts that lack the exon are associated with early flowering phenotypes (Slotte et al., 2009). Although, the proposed effects of the AS products are opposite, these findings provide strong evidence for a role of AS in flowering-time regulation.

Considered at the level of the gene family, the impact of AS on the MIKC subgroup does not appear to dramatically differ from previously established genome-wide patterns. For instance, the dynamic nature of the process is emphasized by the association between AS and fast evolving protein regions and the rapid divergence of AS in recent paralogs, which we find here but is also known on a genome-wide scale. However, usage of available detailed annotation of specific functions to various sequence regions enables the identification of AS events that have an impact on function but that remain hidden behind the globally observed patterns. Such sequence annotations are more specific than for example PFAM domains, and although such knowledge is often not found in databases but somewhat hidden in the literature, for many protein families experimental and predicted sequence regions are known that impact aspects of protein functions. Hence, the results of our study are encouraging for future analysis of AS within other gene families following a similar approach.

CONCLUSION

A large fraction of the AS events in genes encoding MADS domain proteins that are likely to be manifested at the proteome level are predicted to be functional according to our criteria. For many of these cases, indications could be given about which of the aspects of protein complex formation are possibly affected by the AS events. Using additional experimental data (qRT-PCR, overexpression studies) as well as additional bioinformatics analysis (translatome data) we present supporting evidence for the potential biological relevance of the predicted functional AS events. In addition to demonstrating the functional impact of AS on extant MIKC proteins, we also illustrated how the process can potentially introduce new interactions into the network. The analysis we present here provides a starting point for further experimental determination of the precise physiological roles of AS events in plant MADS-box genes.

MATERIAL AND METHODS

Initial data

The identifiers corresponding to the *Arabidopsis* MADS-box genes were extracted from the "gene_families_sep_29_09_update.txt" file, which was downloaded from the TAIR website (www.*Arabidopsis*.org). The genome and predicted proteome of *Arabidopsis thaliana* version TAIR 8.0 were used in this study. cDNA and EST sequences of *Arabidopsis* were downloaded from the PlantGDB webpage (Dong et al., 2004). Gene models and AS events were predicted using our previously described method (Severing et al., 2009). Only those TAIR loci were considered that had at least one annotated protein with an open reading frame (ORF) that was fully supported by cDNA/EST evidence. Transcripts containing premature termination codons (PTCs) located more than 55 nt upstream of the last exon/intron boundary were considered to be NMD targets (Hori and Watanabe, 2007).

Domain annotation

The PFAM (Finn et al., 2006) domain architectures of the MADS-proteins were determined using InterProScan (Zdobnov and Apweiler, 2001). Only those MIKC proteins encoding both the MADS- and K-domains were considered further. A multiple sequence alignment of MADS proteins with clear MIKC structures was build using clustalX2 (Larkin et al., 2007) in order to designate the boundaries in accordance with a previously published structural annotation (Figure 1 in Henschel et al., 2002).

Interaction motifs and protein-protein interaction

Putative interaction motifs and protein-protein interactions were predicted using our recently developed IMSS predictor (van Dijk et al., 2010). In brief, using protein sequences and yeast two-hybrid interaction data, pairs of short sequence motifs are found that occur more often in pairs of interacting proteins compared to pairs of non-interacting proteins. In this motif search, only *Arabidopsis* sequences were used, but the obtained interactions motifs were subsequently found to display reasonably strong conservation. Additional validation was obtained using various bioinformatics analyses of those motifs, and most importantly using experimental mutagenesis on motif sites to change interaction specificities of various MADS- domain proteins. The predictions were only done for protein isoforms of loci that were represented in the previously published interaction map of *Arabidopsis* MADS-box genes (de Folter et al., 2005). Visualization of the SVP1 and SVP3 interaction network was done using Cytoscape version 2.8.1 (Smoot et al., 2011).

Homology searches

The “est_others” file was downloaded from the NCBI blast database ftp site (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/) and translated using the getorf program from the EMBOSS package (Rice et al., 2000) version 5.0. BlastP (Altschul et al., 1997) (complexity filter off; max e-value: 1e^{-10}) searches were performed with the translated sequences against the *Arabidopsis* proteome. Only the best matches involving an *Arabidopsis* MIKC protein were kept, given that the corresponding alignment had an identity of at least 40% and included at least 60% of the residues of the *Arabidopsis* protein.

Identification of AS-induced polymorphism (AIP)

CAP3 (Huang and Madan, 1999) (-p 95 –o 100) was used for clustering ESTs from a single species that had a protein from the same *Arabidopsis* locus as their best blast hit. QualitySNP (Tang et al., 2006) was used for dividing these initial clusters into groups of sequences that were likely to represent the same allelic gene variant. The sequences in these groups were assembled into final transcript contigs using CAP3.

The identification of AIPs relied on the assumption that intron positions within the coding region of genes are generally well conserved. This assumption is based on a previous study in which it was demonstrated that around 94% of introns within conserved regions of proteins are shared between *Arabidopsis* and rice (Roy and Penny, 2007). The positions of putative introns in the subject species were mapped by constructing Needleman-Wunsch (Needleman and Wunsch, 1970) alignments between the subject sequences and their corresponding *Arabidopsis* homologs with known gene structures.

AIPs were identified as indels that coincided with putative intron positions on global alignments between proteins that were predicted to be isoforms of the same gene. It was required that ten alignment positions immediately flanking the polymorphic region contained at least four identical residue pairs and at most one gapped position. It was furthermore required that one of the two isoform sequences could be aligned to the *Arabidopsis* homolog without any gaps around the intron position. As a result of this requirement, AIPs could be represented by amino acid sequences. AIPs were considered to be conserved when the global alignment between corresponding sequences were at least 40% similar.

Confirming predicted functional AS events

Expression of the putative AS events was analyzed by qRT-PCR. Based on known expression patterns of the concerning genes, either cDNA from leaves or carpels was used. For leaf material the full-grown first leafs from five individual plants were taken and carpels were isolated from flowers just prior to opening. RNA was isolated by lithium chloride-phenol-chloroform extraction (Verwoerd et al., 1989) followed by DNase (Invitrogen) treatment. One microgram RNA was used to perform cDNA synthesis using M-MuLV Reverse Transcriptase (Promega). The cDNA made this way was diluted four times and

used for quantitative Real-Time PCR (qRT-PCR) using the SYBR green mix from BioRad. At5g08290, which was determined as "superior reference gene" (Czechowski et al., 2005), was used as reference for the analyses. As control for DNA contamination a minus RT reaction was performed. As second control we performed a "non-template" reaction. All reactions were done in duplo. The oligonuceotides used in the transcript analyses are given in Table S6. Specificity of oligonucleotides was determined by sequencing (DETT sequencing, Amersham) of the amplified fragments.
As additional validation, the short sequence reads from the study of Jiao and Meyerowitz (Jiao and Meyerowitz, 2010) were mapped against the *Arabidopsis* genome using the program TopHat version 1.3.0 (Trapnell et al., 2009). For each of the functional AS events it was determined how many reads supported each of the two transcript variations associated with the event according to a set of rules as used by Wang and co-workers (Wang et al., 2008).

SVP overexpression studies

Entry vectors containing the full length coding sequences of *SVP1* (At2g22540.1) and *SVP3* (At2g22540.2) (van Dijk et al., 2010) were recombined via a Gateway LR reaction (Invitrogen, Carlsbad) with a binary destination vector containing a *CaMV35S* promoter for the purpose of ectopic expression. Generated expression vectors CZN094 (*pCaMV35S::SVP1:NosT*) and CZN756 (*pCaMV35S::SVP3:NosT*) were transformed into *Agrobacterium tumefaciens* strain C58C1/PMP90 followed by transformation of *Arabidopsis thaliana* Col0 wild type plants using the standard floral dip method (Clough and Bent, 1998). Primary transformants were selected based on selective germination and analysis of expression levels of the ectopically expressed *SVP1* and *SVP3* transgenes by qRT-PCR (for oligonucleotide sequences see Table S6). Possible effects on flowering-time were analysed by growing a population of progeny plants (n = 25) from two selected independent primary transformants of CZN094 and CZN756, respectively. As a control Col0 wild type plants were analysed (n = 15). Plants were grown under short day conditions (8 hours light, 16 hours dark; 21 ºC). For each individual plant the number of days from sowing until bolting was scored, as well as the number of rosette leaves at the moment of bolting.

