Isoform-Specific Expression During Embryo Development in Arabidopsis and Soybean

Delasa Aghamirzaie

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Ruth Grene, Chair

Eva Collakova, Co-Chair

Lenwood S. Heath

Song Li

Jason Holliday

April 20th, 2016

Blacksburg, VA

Keywords: Alternative splicing, data analysis, bioinformatics, transcriptomics, RNA-Seq, noncoding RNAs, machine learning, computational biology

Isoform-Specific Expression During Embryo Development in Arabidopsis and Soybean

Delasa Aghamirzaie

ABSTRACT

Almost every precursor mRNA (pre-mRNA) in a eukaryotic organism undergoes splicing, in some cases resulting in the formation of more than one splice variant, a process called alternative splicing. RNA-Seq provides a major opportunity to capture the state of the transcriptome, which includes the detection of alternative spicing events. Alternative splicing is a highly regulated process occurring in a complex machinery called the spliceosome. In this dissertation, I focus on identification of different splice variants and splicing factors that are produced during Arabidopsis and soybean embryo development. I developed several data analysis pipelines for the detection and the functional characterization of active splice variants and splicing factors that arise during embryo development. The main goal of this dissertation was to identify transcriptional changes associated with specific stages of embryo development and infer possible associations between known regulatory genes and their targets. We identified several instances of exon skipping and intron retention as products of alternative splicing. The coding potential of the splice variants were evaluated using CodeWise. I developed CodeWise, a weighted support vector machine classifier to assess the coding potential of novel transcripts with respect to RNA secondary structure free energy, conserved domains, and sequence properties. We also examined the effect of alternative splicing on the domain composition of resulting protein isoforms. The majority of splice variants pairs encode proteins with identical domains or similar domains with truncation and in less than 10% of the cases alternative splicing results in gain or loss of a conserved domain. I constructed several possible regulatory networks that occur at specific stages of embryo development. In addition, in order to gain a better understanding of splicing regulation, we developed the concept of co-splicing networks, as a group of transcripts containing common RNA-binding motifs, which are co-expressed with a specific splicing factor. For this purpose, I developed a multi-stage analysis pipeline to integrate the co-expression networks with *de novo* RNA binding motif discovery at inferred splice sites, resulting in the identification of specific splicing factors and the corresponding cis-regulatory sequences that cause the production of splice variants. This approach resulted in the development of several novel hypotheses about the regulation of minor and major splicing in developing Arabidopsis embryos. In summary, this dissertation provides a comprehensive view of splicing regulation in Arabidopsis and soybean embryo development using computational analysis.

Isoform-Specific Expression During Embryo Development in Arabidopsis and Soybean

Delasa Aghamirzaie

GENRAL ABSTRACT

Almost every precursor mRNA (pre-mRNA) in a eukaryotic organism undergoes splicing, in some cases resulting in the formation of more than one splice variant, a process called alternative splicing. Alternative splicing occurs as the consequence of the action of a complex machinery called the spliceosome containing several splicing factors and splicing related proteins. Alternative splicing is a highly regulated process, yielding splice variants with varying coding potentials, and proteins with altered number and types of functional domains. I developed several RNA-Seq data analysis pipelines at the isoform level to improve the current understanding of splicing in Arabidopsis and soybean embryo development, allowing for hypothesis generation concerning the regulation of seed development. The findings of this dissertation has shed light on the identification of major players of splicing including different splicing factors and their potential targets during embryo development in plants.

Dedicated to my husband, Mahdi Nabiyouni, for his continuous support and love. You've helped me in more ways than anyone else.

To my mom and my little brother, whom I love the most. Mom, you have inspired me the most in my life.

And in memory of my father, who left fingerprints of grace on my life. You shan't be forgotten.

Acknowledgments

This work would not have been possible without the support and encouragement of many people throughout my 4 years of PhD journey between years 2012 and 2016. First, I'd like to thank my advisor, Ruth Grene. To say she has only been my advisor is definitely an understatement. I express my most sincere and gratitude to her who trusted me and provided me with this opportunity to work on the project despite my little knowledge in biology. She supported me like a mother during years of my PhD. Not only she taught me biology and how to do a good science, but also she taught me how to be a good person before being a scientist. She taught me to be humble no matter where I am and what I achieve in my life.

I am extremely grateful to my co-advisor, Eva Collakova, who has been more than a co-advisor and a PI on the project. She has been my dear friend. We had lunch almost every week while she listened to all my concerns and problems. She has a passion for learning new methods. Without her support, I would not be able to accomplish this dissertation.

I am grateful to Lenwood Heath and Song Li, who guided me through computational data analysis in this project, without their advice and help I would not be able to learn deep data analysis of transcriptomics data. I also like to thank Jason Holliday, my committee member, who always had the brightest ideas about how to improve my work. I would like to thank Dhruv Batra who advised me in the development of the CodeWise.

I'd like to thank Erhan Bilal, my mentor in IBM Research, who patiently taught me about cancer drugs and neural networks while I was a summer intern there.

I'd like to thank Andrew Schneider and Yihui Fang, former graduates of Collakova lab for providing me the transcriptomics data for embryo development in Arabidopsis and soybean and doing experimental validations for my predictions.

I'd like to thank Haitham Elmarakeby, Doaa Altarawy, and Hong Tran, my teammates in the AstraZeneca-Sanger drug combination DREAM challenge. They were the most hard-working teammates that I ever had. I learned so much from each of them.

I'd like to thank Mostafa Arefiyan. He was a miracle that came to VT from Brown University and became my lab mate during the last semester. He listened to all my complaints one by one. He also taught me lots of new things about deep learning and computer vision, which I am going to work on for my post doc.