Phylogenetic analysis of Arabidopsis AGAMOUS homologs

A multiple protein sequence alignment of identified *Arabidopsis* AG homologs was constructed using ClustalX2 and trimmed at the column corresponding to the first residue. ClustalX2 was also used for creating a consensus neighbor-joining tree (Saitou and Nei, 1987) from 1000 bootstrap replicates generated from the trimmed alignment. Gapped positions were excluded and a correction was applied for multiple substitutions during the tree construction procedure. The consensus tree topology was edited using the tree explorer from the MEGA package (Kumar et al., 2008) version 4. The following specific varieties or subspecies as annotated in the NCBI taxonomy database (Sayers et al., 2009) are represented in the phylogenetic tree: *Raphanus raphanistrum subsp. raphanistrum*, *Brassica rapa subsp. pekinensis*, *Raphanus sativus var. oleiformis*, *Petunia axillaris subsp. axillaris* and *Pentunia x hybrida*.

Acknowledgements

This work was supported by the BioRange program (SP 3.2.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI), by the Netherlands Organization for Scientific Research (NWO, VENI Grant 863.08.027 to ADJvD) and by a NGI Horizon Breakthrough project (Grant 40-41009-98-8003 to RGHI). G.M. was supported by a scholarship of the Leonardo da Vinci project Unipharma-Graduates (www.unipharmagraduates.it) coordinated by Sapienza University of Rome. We would like to thank Michiel Lammers for help with the qRT-PCRs.

SUPPLEMENTARY FIGURES

```
G.max-1                KEAL--ELSS
G.max-2                KEALVLELSS
C.tetragonoloba-1      REAL--ELSS
C.tetragonoloba-2      REALVMELSS
B.napus-1              REALA-ELNS
B.napus-2              REALAVELNS
R.sativus-1            REALA-ELNS
R.sativus-2            REALAVELNS
A.thaliana-1           REALA-ELSS
A.thaliana-2           REALAVELSS

                            ▲
```

Figure S1. Conserved AIPs in homologs of the Arabidopis *SEPALLATA3 (SEP3)* protein. The intron position corresponding to the AIP of the Arabidopsis *SEP3* isoforms is indicated by the black triangle. *A.thaliana*-1 and *A.thaliana*-2 correspond to *SEP3.1* and *SEP3.2*, respectively.

```
B.rapa.p-short      DNMTNIVDRYEIQHAGELRSLDLAEKTRNYLPHKELLESVKS---------VSVDSLISL
B.rapa.p-long       DNMTNIVDRYEIQHAGELRSLDLAEKTRNYLPHKELLESVKSNLEEPNVDSVSVDSLISL
B.napus-short       DNMTNIVDRYEIQHAGELRSLDLAEKTRNYLPHKELLESVKS---------VSVDSLISL
B.napus-long        DNMTNIIDRYGIQHACELRSLDLAEKTRSYLPHNELLESVKSNLEESNVDNASVDSLISL
MAF3-short          DNMSKIIDRYEIHHADELKALDLAEKIRNYLPHKELLEIVQS-----------VDSLISM
MAF3-long           DNMSKIIDRYEIHHADELKALDLAEKIRNYLPHKELLEIVQSKLEESNVDNVSVDSLISM
MAF2-short          DNMSKIIDRYEIHHADELEALDLAEKTRNYLPLKELLEIVQS-----------VDTLISL
MAF2-long           DNMSKIIDRYEIHHADELEALDLAEKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISL
                                                            ▲
```

Figure S2. Conserved AIPs in Brassica homologs of the Arabidopsis *MADS AFFECTING FLOWERING 2 and -3 (MAF2 and MAF3)* proteins. The conserved intron position corresponding to both the AIPs of the Arabidopsis *MAF2*- and *MAF3* isoforms is indicated by the black triangle. B.rapa.p corresponds to *Brassica rapa subsp. pekinensis.* MAF3-short and MAF3-long correspond to *MAF3.2* and *MAF3.1*, respectively. MAF2-long corresponds to *MAF2.2*.

```
T.officinale-long       LKELKQLETRLERGISKIRSKKHDMILAETENLQKREVELEHHNAFLRSKVQIGERVQQL
T.officinale-short      LKELKQLETRLERGISKIRSKK--------------EVELEHHNAFLRSKVQIGERVQQL
A.thaliana-long         VKELKQVENRLEKAISRIRSKKHELLLVEIENAQKREIELDNENIYLRTKVAEVERYQQH
A.thaliana-short        VKELKQVENRLEKAISRIRSKK--------------EIELDNENIYLRTKVAEVERYQQH
                                              ▲             ▲
```

Figure S3. Conserved AIP for the Arabidopsis *SEEDSTICK (STK)* protein. The introns flanking the skipped exon corresponding to the AIP of the Arabidopsis *STK* isoforms are indicated by black triangles. The *A,thaliana*-short and *A.thaliana*-long isoforms correspond to *STK.1* and *STK.2*, respectively.

```
R.communis-long       EELGELEQELERSVAKVRDRKNELLQQQLDNLRRKERMLEEENGNMYRWIQ-EHRAALEY
R.communis-short      EELGELEQELERSVAKVRDRK--------------ERMLEEENGNMYRWIQ-EHRAALEY
G.hirsutum-short      EELDQLEQELERSVNKVRERK--------------ERMLEEENNNMYRWIQ-EHRAAIEY
G.hirsutum-long       EELDQLEQELERSVNKVRERKNELLQQQLDNLRRKERMLEEENNNMYRWIQ-EHRAAIEH
B.napus-long          HELEGLEQQLEHSVRKVRERKKELLQQQLGNLSRKKRMLEDDNNNMYRWLHDEHRTGVEF
B.napus-short         HELEGLEQQLEHSVRKVRERK-----QQLGNLSRKKRMLEDDNNNMYRWLHDEHRTGVEF
A.thaliana-short      NELDGLERQLEHSVLKVRERK-----QQLENLSRKRRMLEEDNNNMYRWLH-EHRAAMEF
A.thaliana-long       NELDGLERQLEHSVLKVRERKNELMQQQLENLSRKRRMLEEDNNNMYRWLH-EHRAAMEF
                                           △             ▲
```

Figure S4. Conserved AIPs between homologs of the Arabidopsis *B-sister (ABS)* protein. The location of the intron corresponding to the AIP in Arabidopsis is indicated by the grey triangle. The conserved AIP between *R. communis* and *G. hirsutum* corresponds to an entire exon in Arabidopsis. The introns flanking this exon are indicated by the grey and black triangles. The *A.thaliana*-short and *A.thaliana*-long isoforms correspond to *ABS.2* and *ABS.1*, respectively.

```
MAF3  GTAAGT---TAGCTACGAACATCATC--AAAATTCTCTGGAATGCAGTTTTTGATGATAT
MAF2  GTAAGT---TAGCTACGAATATCATT--AAAATTCTTCTGGATGCAGTTTTTGATGTTAT
MAF1  GTAAGTAATTAGCTAAGAACGTCATTCTAATATTCTTCTGGATGCGGTTTTTGGTGTTAT
      ******   ****** ***  ****   ** *****   * **** ******* ** ***

MAF3  TATGGGATAGAATTACTGGTCGAGCCTGAGATAACTCAATGATTTGAATTTCTTAAACTG
MAF2  GAAGGGATAGAAGCACTGGTCGAACCTGAGATAACTCAATGTTTTGAATTTTTCGTACTG
MAF1  GA-AGGATAGAAGCGCTGTTCAAGCCGGAGAAACCTCAATGTTTTGAACTCGTAACACCG
       *  ********   *** ** * ** **** * ******* ****** *  *   ** *

MAF3  GACTTAATTTTCTTAAGTAACAGTTGTTGCATTTTTAGAAAAAACTCAAGAACTTTCACA
MAF2  GACTTAATTCTCTCTAGTTACAGTTATTGCATTTTTCGGAAAA-----------CTCACA
MAF1  AACTTAATTCTCTAGAGTTACAGTTATTGTGTCTACTGGAAAA--TACAAGAACTTCACA
       ******** ***  *** ****** ***  * *   * ****            *****

MAF3  ATCTTTCTGACCATTGCCTTCGTTTTATCCATGTTCAG
MAF2  ATCTTTCTGACCATTCCCTTCGTTTTCTCCATGTGCAG
MAF1  ATCTTTCTGACCATTCC-----TTTTCTTCATGTGCAG
      *************** *     **** * ***** ***
```

A

```
MAF3 IELLVEPEITQ*FEF
MAF2 IEALVEPEITQCFEF
MAF1 IEALFKPEKPQCFEL
```

B

Figure S5. Conservation of the *MADS AFFECTING FLOWERING1 (MAF1)* cryptic exon sequence. A. Multiple sequence alignment of the second intron from the Arabidopsis *MAF1-3* genes. The conserved region corresponding to the cryptic exon from the *MAF1* gene is highlighted by the shaded box. **B.** Multiple sequence alignment of the translated intronic regions from the *MAF2*- and *MAF3* gene and homologous cryptic exon of *MAF1*.

SUPPLEMENTARY TABLES

Table S1. Effect of AS on predicted interaction motifs. The residues from each motif that overlap with AIPs are underlined. Residues that are inserted into motifs are surrounded by square brackets. Multiple overlapping interaction motifs that are affected by the same AIP are stacked. Multiple non-overlapping motifs that are affected by the same AS are numbered.

TAIR 8 Locus	Symbol	Event	Description motif effect	Isoform containing the affected motif
AT1G77080	MAF1/FLM	intron retention	Motif 1: LETVQRLA Motif 2: LPSSSDKK	Motif 1: *MAF1.4* Motif 2: Retained inton
		mutually exclusive exons	Motif 1: KIIDRYEI Motif 2: ELRALDLE Motif 3: QCFELDLE	Motif 1: *MAF1.4* Matif 2: *MAF1.2* Motif3: *MAF1.2*
AT2G22540	SVP	alternative donor	LFEFCSSS EFCSSSMK FCSSSMKE CSSSMKEV SSSMKEVL SSMKEVLE	*SVP.1*
AT2G42830	SHP2	alternative donor	QKR[VK]EIELQ	*SHP-2.2*
AT3G58780	SHP1	alternative acceptor	GTIERYkk	*SHP-1.1*
AT4G09960	STK	exon skipping	Motif 1: RSKKHELL KKHELLLV LLVEIENA Motif 2: QKREIELD	*STK.1*
AT5G23260	ABS-1, ABS-2	alternative acceptor	RERKNELM	*ABS.2*

Chapter 6

General Discussion

Aim of the research

The aim of the research presented in this thesis was to obtain a better understanding of the role of alternative splicing (AS) in expansion of the functional diversity of the proteome in plants. The central question that formed the basis of all research chapters can be formulated as: To what extent is AS used as a mechanism for producing multiple functional proteins from a single gene? The chapters each take a different perspective in answering this question by analyzing different characteristics of AS events and protein isoforms. Here we discuss the results and conclusions from the research chapters within the context of the growing literature about AS in plant and animal species.

Limited conservation of AS events

In Chapter 2, we started from the premise that a conserved AS event is more likely to be functional than one that is not conserved. We specifically considered AS events as mechanical switches that enable the production of multiple proteins from a single gene. Therefore conservation did not only refer to the location and type of the AS event but also to the protein polymorphism that is induced by the event.

The first comparative analysis we performed resulted in the identification of only a limited number of AS events that were conserved between *Arabidopsis thaliana* and *Oryza sativa*. As these species are quite distantly related with an estimated divergence time of ~150 million years (Brendel et al., 2002), we performed a second comparative analysis between the more closely related monocot species *O. sativa* and *Zea mays*. AS was also weakly conserved between these species, which diverged around 50-60 million years ago (Brendel et al., 2002). Our results were consistent with those from another comparative analysis (Wang et al., 2008) that was performed between the legume species *Glycine max*, *Medicago truncatula* and *Lotus japonicus* which shared a common ancestor around 54 million years ago (Lavin et al., 2005). Collectively, taking into account the divergence times between the species analyzed, these studies suggest that most AS events in plants are likely to be lost in less than 50 million years.

High numbers of species-specific AS isoforms and lack of conserved events are common patterns that have also been observed in several comparative analyses between different animal (sub) species (e.g. Harr and Turner, 2010 ;Pan et al., 2005; Calarco et al., 2007). It has even been shown that AS patterns can substantially differ between individuals from the same species (Nembaware et al., 2004; Kwan et al., 2008; Wang et al., 2008). Hence, it is possible that many AS events in plant are also lost within much less than 50 million years. An exception to this trend was observed in the genus *Caenorhabditis*, where AS appeared to be highly conserved between *Caenorhabditis elegans* and *Caenorhabditis briggae,* two species that diverged around 100 million years ago (Barberan-Soler and Zahler, 2008). Nevertheless, when conservation is used as an indicator for function, most comparative analyses performed to date suggest that AS has only a limited impact on the functional diversity of the proteome. In support of this conclusion is the recent proposal that the majority of AS isoforms in mammals, especially those that are species-specific, are more likely to be the result of noise from the splicing machinery than to be functional (Zambelli et al., 2010).

Expression of AS isoforms at the proteome level

Eukaryotes have multiple surveillance mechanisms that prevent the accumulation of toxic proteins in their cells. For instance, the production of truncated proteins can be limited by recognizing end-degrading transcripts that encode premature termination codons. Eukaryotes also have the ability to recognize and degrade proteins which are unable to fold correctly (Goldberg, 2003). It can therefore be assumed that detectable expression of a protein is an indication that such a protein is functional.

High throughput proteomics technologies provide the data that can be used to directly determine which proteins are expressed in a particular biological sample. This type of data is also suitable for assessing the contribution of AS to the proteome diversity. In Chapter 3, we studied the impact of AS on the proteome of *A. thaliana* by analyzing two independent large-scale proteomics datasets (Baerenfaller et al., 2008; Castellana et al., 2008). The aim in this Chapter was to determine whether the predicted contribution of AS to the proteome diversity as determined using transcriptomics data is also observed at the proteome level.

Using the proteomics data, we could confirm the expression of a limited number of AS isoforms. However, for only one third of the loci that were expressed and predicted to produce multiple protein isoforms, it was possible to determine by which specific isoform they were represented. This raised the question whether the low number of confirmed AS events was the result of limited experimental sampling depth or limited contribution of AS events to the proteome diversity of *A. thaliana.*

By performing *in silico* AS detection experiments we showed that AS events were not underrepresented in the experimental datasets. This result indicated that the sampling depth was the most likely reason for the limited numbers of AS events that could be confirmed. It was therefore concluded that many AS events, including those that are species-specific, are manifested at the protein level. This conclusion is partially supported by a recent study in which it was observed that multiple AS transcript isoforms from several loci were actively translated (Jiao and Meyerowitz, 2010). Our study further suggests that the resulting proteins might not be subjected to increased turnover.

In one of the two experimental sets the number of confirmed AS events was higher than the expected number of events as determined by the *in silico* AS detection experiments. An additional analysis revealed that the AS events in this experimental set were biased towards disordered regions of proteins. Our result is consistent with a recent study that showed that AS-induced polymorphisms that distinguish expressed protein isoforms, tend to be biased towards regions that are disordered or located at the protein surface (Hegyi et al., 2011). A possible explanation for these biases is that AS events within these regions can introduce variation without significantly disrupting the remaining structure of the protein. As disordered regions are often involved in mediating molecular interactions (Dunker et al., 2008), AS events within these regions might play a role in regulating these interactions. In summary, under the premise that expression is an indication for function, it is likely that many AS events contribute to functional proteome diversity in plants.

Alternative ways for conserving AS isoforms

In Chapter 4 we investigated the conservation of AS events as mechanical switches that allow the production of multiple distinct protein isoforms from a single gene. The lack of conserved AS events suggests that AS has a limited contribution to functional diversity of the proteome. However, novel AS isoforms can also be conserved without conservation of the corresponding AS event *per se*. In Chapter 4 we addressed this hypothesis by analyzing a large collection of EST/cDNA sequences from hundreds of different species.

We first analyzed the extent to which AS events were conserved within our dataset. We found that only around 16% of all identified AS events were conserved in two or more species. Furthermore, half of the pairs of conserved AS events were detected in only 6% of all analyzed homologous sequence clusters. A possible explanation for such a distribution is that only a few genes have employed AS as a potential mechanism for producing multiple functional proteins. In these rare cases, conservation of AS events appears to have been opted for in evolution.

The first alternative way for conserving AS isoforms is through a process called externalization (Irimia et al., 2010). This involves a gene duplication event that is followed by each of the duplicate genes adopting one of the splice patterns of the parent gene. For only 4% of AS events in our dataset we were able to find complementary paralog pairs. The apparent scarcity of externalization events is consistent with a previous study in animal species in which it was demonstrated that the differences that are typically observed between paralogous sequences are not comparable to those observed between AS isoforms of the same gene (Talavera et al., 2007). It has previously been shown that AS prevalence and gene family size are negatively correlated (Kopelman et al., 2005; Su et al., 2006). It has been suggested that this negative correlation can be attributed to subfunctionalization events involving the externalization of AS isoforms (Su et al., 2006). However, the scarcity of externalization events that we observed does not support such a scenario within the plant kingdom.

A second alternative way of conserving AS events involves the replacement of preexisting isoforms by novel ones. Although we identified several of such isoform replacement cases, a preliminary analysis indicates that usually one isoform dominates while other variants were only represented in a few (and often only one) species. We therefore concluded that isoform replacement is a relatively rare phenomenon in the evolution of plant proteins.