I'd also like to thank my fellow colleagues in TPS, PPWS, and GBCB, Elizabeth Bush, Elizabeth A. Grabau, Boris Vinatzer, Guillaume Pilot, Judy Fielder, Dennie Munson, and David Baven. I like to thank VT students, Karthik Raja Velmurugan, Shuchi Wu, Julien Besnard, Yu Shi, Gunjune Kim, Ying Ni, and Curtis Klumas.

I'd also like to thank all my friends in Blacksburg who were my family here, while we were far from our country. Specially, I want to thank Elmira Hamidi and Behzad Hamidi, who supported me in every step I took in these four years. I also would like to thank my friends, Soheil Kamalzare, Mohammad Khosravi, Narges Dorratoltaj, Pouya Bashivan, and Golnaz Badr.

I'd like to thank my mom and my dad, I owe them everything I have in my life, for their continuous support for every decisions I made and for never telling me I had limits on what I could do. I would never have made it without their encouragement and support. My dad passed away when I was 16 years old and I know how much he wanted me to be successful. Since then, my mom raised my siblings and me with a huge courage in all ups and downs in our life. Mom, I love you the most and I hope I can be a strong woman like you in my life. You are my true role model. Dad, I miss you so much and I will remember you in every step that I take in my life.

Most importantly, I would like to express my sincere gratitude to my husband, Mahdi, for her love, support, and patience. He stood by my side during this journey, before, and hopefully after that as my true best friend. I am so grateful and fortunate to have him in my life. I would also like to thank my in-laws, first for raising such as awesome guy, and second for their support and love.

Delasa Aghamirzaie,

April 2016

Blacksburg, VA

Attributions

Several colleagues contributed to the projects, research, writing, and editing of the manuscript in this dissertation.

All chapters:

Eva Collakova, Ph.D. (Department of Plant Pathology, Physiology and Weed Science) is currently an Associate Professor at PPWS and is the corresponding author on the manuscript. She aided in project development, writing, and editing of the manuscripts.

Ruth Grene, Ph.D. (Department of Plant Pathology, Physiology and Weed Science) is a Professor at PPWS and is a co-author of the manuscript. She aided in project development, writing, and editing of the manuscripts.

Lenwood Heath, Ph.D. (Department of Computer Science) is a professor and co-author on the manuscript. He contributed to the computational analyses.

Chapter 2, 3:

Yihui Fang (Department of Plant Pathology, Physiology and Weed Science) was a master student in PPWS and is a co-author of the manuscript. He performed sample preparations for RNA-Seq analysis of soybean embryos and aided in biological interpretation of the results.

Curtis Kumas (Genetics, Bioinformatics, and Computational Biology), Farzaneh Tabataba (Department of Computer Science) aided in the computational data analysis and contributed to the writing of the manuscript.

Chapter 3:

Mahdi Nabiyouni (Department of Computer Science) aided in visualization of splicing graphs and contributed to the writing of the manuscript.

Chapter 4, 5:

Dhruv Batra, Ph.D. (Bradley Department of Electrical and Computer Engineering) is currently an Assistant Professor in ECE department. He aided in the development of CodeWise and contributed to the writing of the manuscript.

Chapter 6:

Andrew Schneider, Ph.D. (Department of Plant Pathology, Physiology and Weed Science) performed all the experiments and performed biological interpretation of the results.

Chapter 7:

Song Li, Ph.D. (Crop and Soil Environmental Science) is currently an Assistant Professor in CSES department. He aided in the development of co-splicing networks and contributed to the writing of the manuscript.

Chapter 8:

Karthik Raja Velmurugan (Genetics, Bioinformatics, and Computational Biology) is a graduate student in GBCB department. He aided in the data analysis in Expresso Web server and contributed to the writing of the manuscript.

Shuchi Wu, Ph.D. (Department of Horticulture) contributed to the data preparation and contributed to the writing of the manuscript.

Doaa Altarawy (Department of Computer Science) aided in the development of the Expresso Web server and contributed to the writing of the manuscript.

Table of Contents

1 Introduction

The major goals of this dissertation were to apply in-house and publicly available bioinformatics tools to investigate transcript population obtained from RNA-Seq of developing Arabidopsis and soybean embryos. I addressed the following research aims:

1. To identify transcriptional changes associated with the major stages of embryo development;
2. To infer possible associations between known regulatory genes and their targets;
3. To identify products of alternative splicing including coding potential of transcripts (using CodeWise) and domain composition in resulting proteins;
4. To infer possible regulatory networks that occur at specific stages of embryo development;
5. To characterize the role of *VAL1* through computational analysis, a regulatory gene in seed development
6. To group co-expressed and co-spliced transcripts according to specific stages of seed development; and
7. To use *de novo* motif discovery to identify splicing regulatory regions for specific splicing factor proteins.

In addition, a Web server and database, Expresso, was developed to process and integrate curated Arabidopsis ChIP-Seq data.

1.1 Arabidopsis and soybean embryo development

Seeds are an important source of food, feed, biodiesel, and chemicals, being rich in oils, proteins, and carbohydrates (Eva Collakova et al., 2013). Seed development encompasses a series of developmental, metabolic, and physiological processes, controlled by regulatory events, which are only partially understood at the molecular level. These processes take place during the reproductive through maturation stages. Three main processes occur during seed development, which maintains seed viability

until germination. These processes are (i) the development of proper cell and tissue identity during cell differentiation, (ii) the accumulation of reserves to fuel seed germination, and (iii) the acquisition of desiccation tolerance and dormancy to withstand unfavorable conditions that the seed encounters prior to germinating. The early stages of embryo development involve predominantly cell division and differentiation. These processes require a constant active supply of maternally provided oxidizable substrates before the embryonic cells become fully differentiated and photosynthetically active. Nutrient supply by maternal source tissues continues even in green seeds where "photoheterotrophic" metabolism takes place. During the seed-filling stage, cell elongation and accumulation of seed storage compounds constitute the major developmental and metabolic processes. As the water content decreases and the seed biomass increases during the late stages of seed filling, hormonally-regulated seed desiccation- and dormancy-related processes become active.