Our overall conclusion in chapter 4 was that the lack of AS events is not the result of a preference for retaining AS isoforms by means other than retention of the corresponding AS event, i.e. through externalization or isoform replacement. One would expect many AS isoforms to be conserved if AS substantially increases the functional diversity of the proteome. It is therefore unlikely that novel functions often arise through AS and we hypothesize that the majority of AS isoforms are not functional. This conclusion is supported by previous research in which it was shown that many AS events can have a drastic impact on protein structure (Tress et al., 2007; Melamud and Moult, 2009). Disruption of the structure of a protein is very likely to impair its function. An additional analysis of the nucleotide distances between homologs with conserved AS events indicated that the vast majority of AS isoforms are lost within a similar time period as redundant gene copies.

Impact on molecular function

Several genome-wide studies have uncovered interesting patterns that link AS to for instance protein domains (Resch et al., 2004; Hegyi et al., 2011) or evolutionary processes such as the birth and death of exons (Keren et al., 2010). However, in order to understand the biological implications of these patterns, it is of interest to study them not only globally but also within the functional context of individual genes and gene families. In Chapter 5, we performed such a detailed analysis of AS within the well studied MIKC subgroup (Type II) of the MADS box transcription factor family in plants. In order to study the functional impact of AS within this family we specifically made use of the fact that the entire primary sequence of MADS proteins is divided into regions with specifically assigned functions.

For instance, it was observed in Chapter 2 that, similarly as in animal species (Kriventseva et al., 2003), AS events in plants tend to occur more frequently outside than within the boundaries of predicted protein domains. Protein regions with assigned protein domains are often better preserved than protein regions without assigned protein domains. Under the assumption that highly conserved protein regions are more likely to be functional than more variable, less conserved regions, the biased location of AS events can be interpreted as a preference for avoiding AS induced polymorphisms in functional regions of proteins. However, a fast evolving protein region does not mean that it does not have a conserved function.

Although, the I-domain is less conserved compared to other regions of MADS proteins, it plays an important role in determining the dimerization interaction specificities of MADS proteins. AS events within this domain can, according to both *in silico* predictions and experimental data, affect the interaction specificities of MADS proteins. However, there is currently not enough data available to determine whether the elevated levels of AS within the I-domain are the result of functional benefit or coincidence with the fast evolution of this domain.

The majority of AS-induced polymorphisms in the MADS protein family were located at the boundary of – or fully within the K-box domains of these proteins. Although the number of pairwise conservation cases was high given the results of previous comparative analyses, they were mostly identified between very closely related species. Nevertheless, it has for at least two of these cases been shown that the isoforms associated with the conserved AS induced polymorphism not only formed different higher-order complexes but also caused different phenotypes.

Finally we demonstrated the contribution of AS in generating diversity on both the sequence and the interaction-network level in this family. Through the selective inclusion of either an ancient – or relatively young exon, two protein isoforms with different interaction motifs in the I-domain are produced from the *MADS AFFECTING FLOWERING 1 (MAF1)* gene. This AS event is an example of how a gene is provided with an internal paralog that allows for evolution of novel functions (interactions) without the loss of the existing function. In accordance with previous studies in which it was shown that AS patterns rapidly and extensively diverge after gene duplication events (Kopelman et al., 2005; Su et al., 2006; Zhang et al., 2010), the capability of including the young exon has been lost in very recent duplicates of the MAF1 gene. In summary, in Chapter 5, we provided examples of the functional implications for individual genes that underlie genome wide trends observed for AS.

Concluding remarks

It has been suggested that the majority of AS isoforms within a cell is likely to be the result from noise in the splicing machinery (e.g. Pickrell et al., 2010). The novel AS isoforms that arise from this splicing noise are internal paralogs which can evolve novel functions (Modrek and Lee, 2003) while original function of the corresponding gene is retained. From this perspective AS can be viewed as a mechanism that increases genetic diversity. Even if the majority of AS isoforms result from noise, their expression at the proteome level (Chapter 3) is required in order for selection to “identify” the potential functional ones. Expression of non-functional proteins might not be a problem as long as it does not interfere with cellular processes and excessively increase metabolic load. *In silico* identification of those AS isoforms that are likely to be functional is not trivial and requires detailed analyses of individual genes or gene families (Chapter 5). The results from our large scale AS isoform conservation analyses (chapter 2 and 4) indicate that the majority of AS isoforms are unlikely to be functional and simply lost within a similar time frame as redundant gene copies that originate through gene duplication events (Chapter 4)

Summary

Splicing is one of the key processing steps during the maturation of a gene’s primary transcript into the mRNA molecule used as a template for protein production. Splicing involves the removal of segments called introns and re-joining of the remaining segments called exons. It is by now well established that not always the same segments are removed from a gene’s primary transcript during the splicing process. The consequence of this splicing variation, termed Alternative Splicing (AS), is that multiple distinct mature mRNA molecules can be produced from a single gene.

One of the two biological roles that are ascribed to AS is that of a mechanism which enables an organism to produce multiple functionally distinct proteins from a single gene. Alternatively, AS can serve as a means for controlling gene expression at the post-transcriptional level. Although many clear examples have been reported for both roles, the extent to which AS increases the functional diversity of the proteome, regulates gene expression or simply reflects noise in splicing machinery is not well known.

Determining the full functional impact of AS by designing and performing wet-lab experiments for all AS events is unfeasible and bioinformatics approaches have therefore widely been used for studying the impact of AS at a genome-wide scale. In this thesis four bioinformatics studies are presented that were aimed at determining the extent to which AS is used in plants as a mechanism for producing multiple distinct functional proteins from a single gene. Each chapter uses a different method for analyzing specific properties of AS.

Under the premise that functional genetic features are more likely to be conserved than non-functional ones, AS events that are present in two or more species are more likely to be biologically relevant than those that are confined to a single species. In chapter 2 we analyzed the conservation of AS by performing a comparative analysis between three divergent plant species. The results of that study indicated that the vast majority of AS events does not persist over long periods of evolution. We concluded, based on this lack of conservation, that AS only has a limited impact on the functional diversity of the proteome in plants. Following this conclusion, it can hypothesized that the variation that AS induces at the transcriptome level is not likely to be manifested at the protein level. In chapter 3 we tested this hypothesis by analyzing two independent proteomics datasets. This type of data can be used to directly identify proteins present in a biological sample. Our results indicated that the variation induced by AS at the transcriptome level is also manifested at the protein level. We concluded that either many AS events have a confined species-specific (not conserved) function or simply produce protein variants that are stable enough to escape rapid turn-over.

Another method for determining whether AS increases the functional diversity of the proteome is by determining whether protein sequence variations that are typically induced by AS are common within the plant kingdom. We found (chapter 4) that this is not the case in plants and concluded that novel functions do not frequently arise through AS. We also found that most of the AS-induced variation is lost, similarly as for redundant gene copies, within a very short evolutionary time period.

One limitation of genome-wide analyses is that these capture only the more general patterns. However, the functional impact of AS can be very different in different genes or gene-families. In order fully assess the functional impact of AS, it is therefore important to also study the process within the functional context of individual genes or gene families. In chapter 5 we demonstrated this concept by performing a detailed analysis of AS within the MADS-box gene family. We were able to provide clues as to how AS might impact the protein-protein interaction capabilities of individual MADS proteins. Some of our predictions were supported by experimental evidence. We further showed how AS can serve as an evolutionary mechanism for experimenting with novel functions (novel interactions) without the explicit loss of existing functions.

The overall conclusion, based on the performed analyses is as follows: AS primarily is a consequence of noise in the splicing machinery and results in an increased diversity of the proteome. However, only a small fraction of the proteins resulting from AS will have beneficial functions and are subsequently selected for during evolution. The large remaining fraction is, similarly as for redundant gene-copies, lost within a very short evolutionary time period after its emergence.

Samenvatting

Splicing is een essentiële stap in het proces waarbij van een primair gen transcript een mRNA molecuul wordt gemaakt dat als blauwdruk dient voor de productie van eiwitten. Tijdens de splicing reactie worden segmenten, genaamd intronen, uit het primaire transcript verwijderd en worden de overgebleven segmenten, genaamd exonen, weer aan elkaar geplakt. Het is algemeen bekend dat niet altijd dezelfde segmenten van een primair transcript tijdens de splicing reactie worden verwijderd. Deze variatie in de splicing reactie wordt Alternatvie Splicing (AS) genoemd en heeft als gevolg dat er verschillende mRNA moleculen gemaakt kunnen worden van een enkel gen.