1.2 Transcriptomics

Many genes in eukaryotes produce more than one isoform as a consequence of alternative splicing (Kelemen et al., 2013). The notion of getting an extensive snapshot of the state of the transcriptome under a given condition is now an accepted part of contemporary experimental biology. Transcriptomes can be captured using microarrays or RNA-Seq. The disadvantage of microarrays is that, in most cases, the sequences present on the solid support represented exclusively protein coding sequences in organisms, and oftentimes, even the nucleic acid populations are incomplete. Furthermore, because of the technology, cross-hybridization among closely related sequences could occur. The advantage of RNA-Seq technology is the lack of dependence on any prior knowledge of the genome involved, allowing the possibility of the discovery of transcriptional activity within both annotated and un-annotated regions of the genome. Transcripts can be coding or noncoding, genic or intergenic, and sense or antisense.

RNA-Seq also facilitates identification of alternative splicing events as a major source of transcript diversity. Alternative splicing results in the production of more than one transcript or splice variant from a single coding locus. Splice variants can be coding

or noncoding. Coding splice variants have the potential to be translated into proteins or regulatory peptides that can contain, or lack, known domains important for function, regulation, interaction with other molecules, and/or subcellular localization (James et al., 2012; E. I. Severing et al., 2012). In contrast, noncoding splice variants may perform regulatory functions or they may be subjected to degradation or nonsense mediated decay (NMD) processes. Deep RNA-Seq (often >7X coverage) is necessary for accurate detection of splice variants, single-nucleotide polymorphisms, small RNAs, antisense transcripts, and long noncoding RNA (lncRNA) (Clark et al., 2011). Many transcripts including lncRNAs, small RNAs, and antisense transcripts are usually expressed at very low levels, which make their detection only possible via deep RNA-Seq approaches.

1.3 RNA-Seq data analysis principles

The enormous quantities of transcriptomic data obtained from RNA-Seq studies require computational expertise for organization, mining, and analysis of the data. Various tools have been developed and are still being developed to address different aspects of the RNA-Seq data analysis problem, including mapping the reads to the genome, transcript assembly and differential expression analysis (Wolf, 2013). RNA-Seq data analysis workflows are divided into reference-based methods and *de novo* assembly methods based on availability of the reference genome. Reference-based methods start with mapping the reads to the reference genome, while *de novo* assembly methods assemble reads into contigs based on their overlaps. Various pipelines and workflows are available for both methods (Martin & Wang, 2011). I used several of the widely used tools such as the Tuxedo Suite pipeline (C. Trapnell, Roberts, et al., 2012), StringTie (Mihaela Pertea et al., 2015), and Limma (Ritchie et al., 2015) for different stages of RNA-Seq data analysis of Arabidopsis and soybean embryo development (Aghamirzaie et al., 2013; Eva Collakova et al., 2013; Schneider et al., 2015). These tools are briefly explained here.

- **Tophat** (C. Trapnell, Pachter, & Salzberg, 2009). Tophat is a fast splice junction mapper for RNA-Seq reads. Tophat and its later version, Tophat2 (D. Kim et al.,

2013), have the ability to identify novel splice sites through direct mapping to known transcripts. Tophat2 uses a multi-stage pipeline for mapping the reads to the genome. First, at the genome alignment stage, it maps reads spanning a single exon. The reads that span multiple exons are unmapped at this point. Next, at the splice-aware alignment stage, multi-exon spanning reads are split into segments. Segments are aligned to the genome and potential novel splice sites are identified, based on segment alignment results. Sequences flanking a splice site are identified and segments are aligned to them. Aligned segments are stitched together.

- **Cufflinks** (Roberts, Pimentel, Trapnell, & Pachter, 2011; C. Trapnell, Roberts, et al., 2012). Cufflinks uses the reads that have been mapped to the genome in Tophat to assemble transcripts. Each pair of fragment reads is a single alignment. The first step in Cufflinks is finding pairs of incompatible fragments, the fragments that originated from distinct mRNA isoforms. Then, it builds an overlap graph, in which each fragment is a node. Fragments are connected if they are compatible and if their alignments overlap in the genome. Cufflinks follows Dilworth's Theorem, which states that the minimum number of transcripts is equivalent to the number of mutually incompatible reads. Cufflinks parses the overlap graph to produce a minimal set of paths covering all transcripts. At the next step, abundance of the assembled transcripts is estimated using FPKM (fragments per kilo base of transcript per million fragments mapped).

- **Cuffmerge** (C. Trapnell, Roberts, et al., 2012). Cufflinks provides a transcriptome assembly for each condition in the experiment. These different transcriptome assemblies by Cufflinks or any other transcript assembly tool, in GTF format, can be combined together using Cuffmerge, which is available in the Cufflinks package, to gain a unique reference transcriptome (in GTF format) specific to the data. This newly assembled transcriptome contains known (present in the reference genome) and novel (not present in the reference genome and specific to this RNA-Seq data set)

transcripts. Novel transcripts include novel isoforms of known genes, antisense transcripts, and intergenic transcripts.

- **Cuffcompare** (C. Trapnell, Roberts, et al., 2012). Since transcriptome/genomic annotations are incomplete, RNA-Seq studies reveal novel genes and novel transcripts. Some of the transcripts may be expressed at low levels, so they may be partially assembled. The Cuffcompare tool uses a reference genome together with a transcriptome assembly from Cufflinks or merged transcriptome from Cuffmerge and assigns class codes to the different inferred transcripts. These class codes serve as a classification basis for the various assembled transcripts with reference to transcripts with well-characterized splicing patterns (class "="). It is noteworthy that assignment of class codes in Cuffcompare is prioritized. For example, when an isoform has a novel splice junction, it is classified as class "j", although its structure may fall into other lower priority classes as well. Class "j" transcripts are potentially novel isoforms, in that they have at least one novel splice junction and at least one splice junction shared with the reference transcript. Class "o" transcripts are assembled transcripts that show exonic overlap with the reference transcript, but do not fall into other higher priority class such as "c" or "j". Class "c" stands for "contained" and is used when a transcript has a high exonic overlap with a known transcript. Transcripts in classes "x" and "s" have exonic and intronic overlap, respectively, with the reference transcript on the opposite (antisense) strand. Class "i" transcripts are those for which some sequence falls entirely within a reference intron. These transcripts are representative of intron retention events, in which the transcript did not fall into other higher priority classes.
- **Cuffdiff** (Trapnell et al., 2013). Cuffdiff is part of the Cufflinks package. It takes aligned reads (Tophat output) and identifies differentially expressed transcripts in the requested conditions. Cuffdiff assumes that the variability of a transcript count depends both on its expression and splicing structure. Cuffdiff models variability in the number of fragments generated by each transcript across replicates. Cuffdiff also provides accurate abundance estimation at the isoform level. It captures uncertainty in a transcript fragment count as a beta distribution and the over-dispersion in this count

with a negative binomial and mixes these two distributions together. The resulting mixture is a negative binomial distribution.

- **StringTie** (Mihaela Pertea et al., 2015). StringTie uses a genome-guided transcript assembly along with some principles from *de novo* assembly to improve transcript assembly. Cufflinks does not consider transcript abundance while parsing the overlap graph. Therefore, it may not always reconstruct the correct set of isoforms. StringTie assembles transcripts and estimates their abundances simultaneously. StringTie groups reads into clusters, then creates a splice graph for each cluster. It identifies transcripts from splice graphs and creates a specific flow in the network for each transcript. It estimates transcript abundance based on a maximum flow algorithm. Therefore, transcript abundance is estimated from the path in the splice graph that has the heaviest coverage.
- **Limma** (Ritchie et al., 2015; Gordon K Smyth, 2005). Limma is an R/Bioconductor package for performing differential expression analysis of gene expression data. Limma was originally developed for microarray data analysis and has now been expanded to include the analysis of RNA-Seq data (Gordon K Smyth, 2005). The main difference between Limma and other differential expression analysis tools is that Limma uses linear models to analyze the entire data set, rather than performing individual pairwise comparisons. Linear models include time course analysis and regression splines. Limma fits separate models per gene or transcript. Limma uses parametric empirical Bayes for modeling mean-variance among genes/transcripts. For differential analysis of RNA-Seq data, Voom normalization needs to be performed prior to Limma to produce mean-variance for transcripts. Voom uses normal linear modeling strategies to estimate the mean-variance for raw counts (Law, Chen, Shi, & Smyth, 2014).

1.4 Splicing mechanism

Splicing mainly occurs in the nucleus using a complex machinery called the spliceosome which includes a large number of small nuclear ribonucleoproteins (snRNPs) capable of interacting with small nuclear RNAs (snRNAs), proteins important for structural integrity of spliceosome, and proteins called splicing factors that provide specificity of splicing (Clancy, 2008). During the splicing process, some parts of

transcripts, mainly introns or parts of exons, are removed from the primary precursor RNAs (pre-mRNAs) at splice sites. Splice sites are present at 5'- and 3'- ends of introns and exons, which contain some conserved sequences that may recruit splicing related proteins (SRPs) and snRNPs to specific RNA binding sites. Most commonly, splice sites contain the dinucleotide GU at their 5' end, and AG at their 3' end (Clancy, 2008). Interaction of SRPs with exonic or intronic enhancers and silencers through binding or unbinding to *cis*-regulatory elements can positively or negatively affect the splicing process (Clancy, 2008).

1.5 RNA motif search

It is well established that presence of common dinucleotide sequences at the 5' and 3' end of the splice sites is necessary but not sufficient for accurate detection of splice junctions, and existence of *cis*-regulatory elements at the exonic and intronic splice sites are necessary for splicing to happen (Chasin, 2008). Splicing factors bind to these short cis-regulatory elements, which can positively or negatively affect the splicing process. There are several tools available to predict *cis*-regulatory elements, including exonic/exonic splicing enhancers/silencers. Some tools such as SFmap (Paz, Akerman, Dror, Kosti, & Mandel-Gutfreund, 2010), SpliceAid (Piva, Giulietti, Nocchi, & Principato, 2009) and SROOGLE (Schwartz, Hall, & Ast, 2009) search for experimentally validated motifs in a given RNA sequence. The MEME suite can also be used for identification of de novo motifs in the splice junction (Bailey, Johnson, Grant, & Noble, 2015). In addition, the TomTom tool available in the MEME suite can be used to search for known motifs in the RNA sequences (Bailey et al., 2015). RegRNA (H.-Y. Huang, Chien, Jen, & Huang, 2006) finds motifs based on their homology to known motifs. In general, demonstrating RNA-protein interactions experimentally is feasible using different crosslinking and immunoprecipitation (CLIP) methods. However, CLIP methods don't work for all systems and often affected by mild sequence bias. Although search for known motifs and the use of CLIP data are important for finding *cis*-regulatory elements, they are limited to a few very well-known splicing factors such as SF2/ASF, SC35, SRp40, SRp5m and PTB (Cereda et al., 2014). Some other methods such as RNAmotifs (Cereda et al., 2014), have been developed to find multivalent RNA motifs for differentially spliced exons. Multivalent RNA motifs are defined as the motifs that

"enable RBPs to achieve high-affinity binding by cooperative interactions between multiple RNA-binding domains and the clustered short RNA motifs" (Cereda et al., 2014).