Een van de twee biologische rollen die aan AS wordt toegeschreven is dat van een mechanisme dat het mogelijk maakt om verschillende eiwitten te produceren van een enkel gen. Daarnaast kan AS ook dienen als een mechanisme waarme het expressie niveau van mRNAs gereguleerd wordt nadat het transcriptie proces al voltooid is. Alhoewel er duidelijke voorbeelden van de bovengenoemde functies beschreven zijn, is het niet bekend in hoeverre AS de eiwit diversiteit vergroot, de expressie van genen reguleert of enkel het gevolg is van ruis tijdens de splicing reactie. De bioinformatica speelt een belangrijke rol bij het beantwoorden van deze vraag omdat het ondoenlijk is om de functionele impact van AS te bepalen door voor elke afzonderlijk AS instantie experimenten te ontwerpen en uit te voeren.

In dit proefschrift worden vier bioinformatica studies gepresenteerd die allen als doel hadden om een beter beeld te krijgen van de mate waarin planten AS gebruiken om meerdere functionele eiwitten te produceren van een enkel gen. In elke studie wordt een andere methode gebruikt om specifieke eigenschappen van AS te bestuderen.

Onder de aanname dat de functionele elementen in een genoom sterker geconserveerd zijn dan niet-functionele elementen, zal een AS instantie die in twee of meer organismen voorkomt eerder biologisch relevant zijn dan een AS instantie die maar in één enkele soort voorkomt. In hoofdstuk 2 wordt een grootschalige studie gepresenteerd waarin AS in drie verschillende planten werd vergeleken. De resultaten van dat onderzoek wezen uit dat de meeste AS instanties niet lang blijven bestaan in de evolutie. We hebben, op basis van dit gebrek aan conservering, geconcludeerd dat AS in planten maar een beperkte bijdrage levert aan de functionele diversiteit van eiwitten.

Deze conclusie heeft vervolgens tot de hypothese geleid dat de door AS geïnduceerde variatie op mRNA niveau mogelijk niet tot uiting komt op eiwit niveau. We hebben deze hypothese getoetst (hoofstuk 3) door twee onafhankelijke proteomics datasets te analyseren. Dit type data kan worden gebruikt om te bepalen welke eiwitten er in een biologisch monster aanwezig waren. De resultaten van onze analyse wezen erop dat de variatie die AS op het mRNA niveau induceert voor een groot gedeelte ook tot uitdrukking komt op eiwit niveau. We hebben op basis van deze resultaten geconcludeerd dat de meeste AS instanties ófwel een beperkte, soort-specifieke functie hebben, of in eiwitten resulteren die zodanig stabiel zijn dat ze niet snel worden afgebroken.

Een andere manier om te bepalen of AS het repertoire van functionele eiwitten vergroot is door te kijken of de typische variatie die het gevolg is van AS veelvuldig voorkomt binnen het plantenrijk. De resultaten van de studie die in hoofdstuk 4 wordt gepresenteerd wees er echter op dat dit niet het geval is. We zijn daardoor tot de conclusie gekomen dat de variatie die AS introduceert maar in enkele gevallen tot nieuwe functies leidt. Een verdere analyse wees erop dat de meeste AS produkten waarschijnlijk, net als redundante gen- kopieën, binnen een zeer korte evolutionaire periode verdwijnen.

Een van de beperkingen van genoomwijde analyses is dat ze over het algemeen alleen de veel voorkomende patronen identificeren. Aangezien de gevolgen van AS erg kunnen verschillen per gen of gen familie, kan de volledige impact van AS pas begrepen worden als het proces binnen de functionele context van individuele genen of gen-families wordt geplaatst. In hoofdstuk 5 wordt dit concept geïllustreerd met een gedetailleerde analyse van AS instanties binnen de MADS box gen familie. We konden een aantal indicaties geven over de mogelijke gevolgen van AS voor de eiwit-eiwit interactie eigenschappen van enkele genen van deze familie. Een aantal van deze voorspellingen wordt ondersteund door experimentele data. We laten ook zien hoe AS de mogelijkheid biedt om met nieuwe functies (nieuw interacties) te experimenteren zonder dat de reeds bestaande functies (interacties) verloren gaan.

De algemene conclusie die uit de studies in dit proefschrift kan worden getrokken is als volgt: AS ontstaat primair als gevolg van fouten (ruis) tijdens de splicing reactie en vergroot het eiwit repertoire. Slechts een kleine fractie van de additionele eiwitten heeft een gunstige functie, waarvoor geselecteerd word in de evolutie. De andere en grotere fractie zal, net als redundante gen-kopieen, binnen een zeer korte evolutionaire periode na hun ontstaan weer verdwijnen.

References

Adams MD et. al (2000) The genome sequence of Drosophila melanogaster. Science 287: 2185-2195

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207

Airoldi CA, Bergonzi S, Davies B (2010) Single amino acid change alters the ability to specify male or female organ identity. Proc Natl Acad Sci U S A 107: 18898-18902

Alt FW, Bothwell AL, Knapp M, Siden E, Mather E, Koshland M, Baltimore D (1980) Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. Cell 20: 293-301

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Ribas de Pouplana L, MartÃ-nez-Castilla Ln, Yanofsky MF (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proceedings of the National Academy of Sciences of the United States of America 97: 5328-5333

Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: structures, functions, and evolution. J Struct Biol 134: 117-131

Artamonova, II, Gelfand MS (2007) Comparative genomics and evolution of alternative splicing: the pessimists' science. Chem Rev 107: 3407-3430

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29

Baek JM, Han P, Iandolino A, Cook DR (2008) Characterization and comparison of intron structure and alternative splicing between Medicago truncatula, Populus trichocarpa, Arabidopsis and rice. Plant Mol Biol 67: 499-510

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320: 938-941

Balasubramanian S, Sureshkumar S, Lempe J, Weigel D (2006) Potent Induction of Arabidopsis thaliana Flowering by Elevated Growth Temperature. PLoS Genet 2: e106

Balasubramanian S, Weigel D (2006) Temperature Induced Flowering in Arabidopsis thaliana. Plant Signal Behav 1: 227-228

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465: 53-59

Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res 18: 1381-1392

Barberan-Soler S, Zahler AM (2008) Alternative splicing and the steady-state ratios of mRNA isoforms generated by it are under strong stabilizing selection in Caenorhabditis elegans. Mol Biol Evol 25: 2431-2437

Becker A, Theißen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Molecular Phylogenetics and Evolution 29: 464-489

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270: 2411-2414

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 74: 3171-3175

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872-2877

Birzele F, Csaba G, Zimmer R (2008) Alternative splicing and protein structure evolution. Nucleic Acids Res 36: 550-558

Blakeley P, Siepen JA, Lawless C, Hubbard SJ (2010) Investigating protein isoforms via proteomics: a feasibility study. Proteomics 10: 1127-1140

Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126: 37-47

Brendel V, Kurtz S, Walbot V (2002) Comparative genomics of Arabidopsis and maize: prospects and limitations. Genome Biol 3: REVIEWS1005

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

Caicedo AL, Richards C, Ehrenreich IM, Purugganan MD (2009) Complex Rearrangements Lead to Novel Chimeric Gene Fusion Polymorphisms at the Arabidopsis thaliana MAF2-5 Flowering Time Gene Cluster. Mol Biol Evol 26: 699-711

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proceedings of the National Academy of Sciences of the United States of America 101: 15670-15675

Calarco JA, Xing Y, Caceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, Blencowe BJ (2007) Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev 21: 2963-2975

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7: 327

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3: 285-298

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 105: 21034-21038

Causier B, Castillo R, Zhou J, Ingram R, Xue Y, Schwarz-Sommer Z, Davies B (2005) Evolution in Action: Following Function in Duplicated Floral Homeotic Genes. 15: 1508-1512

Causier B, Schwarz-Sommer Z, Davies B (2010) Floral organ identity: 20 years of ABCs. Seminars in Cell & Developmental Biology 21: 73-79

Chacko E, Ranganathan S (2009) Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. BMC Genomics 10 Suppl 1: S5

Chang KY, Georgianna DR, Heber S, Payne GA, Muddiman DC (2010) Detection of alternative splice variants at the proteome level in Aspergillus flavus. J Proteome Res 9: 1209-1217

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977) An amazing sequence arrangement at the 52 ends of adenovirus 2 messenger RNA. 12: 1-8

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104: 19428-19433

Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol 8: R64

Clark TA, Sugnet CW, Ares M, Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science 296: 907-910

Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J 16: 735-743

Cox J, Mann M (2007) Is proteomics the new genomics? Cell 130: 395-398

Crick F (1970) Central dogma of molecular biology. Nature 227: 561-563

Cuperlovic-Culf M, Belacel N, Culf AS, Ouellette RJ (2006) Data analysis of alternative splicing microarrays. Drug Discov Today 11: 983-990

Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. Plant Physiol 139: 5-17

De Bodt S, Raes J, Van de Peer Y, Theißen G (2003) And then there were many: MADS goes genomic. Trends in Plant Science 8: 475-483

de Folter S, Immink RGH, Kieffer M, Parenicova L, Henz SR, Weigel D, Busscher M, Kooiker M, Colombo L, Kater MM, Davies B, Angenent GC (2005) Comprehensive Interaction Map of the Arabidopsis MADS Box Transcription Factors. Plant Cell 17: 1424-1433

De Kee DW, Gopalan V, Stoltzfus A (2007) A sequence-based model accounts largely for the relationship of intron positions to protein structural features. Mol Biol Evol 24: 2158-2168

DFCI Gene Indices Software Tools. *In*,

Dong Q, Schlueter SD, Brendel V (2004) PlantGDB, plant genome database and analysis tools. Nucleic Acids Res 32: D354-359

Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN (2008) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9 Suppl 2: S1

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L (1980) Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell 20: 313-319

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797

Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17: 419-437

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93-104

Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol 7: R35

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20: 45-58

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-251

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8: 967-974

Folter Sd, Shchennikova AV, Franken J, Busscher M, Baskar R, Grossniklaus U, Angenent GC, Immink RGH (2006) A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. The Plant Journal 47: 934-946

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc Natl Acad Sci U S A 102: 16176-16181

Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19: 521-532

Gelfand MS, Mironov AA, Pevzner PA (1996) Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A 93: 9061-9066

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

Gingras A-C (2009) 35 years later, mRNA caps still matter. Nat Rev Mol Cell Biol 10: 735-735

Goldberg AL (2003) Protein degradation and protection against misfolded or damaged proteins. Nature 426: 895-899

Graveley BR (2005) Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell 123: 65-73

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503-510

Haas BJ, Zody MC (2010) Advancing RNA-Seq analysis. Nat Biotechnol 28: 421-423

Hallegger M, Llorian M, Smith CW (2010) Alternative splicing: global insights. FEBS J 277: 856-866

Harr B, Turner LM (2010) Genome-wide analysis of alternative splicing evolution among Mus subspecies. Mol Ecol 19 Suppl 1: 228-239

Hartmann U, Höhmann S, Nettesheim K, Wisman E, Saedler H, Huijser P (2000) Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. The Plant Journal 21: 351-360

Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA (2002) Splicing graphs and EST assembly problem. Bioinformatics 18 Suppl 1: S181-188

Hegyi H, Kalmar L, Horvath T, Tompa P (2010) Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. Nucleic Acids Res

Henschel K, Kofuji R, Hasebe M, Saedler H, Munster T, Theissen G (2002) Two ancient classes of MIKC-type MADS-box genes are present in the moss Physcomitrella patens. Mol Biol Evol 19: 801-814

Hertel KJ (2008) Combinatorial Control of Exon Recognition. Journal of Biological Chemistry 283: 1211-1215

Hiller M, Platzer M (2008) Widespread and subtle: alternative splicing at short-distance tandem sites. Trends Genet 24: 246-255

Holm S (1979) A SIMPLE SEQUENTIALLY REJECTIVE MULTIPLE TEST PROCEDURE. Scandinavian Journal of Statistics 6: 65-70

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

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

Hughes AL, Friedman R (2007) Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm Caenorhabditis elegans. Genetica

Iida K, Go M (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. Mol Biol Evol 23: 1085-1094

Immink R, Tonaco I, de Folter S, Shchennikova A, van Dijk A, Busscher-Lange J, Borst J, Angenent G (2009) SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biology 10: R24

Immink RGH, Kaufmann K, Angenent GC (2010) The [`]ABC' of MADS domain protein behaviour and interactions. Seminars in Cell & Developmental Biology 21: 87-93

Irimia M, Maeso I, Gunning PW, Garcia-Fernandez J, Roy SW (2010) Internal and external paralogy in the evolution of tropomyosin genes in metazoans. Mol Biol Evol 27: 1504-1517

Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW (2008) Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol 25: 375-382

Irimia M, Rukov JL, Penny D, Roy SW (2007) Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. BMC Evol Biol 7: 188

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

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302: 2141-2144

Jones P, Cote RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D, Hermjakob H (2008) PRIDE: new developments and new datasets. Nucleic Acids Res 36: D878-883

Kalyna M, Lopato S, Voronin V, Barta A (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. Nucleic Acids Res 34: 4395-4405

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14: 331-342

Kan Z, States D, Gish W (2002) Selecting for functional alternative splices in ESTs. Genome Res 12: 1837-1845

Kaufmann K, Anfang N, Saedler H, Theissen G (2005) Mutant analysis, protein–protein interactions and subcellular localization of the Arabidopsis Bsister (ABS) protein. Molecular Genetics and Genomics 274: 103-118

Kaufmann K, Melzer R, Theißen G (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene 347: 183-198

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11: 345-355

Kim E, Goren A, Ast G (2008) Alternative splicing: current perspectives. Bioessays 30: 38-47

Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35: 125-131

Kim S, Koh J, Ma H, Hu Y, Endress PK, Hauser BA, Buzgo M, Soltis PS, Soltis DE (2005) Sequence and expression studies of A-, B-, and E-class MADS-box homologues in Eupomatia (Eupomatiaceae): Support for the bracteate origin of the calyptra. International Journal of Plant Sciences 166: 185-198

Kitahara K, Matsumoto S (2000) Rose MADS-box genes 'MASAKO C1 and D1' homologous to class C floral identity genes. Plant Sci 151: 121-134

Knowles DG, McLysaght A (2006) High rate of recent intron gain and loss in simultaneously duplicated Arabidopsis genes. Mol Biol Evol 23: 1548-1557

Kofuji R, Sumikawa N, Yamasaki M, Kondo K, Ueda K, Ito M, Hasebe M (2003) Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. Mol Biol Evol 20: 1963-1977

Kondrashov FA, Koonin EV (2001) Origin of alternative splicing by tandem exon duplication. Hum Mol Genet 10: 2661-2669

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

Kornblihtt AR (2006) Chromatin, transcript elongation and alternative splicing. Nat Struct Mol Biol 13: 5-7

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S (2003) Increase of functional diversity by alternative splicing. Trends Genet 19: 124-128

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9: 299-306

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J (2008) Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40: 225-231

Kyozuka J, Harcourt R, Peacock WJ, Dennis ES (1997) Eucalyptus has functional equivalents of the Arabidopsis AP1 gene. Plant Molecular Biology 35: 573-584

Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of alternative splicing. Curr Opin Struct Biol 14: 273-282

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8: 477-486

Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Systematic Biology 54: 575-594

Lee C, Wang Q (2005) Bioinformatics analysis of alternative splicing. Brief Bioinform 6: 23-33

Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 11: 1561-1567

Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100: 189-192

Li D, Liu C, Shen L, Wu Y, Chen H, Robertson M, Helliwell CA, Ito T, Meyerowitz E, Yu H (2008) A repressor complex governs the integration of flowering signals in Arabidopsis. Dev Cell 15: 110-120

Li J, Li X, Guo L, Lu F, Feng X, He K, Wei L, Chen Z, Qu LJ, Gu H (2006) A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice. J Exp Bot 57: 1263-1273

Li Q, Hunt AG (1997) The Polyadenylation of RNA in Plants. Plant Physiol 115: 321-325

Lightfoot D, Malone K, Timmis J, Orford S (2008) Evidence for alternative splicing of MADS-box transcripts in developing cotton fibre cells. Molecular Genetics and Genomics 279: 75-85

Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A 98: 11193-11198

Lin H, Ouyang S, Egan A, Nobuta K, Haas BJ, Zhu W, Gu X, Silva JC, Meyers BC, Buell CR (2008) Characterization of paralogous protein families in rice. BMC Plant Biol 8: 18

Lin H, Zhu W, Silva JC, Gu X, Buell CR (2006) Intron gain and loss in segmentally duplicated genes in rice. Genome Biol 7: R41

Liu S, Altman RB (2003) Large scale study of protein domain distribution in the context of alternative splicing. Nucleic Acids Res 31: 4828-4835

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

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155

Maquat LE (2005) Nonsense-mediated mRNA decay in mammals. J Cell Sci 118: 1773-1776

Masiero S, Li MA, Will I, Hartmann U, Saedler H, Huijser P, Schwarz-Sommer Z, Sommer H (2004) INCOMPOSITA: a MADS-box gene controlling prophyll development and floral meristem identity in Antirrhinum. Development 131: 5981-5990

Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6: 386-398

McGlincy NJ, Smith CW (2008) Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends Biochem Sci 33: 385-393

McGuire AM, Pearson MD, Neafsey DE, Galagan JE (2008) Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol 9: R50

Melamud E, Moult J (2009) Stochastic noise in splicing machinery. Nucleic Acids Res 37: 4873-4886

Melamud E, Moult J (2009) Structural implication of splicing stochastics. Nucleic Acids Res 37: 4862-4872

Miller OL, Jr., Hamkalo BA, Thomas CA, Jr. (1970) Visualization of bacterial genes in action. Science 169: 392-395

Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res 9: 1288-1293

Mo F, Hong X, Gao F, Du L, Wang J, Omenn GS, Lin B (2008) A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. BMC Bioinformatics 9: 537

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

Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34: 177-180

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

Montag K, Salamini F, Thompson RD (1995) ZEMa, a member of a novel group of MADS box genes, is alternatively spliced in maize endosperm. Nucl. Acids Res. 23: 2168-2177

Moraes TF, Edwards RA, McKenna S, Pastushok L, Xiao W, Glover JN, Ellison MJ (2001) Crystal structure of the human ubiquitin conjugating enzyme complex, hMms2-hUbc13. Nat Struct Biol 8: 669-673

Morello L, Breviario D (2008) Plant spliceosomal introns: not only cut and paste. Curr Genomics 9: 227-238

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349

Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2006) Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics 22: 1211-1216

Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23: 198-199

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453

Nembaware V, Wolfe KH, Bettoni F, Kelso J, Seoighe C (2004) Allele-specific transcript isoforms in human. FEBS Lett 577: 233-238

Ner-Gaon H, Fluhr R (2006) Whole-genome microarray in Arabidopsis facilitates global analysis of retained introns. DNA Res 13: 111-121

Nesi N, Debeaujon I, Jond C, Stewart AJ, Jenkins GI, Caboche M, Lepiniec L (2002) The TRANSPARENT TESTA16 Locus Encodes the ARABIDOPSIS BSISTER MADS Domain Protein and Is Required for Proper Development and Pigmentation of the Seed Coat. Plant Cell 14: 2463-2479

Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457-463

Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T (2010) Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of Trypanosoma brucei. PLoS Pathog 6

Nurtdinov RN, Artamonova, II, Mironov AA, Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12: 1313-1320

O'Toole N, Hattori M, Andres C, Iida K, Lurin C, Schmitz-Linneweber C, Sugita M, Small I (2008) On the expansion of the pentatricopeptide repeat gene family in plants. Mol Biol Evol 25: 1120-1128

Olsen JV, Ong SE, Mann M (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 3: 608-614

Pan Q, Bakowski MA, Morris Q, Zhang W, Frey BJ, Hughes TR, Blencowe BJ (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet 21: 73-77

Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev 20: 153-158

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

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ, Blencowe BJ (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16: 929-941

Paquette S, Moller BL, Bak S (2003) On the origin of family 1 plant glycosyltransferases. Phytochemistry 62: 399-413

Paquette SM, Bak S, Feyereisen R (2000) Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of Arabidopsis thaliana. DNA Cell Biol 19: 307-317

Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L (2003) Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World. Plant Cell 15: 1538-1551

Patthy L (1999) Genome evolution and the evolution of exon-shuffling--a review. Gene 238: 103-114

Pickrell JK, Pai AA, Gilad Y, Pritchard JK (2010) Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. PLoS Genet 6: e1001236

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C (2007) Rapid, transcript-specific changes in splicing in response to environmental stress. Mol Cell 27: 928-937

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21: 3435-3438

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. *In*, Vienna, Austria

Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL (2003) Analysis of the Arabidopsis MADS AFFECTING FLOWERING Gene Family: MAF2 Prevents Vernalization by Short Periods of Cold. Plant Cell 15: 1159-1169

Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol 58: 267-294

Reeves PA, He Y, Schmitz RJ, Amasino RM, Panella LW, Richards CM (2007) Evolutionary Conservation of the FLOWERING LOCUS C-Mediated Vernalization Response: Evidence From the Sugar Beet (Beta vulgaris). Genetics 176: 295-307

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041-1052

Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res 3: 76-83

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277

Riechmann JLaM, E.M. (1997) MADS domain proteins in plant development. Biological Chemistry 378: 1079-1118

Rizzon C, Ponger L, Gaut BS (2006) Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in <italic>Arabidopsis</italic> and Rice. PLoS Comput Biol 2: e115

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A 103: 8390-8395

Rosti S, Denyer K (2007) Two paralogous genes encoding small subunits of ADP-glucose pyrophosphorylase in maize, Bt2 and L2, replace the single alternatively spliced gene found in other cereal species. J Mol Evol 65: 316-327

Roy SW, Irimia M (2008) Intron mis-splicing: no alternative? Genome Biol 9: 208

Roy SW, Penny D (2007) Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. Mol Biol Evol 24: 171-181

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815

Sammeth M, Foissac S, Guigo R (2008) A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 4: e1000147

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37: D5-15

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

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37: 501-506

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. 101: 671-684

Scortecci KC, Michaels SD, Amasino RM (2001) Identification of a MADS-box gene, FLOWERING LOCUS M, that represses flowering. The Plant Journal 26: 229-236

Severing EI, van Dijk AD, Stiekema WJ, van Ham RC (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC Genomics 10: 154

Sharp PA (1994) Split genes and RNA splicing. Cell 77: 805-815

Shepard PJ, Hertel KJ (2008) Conserved RNA secondary structures promote alternative splicing. RNA 14: 1463-1469

Shepard PJ, Hertel KJ (2010) Embracing the complexity of pre-mRNA splicing. Cell Res 20: 866-868

Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. Nucleic Acids Res

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

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432

Sterck L, Rombauts S, Vandepoele K, Rouze P, Van de Peer Y (2007) How many genes are there in plants (... and why are they there)? Curr Opin Plant Biol 10: 199-203

Su Z, Wang J, Yu J, Huang X, Gu X (2006) Evolution of alternative splicing after gene duplication. Genome Res 16: 182-189

Sugnet CW, Kent WJ, Ares M, Jr., Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput: 66-77

Sun H, Chasin LA (2000) Multiple splicing defects in an intronic false exon. Mol Cell Biol 20: 6414-6425

Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X (2007) The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol 3: e33

Taneri B, Snyder B, Novoradovsky A, Gaasterland T (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. Genome Biol 5: R75

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7: 438

Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V (2007) Improving gene annotation using peptide mass spectrometry. Genome Res 17: 231-239

The C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C.&nbsp;elegans: A Platform for Investigating Biology. Science 282: 2012-2018

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111

Tress ML, Bodenmiller B, Aebersold R, Valencia A (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol 9: R162

Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PL, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, Lopez G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Storling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT, Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A (2007) The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A 104: 5495-5500

Tress ML, Wesselink JJ, Frankish A, Lopez G, Goldman N, Loytynoja A, Massingham T, Pardi F, Whelan S, Harrow J, Valencia A (2008) Determination and validation of principal gene products. Bioinformatics 24: 11-17

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16: 203-211

Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41: 415-427

Valenzuela A, Talavera D, Orozco M, de la Cruz X (2004) Alternative splicing mechanisms for the modulation of protein function: conservation between human and other species. J Mol Biol 335: 495-502 van Dijk AD, Morabito G, Fiers M, van Ham RC, Angenent GC, Immink RG (2010) Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. PLoS Comput Biol 6: e1001017

van Dijk AD, ter Braak CJ, Immink RG, Angenent GC, van Ham RC (2008) Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control. Bioinformatics 24: 26-33

Verwoerd TC, Dekker BM, Hoekema A (1989) A small-scale procedure for the rapid isolation of plant RNAs. Nucleic Acids Res 17: 2362

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

Wang BB, O'Toole M, Brendel V, Young ND (2008) Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. BMC Plant Biol 8: 17

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476

Wang P, Yan B, Guo JT, Hicks C, Xu Y (2005) Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc Natl Acad Sci U S A 102: 18920-18925

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

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

Wen R, Torres-Acosta JA, Pastushok L, Lai X, Pelzer L, Wang H, Xiao W (2008) Arabidopsis UEV1D promotes Lysine-63-linked polyubiquitination and is involved in DNA damage response. Plant Cell 20: 213-227

Werner JD, Borevitz JO, Uhlenhaut NH, Ecker JR, Chory J, Weigel D (2005) FRIGIDA-independent variation in flowering time of natural Arabidopsis thaliana accessions. Genetics 170: 1197-1207

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875

Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci U S A 102: 13526-13531

Xing Y, Resch A, Lee C (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. Genome Res 14: 426-441