1.6 Alternative splicing

Alternative splicing is a post-transcriptional process that produces more than one mRNA from a single pre-mRNA. Alternative splicing is a highly regulated cellular process, giving rise to the production of multiple transcripts – either coding or noncoding – from a single coding locus (Mo Chen & Manley, 2009). During alternative splicing, different combinations of exons and introns may be included within, or excluded from, the final mRNA, resulting in a splice variant. Alternative splicing can produce different coding isoforms that may code for different proteins or may produce noncoding RNAs through premature stopping of translation (Figure 1.1). noncoding RNAs might be involved in some regulatory mechanisms at the RNA level or they might be subject to nonsense mediated decay (NMD) (Jeremy E Wilusz & Wilusz, 2014). Alternative splicing is categorized into five basic models: (i) exon skipping, (ii) mutually exclusive exons, (iii) alternative donor site, (iv) alternative acceptor site, and (v) intron retention (E. Kim, Magen, & Ast, 2007). Accurate identification of splice variants and their abundance is one of the current challenges of RNA-Seq workflows (R. Liu, Loraine, & Dickerson, 2014). Several tools are currently available for the identification of alternative splicing from high throughput sequencing technologies: (i) spliced aligners such as TopHat (D. Kim et al., 2013), MapSplice (K. Wang et al., 2010), and SpliceMap (Au, Jiang, Lin, Xing, & Wong, 2010) are able to identify splice sites from reads, (ii) spliced assemblers such as Cufflinks (C. Trapnell, Roberts, et al., 2012), Scripture, and Trinity are able to assemble mapped reads into splice variants using reference genome as a guide or independent of reference genome via finding overlaps between reads, and (iii) differential splicing analysis tools such as DEXSeq (Anders, Reyes, & Huber, 2012).

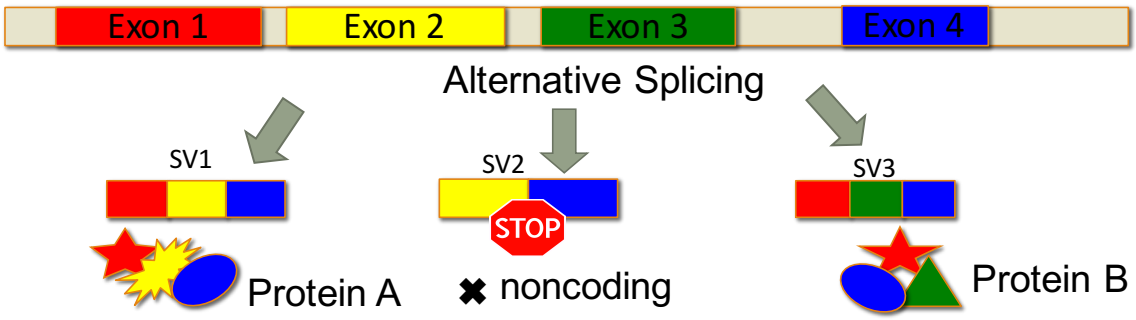

Figure 1.1: Alternative splicing can produce splice variants with different coding potentials. Coding SVs can produce different proteins with different domains. SVs can be noncoding due to the presence of premature stop codons on the final transcript.

1.7 Long noncoding RNAs

Functions of coding genes have been studied for many years. However, efforts to characterize the function of noncoding transcripts, particularly lncRNAs and long intergenic noncoding (lincRNAs), have only been started recently. lncRNAs have been implicated in many diseases including cancer progression, and the regulation (often inhibition) of a diverse array of biological processes, such as dosage compensation, cell cycle control, and development (Mercer & Mattick, 2013; Rinn & Chang, 2012). Some classes of short ncRNAs (<200 nucleotides in length) are already accepted as fundamental players in gene regulation; these include small interfering RNAs (siRNAs), microRNAs (miRNAs), and PIWI-interacting RNAs (piRNAs) (Pelechano & Steinmetz, 2013). There is evidence suggesting that lncRNAs can act either through short peptides or as functional regulatory RNAs (Chew et al., 2013; Guttman, Russell, Ingolia, Weissman, & Lander, 2013; Lindsey, Casson, & Chilley, 2002). Despite the fact that there is no consensus in the literature from ribosome profiling data as to whether these lncRNAs encode short peptides or not, it has been apparent that mRNA is involved in the regulation of gene expression through different mechanisms, such as chromatin modification and RNA-protein interactions (Guttman et al., 2013; Ponjavic, Ponting, & Lunter, 2007; Jeremy E. Wilusz, Sunwoo, & Spector, 2009; L. Yang, Froberg, & Lee, 2014). These ncRNAs perform their regulatory functions primarily through transcriptional interference, sense and antisense hybridization, interactions with RNA-binding proteins, and/or serving as precursors for small regulatory RNAs (Jeremy E. Wilusz et al., 2009; L. Yang et al., 2014).

In plants, noncoding transcripts have been reported to be involved in the regulation of development, flowering, root meristem development, and responses to stress, e.g., cold (Amor et al., 2009; Boerner & McGinnis, 2012; Di et al., 2014; W. Zhang et al., 2014). Two of the well-known examples of functional noncoding RNAs in plants are COLDAIR and COOLAIR (Cregg, Murphy, & Mardinoglu, 2012; Kelemen et al., 2013; J. Liu et al., 2012). COLDAIR is a long intronic ncRNA implicated in the epigenetic repression of FLC during vernalization, and COOLAIR is a ncRNA fully encompassing FLC in the antisense direction, which is alternatively polyadenylated and spliced (Ietswaart, Wu, & Dean, 2012).

1.8 Available tools for coding potential assessment

- **CPC** (Kong et al., 2007). CPC is a support vector machine (SVM) classifier for assessing the coding potential of transcripts. CPC is trained using six sequence features. The first three features assess the extent and quality of an ORF in a transcript. CPC uses framefinder to identify the longest ORF within three forward frames. It parses framefinder output and reports log-odds score, coverage and integrity of the predicted ORF. The integrity of an ORF is defined as an ORF with a start codon and an in-frame stop codon. The next three features include parsing BLASTX results against the UniProt database using an E-value cutoff of 1e-10. The number of hits, the hit score, and the frame score are used as three other features. The hit score was defined as sum(log(E-value)) for all three frames. The frame score is variance of sum(log(E-value)). CPC is trained using 5,610 protein coding transcripts from Swiss-Prot eukaryotic proteins and 2,670 noncoding RNAs from RNAdb and NONCODE. CPC showed 95.7% accuracy on a test data set. However, CPC was developed in 2007 and has not been updated afterward.
- **PhyloCSF** (Lin, Jungreis, & Kellis, 2011). PhyloCSF is a comparative genomics-based tool. It determines whether a sequence is coding or noncoding based on phylogenetic codon models. Based on multi-species nucleotide alignment, it determines how likely a sequence is to be a conserved protein coding sequence. PhyloCSF relies on genome alignment models for estimation of branch lengths, codon frequencies and codon substitutions.

- **CPAT** (Liguo Wang et al., 2013). CPAT is a logistic regression based tool trained on four features: ORF size, ORF coverage, the Fickett testcode statistic and hexamer usage bias. The Fickett testcode depends on nucleotide composition and codon usage bias in coding and noncoding RNAs. Hexamer usage bias depends on adjacent amino acids in proteins. CPAT is an alignment free coding potential assessment tool and is available for human, mouse, fly, and zebrafish. Because it is alignment free, it performs very rapidly compared with alignment-based methods such as CPC and PhyloCSF. A considerable number of genic lncRNAs have overlaps to coding RNAs. Therefore, similarity search scores will be high for these transcripts.
- **iSeeRNA** (Sun et al., 2013). iSeeRNA is an SVM-based classifier for the detection of lincRNAs in human and mouse. It was trained with 5,079 human lincRNAs, 24,960 human's coding RNAs, 889 mouse' lincRNA, and 15,121 mouse coding RNAs. iSeeRNA is built using 10 features. The first feature is the conservation score, which is calculated using averaging phastCons (Siepel et al., 2005) score files from UCSC per transcript. The second and third features are ORF length and ORF proportion. Seven other features constitute seven di- or tri-nucleotide sequences (GC, CT, TAG, TGT, ACG and TCG). iSeeRNA is available in a Web server and accepts GFF/GTF or BED formats.

1.9 Overview and outline

In Chapter 2, I report on detailed temporal transcriptional and metabolic changes in soybean embryos to gain a systems biology view of developmental and metabolic changes that occur during seed development. The first major transition involved a switch between heterotrophic and photoheterotrophic metabolism in dividing and elongating cells, respectively. The second transition involved the onset of maturation and desiccation tolerance during seed filling. Clustering of metabolite and transcriptomic data analysis revealed clusters of functionally related metabolites and transcripts active in these different developmental and metabolic programs.

In Chapter 3, I report on the analyses of an RNA sequencing data set which was used to provide comprehensive information about transcriptional and post-transcriptional events that take place in developing soybean embryos. Results of the analysis lead to the identification of different classes of alternatively spliced isoforms and corresponding

changes in their levels on a global scale that occur during soybean embryo development. Alternative splicing was associated with transcripts involved in various metabolic and developmental processes, including central carbon and nitrogen metabolism, induction of maturation and dormancy, and splicing itself.

In Chapter 4, I report on the development of CodeWise, a plant-specific coding potential calculator for the prediction of the coding status of transcripts. CodeWise uses several features, including sequence, conserved domains, and RNA secondary structure, to determine whether a transcript is coding or noncoding.

In Chapter 5, I present an integrated computational framework to predict isoform-specific functions of plant transcripts. This framework includes CodeWise, together with other tools, which focus on global sequence similarity, functional domains, the construction of co-expression networks, and the inference of possible signaling pathways. This framework was applied, first, to all detected transcripts (103,106), out of which 13% was predicted by CodeWise to be noncoding RNAs in developing soybean embryos. It was then used to investigate the role of AS during soybean embryo development. A population of 2,938 alternatively spliced and differentially expressed splice variants was analyzed and mined with respect to timing of expression. Inferred signaling pathways included abscisic acid- and FUSCA3-related transcripts, several of which were classified as noncoding and/or antisense transcripts. Noncoding and antisense transcripts likely play important regulatory roles in seed maturation- and desiccation-related signaling in soybean.

In Chapter 6, I report on a study focusing on the previously unknown role of VAL1 in seed development and metabolism. The *VAL* gene family encodes repressors of the seed maturation program in germinating seeds, although they are also expressed during seed maturation. I present the results of an analysis of an RNA-Seq data set obtained from the developing embryos of two Arabidopsis genotypes, Col-0 and a *val1* mutant. Reverse genetics revealed that *val1* mutant seeds accumulated elevated levels of proteins compared to the wild type, suggesting that VAL1 functions as a repressor of seed metabolism. None of the transcripts encoding the core well studied regulatory network

were affected in the mutants. Instead, activation of *VAL1* appears to result in the repression of a subset of seed maturation genes downstream of the core regulators.