Yang Y, Fanning L, Jack T (2003) The K domain mediates heterodimerization of the Arabidopsis floral organ identity proteins, APETALA3 and PISTILLATA. Plant J 33: 47-59

Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB (2005) Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A 102: 2850-2855

Yu WP, Brenner S, Venkatesh B (2003) Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. Trends Genet 19: 180-183

Yuan YX, Wu J, Sun RF, Zhang XW, Xu DH, Bonnema G, Wang XW (2009) A naturally occurring splicing site mutation in the Brassica rapa FLC1 gene is associated with variation in flowering time. J Exp Bot 60: 1299-1308

Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K, Itoh T, Imanishi T, Gojobori T, Go M (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. Gene 380: 63-71

Zambelli F, Pavesi G, Gissi C, Horner DS, Pesole G (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 11: 534

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

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

Zhang J (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution 18: 292-298

Zhang JZ, Li ZM, Mei L, Yao JL, Hu CG (2009) PtFLC homolog from trifoliate orange (Poncirus trifoliata) is regulated by alternative splicing and experiences seasonal fluctuation in expression level. Planta 229: 847-859

Zhang PG, Huang SZ, Pin AL, Adams KL (2010) Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis. Mol Biol Evol 27: 1686-1697

Acknowledgements

To be honest I found this to be the most difficult section of my thesis to write. I am going to start by thanking all my (ex) colleagues for all the coffee breaks, work- and literature discussions. I wish all the current Ph.D. students the best of luck in finishing their thesis.

I want to thank my promotor Willem Stiekema for giving me the opportunity to do this Ph.D. study and for supervising the thesis. Roeland van Ham; we had many nice discussions but also a number of disagreements. In the end, I'm very grateful that you stayed committed to supervise my thesis even when you left PRI. Many thanks for all your input and feedback which resulted in the completion of this thesis.

Aalt-Jan van Dijk and Richard Immink; ADJ, I especially want to thank you for all your input in the research chapters of this thesis. I also want to thank you for all the feedback you gave me on my sometimes "crazy" ideas. Richard, thank you for your input on chapter five of this thesis. It has been an honor working with the two of you.

Especially in BioInformatics departments it is crucial that all available systems are up and running smoothly. I therefore wish to thank the SysOp team of Applied Bioinformatics for providing working servers and clusters on which I could run many of my analyses. Bas, thank you for handling all the (many) database requests and other small technical issues. Jan, thanks for all the handy linux tips and also for "forcing" me to install Dropbox because it finally saved me a lot of time.

As some people (or many) know I'm not very good at keeping my administration organized. I therefore want to thank Maria Augustijn for always knowing my PhD. administration better than me and Douwe Zuidema for always helping me with my TSP.

I also want to thank my long time (ex) roommates, Erwin, Ate, Patrick and Yiannis. Ate, good luck with finishing your thesis. Yiannis; thanks for all the nice discussions about "integrating-out" probabilities and for all the (many) drinks in the centre of Wageningen. I also thank you for your help on the final stages of my thesis. Anand, I want to thank you for all the walks around the neighborhood.

Andre and Jakub, thanks for your friendship and for being my paranymphs. Jakub, I also thank you and Eva, for all the good food and bottles of Casillero del Diablo in Utrecht. Also thanks for letting me use your computer as an ssh-client on Sunday mornings. I wish you the best of luck in Barcelona.

Finally, I want to thank both my parents for all their support. This was very important to me during the time I was working on my thesis.

List of publications

Boxma B, Ricard G, van Hoek AH, Severing E, Moon-van der Staay SY, et al. (2007) The [FeFe] hydrogenase of Nyctotherus ovalis has a chimeric origin. BMC Evol Biol 7: 230.

Severing EI, van Dijk AD, Stiekema WJ, van Ham RC (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC Genomics 10: 154.

Severing EI, van Dijk A, van Ham RC (2011) Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. BMC Plant Biol 11: 82.

Severing EI, van Dijk AD, van Ham RC. Analysis of conservation of alternative splicing indicates limited contribution to protein diversity in the plant kibgdom. *In preparation*

Severing EI, van Dijk AD, Morabito G, Busscher-Lange J, Immink RG, van Ham RC. Predicting the impact of Alternative Splicing on Plant MADS Domain Protein Function. *In revision*.

Curriculum vitae

Edouard Severing was born on the 11th of October 1977 in Dordrecht, the Netherlands. In 1977 he obtained his high school degree at the Radulphus College on Curaçao and went to study biology at the Radboud University in Nijmegen. During the specialisation phase of his studies, Edouard did his first internship at the evolutionary microbiology department headed by Johannes Hackstein and later his second internship at the comparative genomics group at the CMBI in Nijmegen headed by Martijn Huynen. In 2005, Edouard started his Ph.D. in the Applied Bioinformatics department of Plant Research International in Wageningen under the supervision of Roeland van Ham and Willem Stiekema. Currently he has a post-doc position at the laboratory of bioinformatics from Wageningen University.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: **Edouard Severing**
Date: **7 December 2011**
Group: **Applied Bioinformatics, Wageningen University & Research centre**

	1) Start-up phase	date
▶	**First presentation of your project**	
	Alternative splicing in Plants	Dec 11, 2005
▶	**Writing or rewriting a project proposal**	
	Alternative splicing in Plants	Feb 2006
▶	**Writing a review or book chapter**	
▶	**MSc courses**	
▶	**Laboratory use of isotopes**	
	Subtotal Start-up Phase	_7,5 credits*_

	2) Scientific Exposure	date
▶	**EPS PhD student days**	
	EPS PhD student day (Wageningen)	Sep 19, 2006
	EPS PhD student day (Wageningen)	Sep 13, 2007
▶	**EPS theme symposia**	
	Theme 4 symposium 'Genome Plasticity' (Wageningen)	Dec 12, 2008
	Theme 4 symposium 'Genome Plasticity' (Nijmegen)	Dec 11, 2009
▶	**NWO Lunteren days and other National Platforms**	
	NBIC congress (Ede)	Apr 24, 2006
	CBSG Workshop Genomic tools (Wageningen)	Feb 20, 2007
	NBIC congress (Amsterdam)	Apr 17-18, 2007
	BioRange (Lunteren)	Mar 05-06, 2008
	NBIC meeting (Lunteren)	Mar 17-18, 2009
	Biorange meeting Arnhem	2009
▶	**Seminars (series), workshops and symposia**	
	Evolutionary BioInformatics Symposium (UTRECT)	Nov 16, 2006
	CytoScape symposium (AMSTERDAM)	Nov 09, 2007
	Galaxy workshop	Sep 02, 2010
	EPS siminar (Hilberit)	Sep 08,2010
	PSG joint bioinformatics meetings	Jan-Jul 2006
	EPS Seminar Dr. Bas Haring	Sep 16, 2010
	EPS Seminar Dr. Adam Price	Sep 17, 2010
▶	**Seminar plus**	
▶	**International symposia and congresses**	
	Second Benelux BioInformatics Conference, Wageningen, NL	Oct 17-18, 2006
	ECCB/ISMB 2007 (Vienna, Austria)	Jul 21-25, 2007
	ECCB/ISMB 2007 / Special Interest Group Meetings	Jul 19-20, 2007
	Benelux BioInformatics Conference, Maastricht	Dec 15-16, 2008
▶	**Presentations**	
	PSG joint bioinformatics (oral presentation)	2006
	BU meeting (oral)	2007 and 2009
	Benelux Bioinformatics Conference Maastrict (poster)	Dec 15-16, 2008
	Biorange meeting Arnhem (oral)	2009
▶	**IAB interview**	Dec 05, 2008
▶	**Excursions**	
	Subtotal Scientific Exposure	_14,2 credits*_

	3) In-Depth Studies	date
▶	**EPS courses or other PhD courses**	
	Molecular phylogenies: reconstruction & interpretation	Oct 15-19, 2006
	Multivariate Analysis	Apr 19-27, 2006
▶	**Journal club**	
	Literature discussion within the Bioinformatics group	2005-2009
▶	**Individual research training**	
	Subtotal In-Depth Studies	_6,0 credits*_

	4) Personal development	date
▶	**Skill training courses**	
	Phd competence assessment	Apr 11, 2006
	Interpersonal communication	Nov 2007
	Career Orientation	Jun 2009
	Information literacy PhD + EndNote Introduction	Dec 07, 2009
▶	**Organisation of PhD students day, course or conference**	
	CBSG Workshop Genomic tools (Wageningen)	Feb 20, 2007
▶	**Membership of Board, Committee or PhD council**	
	Subtotal Personal Development	_4,5 credits*_

TOTAL NUMBER OF CREDIT POINTS*	32,2

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

* A credit represents a normative study load of 28 hours of study.