Chapter 7 reports on the analysis of co-splicing events on developing Arabidopsis embryos to understand splicing regulation in spliceosome. I developed a multi-stage pipeline to integrate co-expression and *de novo* motif discovery to infer co-splicing networks in Arabidopsis developing embryos.

Chapter 8 reports on development of Expresso database and Web server as a tool for integration of available ChIP-Seq data *in Arabidopsis thaliana*.

Chapter 9 summarizes the dissertation with the significance of the work and future prospects.

2 Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine max) Embryos[1]

This chapter is produced from (Eva Collakova et al., 2013), which is an open access journal.

Collakova, E., **Aghamirzaie, D.**, Fang, Y., Klumas, C., Tabataba, F., Kakumanu, A., . . . Grene, R. (2013). Metabolic and transcriptional reprogramming in developing soybean (*Glycine max*) embryos. *Metabolites, 3*, 347-372. doi:10.3390/metabo3020347

Abstract. Soybean (*Glycine max*) seeds are an important source of seed storage compounds, including protein, oil, and sugar used for food, feed, chemical, and biofuel production. We assessed detailed temporal transcriptional and metabolic changes in developing soybean embryos to gain a systems biology view of developmental and metabolic changes and to identify potential targets for metabolic engineering. Two major developmental and metabolic transitions were captured enabling identification of potential metabolic engineering targets specific to seed filling and to desiccation. The first transition involved a switch between heterotrophic and photoheterotrophic metabolism in dividing and elongating cells, respectively. The second transition involved the onset of maturation and desiccation tolerance during seed filling. Clustering of metabolite and transcriptomic data analysis revealed clusters of functionally related metabolites and transcripts active in these different developmental and metabolic programs. The gene clusters provide a resource to generate predictions about the associations and interactions of unknown regulators with their targets based on "guilt-by-association" relationships. The inferred regulators also represent potential targets for future metabolic engineering of relevant pathways and steps in central carbon and nitrogen metabolism in soybean embryos and drought and desiccation tolerance in plants.

[1] Delasa Aghamirzaie was involved in all the computational data analysis. Yihui Fang was involved in growing soybean seeds and preparing RNA-Seq libraries. Lenwood Heath aided in the computational analysis. Ruth Grene and Eva Collakova performed all the biological data mining of the results.

Keywords: central carbon and nitrogen metabolism; plant metabolic engineering; RNA sequencing; seed storage compounds; soybean; systems biology; transcriptomics; untargeted and targeted metabolomics

2.1 Introduction

Seed is an important source of food, feed, biodiesel, and chemicals, because it is rich in oils, proteins, and carbohydrates (Clemente & Cahoon, 2009; Weselake et al., 2009). These products are referred to as seed storage reserves, as they represent carbon-, nitrogen-, and energy-rich molecules needed for seed germination before the onset of photosynthesis in seedlings (Eastmond & Graham, 2001; Graham, 2008; Penfield, Graham, & Graham, 2005). Oil and carbohydrates provide carbon and energy, while the seed storage proteins are the major source of nitrogen during germination. Typically, soybean seeds contain about 18% of oil (triacylglycerols), 42% of protein (2S albumins and 7S and 11S globulins), and other components, including carbohydrates, and storage and cell-wall polysaccharides, comprising the rest of the seed biomass (Clemente & Cahoon, 2009; D.W. Meinke, Chen, & Beachy, 1981). However, seed composition is a variable trait. For instance, oil content can vary from 6.5% to 28.7% in dry soybean seeds depending on the plant genetic background and growth conditions (Weselake et al., 2009), suggesting that central carbon metabolism (CCM) leading to the accumulation of seed storage compounds in developing seeds is quite flexible and should be amenable to metabolic engineering.

Seed development in general encompasses a chronological series of developmental, metabolic, and physiological processes, controlled by relevant, and only partially understood, regulatory events. In soybean, these processes take place during the reproductive R3 (beginning pod) through R8 (full maturity) developmental stages at the whole plant, organ, tissue, cell, and molecular levels (Fehr, Caviness, Burmood, & Pennington, 1971; Brandon H. Le et al., 2007; D.W. Meinke et al., 1981). Development and metabolism in an oilseed such as soybean is relatively simple compared to events taking place in other parts of the plant, since the sole purpose of the seed is species propagation. This translates into the involvement of three main processes during seed development that maintain seed viability until germination. These processes include: (i) the development of proper cell and tissue identity during cell differentiation and

maintaining genetic information about the species, (ii) the accumulation of reserves to fuel seed germination, and (iii) the ability to withstand unfavorable conditions that the seed encounters prior to germinating.

The early stages of embryo development (R3 reproductive stage) involve predominantly cell division and differentiation, starting with cell division in the fertilized egg, followed by a combination of highly regulated cell division and differentiation processes during tissue and organ formation (Brandon H. Le et al., 2007; D.W. Meinke et al., 1981). These processes require a constant active supply of maternally provided oxidizable substrates before the embryonic cells become fully differentiated and photosynthetically active (Hill, Morley-Smith, & Rawsthorne, 2003; Hill & Rawsthorne, 2000). Nutrient supply by maternal source tissues continues even in green seeds, as they show a very active photoheterotrophic type of metabolism (D.K. Allen, Ohlrogge, & Shachar-Hill, 2009; Borisjuk et al., 2005; Munier-Jolain, Munier-Jolain, Roche, Ney, & Duthion, 1998; Rolletschek et al., 2005; Ruuska, Schwender, & Ohlrogge, 2004). During this seed-filling stage (R4 – 7 reproductive stages), cell elongation and accumulation of seed storage compounds represent the major developmental and metabolic processes (D.K. Allen et al., 2009; D. K. Allen, Shachar-Hill, & Ohlrogge, 2007; Bates, Durrett, Ohlrogge, & Pollard, 2009; Iyer et al., 2008 Sriram, 2004 #42). As the water content decreases and the seed biomass increases during late stages of seed filling, hormonally-regulated seed desiccation- and dormancy-related processes become active (Angelovici, Galili, Fernie, & Fait, 2010; Blochl, Grenier-de March, Sourdioux, Peterbauer, & Richter, 2005; R. Finkelstein, Reeves, Ariizumi, & Steber, 2008; Gutierrez, Van Wuytswinkel, Castelain, & Bellini, 2007).

Each of these processes and their transitions are highly regulated and coordinated at multiple levels. This complicates the development of metabolic engineering strategies for altering the levels of seed storage compounds and their composition to our advantage and for engineering drought-resistant plants by exploring molecular basis for seed desiccation tolerance. The three stages of embryo development all involve CCM. As such, the potential candidate seed-filling-specific CCM and seed desiccation and dormancy genes targeted for metabolic engineering of seed storage compounds and drought tolerance need to be distinguished from those involved in early embryogenesis. We were interested

in obtaining a systems biology perspective of transcriptional and metabolic reprogramming in developing soybean embryos to capture these transitions with associated genes. Here we present and discuss results from a detailed untargeted metabolomic and comprehensive transcriptomic time-course experiment that encompasses these transitions in developing soybean embryos. We used extensive computational analyses to integrate these diverse temporal data sets to identify unique metabolite and gene expression patterns to gain a better understanding of transitions between these developmental stages and to identify seed filling- and desiccation tolerance-specific genes.

2.2 Results and Discussion

2.2.1 Metabolic Reprogramming in Developing Soybean Embryos[1]

2.2.1.1 Oil and Protein Accumulation in Developing Soybean Embryos

Soybean plants at reproductive stages R3 and 4 started to produce pods 7 to 10 days after anthesis. During the reproductive growth stage R5 (additional 5 – 7 days after anthesis), soybean plants carried pods containing young green seeds (3 mm, day 0 of the time course) that had already started to accumulate seed storage oils and proteins. We followed changes in the levels of proteins and oil-derived fatty acids during seed filling, starting with young embryos (time point 5 days, corresponding to 17- to 22-day-old embryos) and ending with maturing and desiccating embryos (time point 55 days, corresponding to 67- to 72-day-old embryos). The accumulation of these major seed storage compounds increased in a nearly linear manner in developing soybean embryos until day 25 when a plateau was reached for both fatty acids and proteins. With respect to protein levels, the plateau was maintained until the end of the time course, while the levels of individual fatty acids started to decrease after day 40 in the time course in desiccating embryos (Figure 1.1). Oil degradation during oilseed embryo development is well documented (S. Baud, J. P. Boutin, M. Miquel, L. Lepiniec, & C. Rochat, 2002;

[1] Yihui Fang and Eva Collakova performed metabolomics analysis in soybean developing embryos.

Chia, Pike, & Rawsthorne, 2005) and will be discussed in the context of the peroxisomal glyoxylate cycle.

2.2.1.2 Polar Metabolomics in Developing Soybean Embryos

Embryos at early stages of development are metabolically very active and accumulate a variety of polar metabolites of central carbon and nitrogen metabolism, including sugars, sugar alcohols, sugar acids, amino acids, organic amines and alcohols, carboxylic acids, and phenolic compounds. We were able to detect 55 of these various metabolites by using untargeted metabolomics via gas chromatography-mass spectrometry (GC-MS) and targeted AccQ●Tag™ amino acid analysis via Waters Ultra Performance Liquid Chromatography (UPLC) coupled to fluorescent detection (FLD). We used these two different, partially complementary approaches because some amino acids were not easily detected at low levels with GC-MS because they do not derivatize and analyze well as trimethylsilyl derivatives. In fact, majority of metabolites were present at high levels in young embryos and kept gradually disappearing in aging embryos. Many metabolites, particularly amino acids, could no longer be detected in maturing and desiccating embryos by GC-MS. UPLC-FLD provided a more sensitive solution than GC-MS and many amino acids could be detected even at day 55 of the time course. However, UPLC-FLD lacks selectivity and the confirmation of the presence of individual amino acids in these samples by a different method providing some kind of spectral information about their identity is typically needed due to the high occurrence of coelution of amino acids and organic amines in complex biological samples. The low levels at which these amino acids were detected in older embryos, point to the satisfactory separation of these compounds by UPLC-FLD.

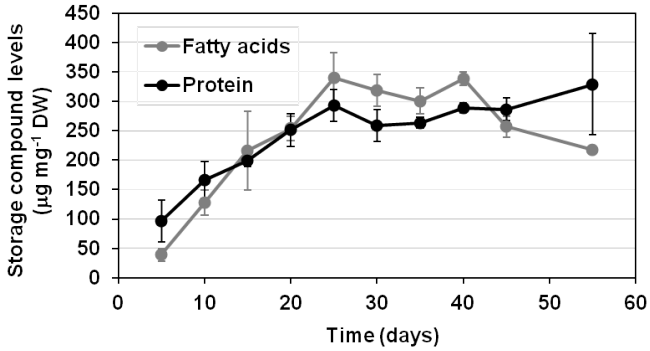

Figure 2.1. Changes in fatty acid and total protein levels in developing soybean embryos. Fatty acids from hydrolyzed oil and total proteins were analyzed by GC-FID as fatty acid methylesters and by a fluorescent hydrophobic protein assay, respectively, as described in the Experimental Section. Young 5-day-old embryos already accumulate seeds storage compounds and the levels of these compounds gradually increased in developing soybean embryos from day 5 to 25.

While simple sugars and amino acids were disappearing in maturing embryos, storage and desiccation oligosaccharides (e.g., galactinol and other sugar alcohols and raffinose) showed the opposite trend and accumulated in maturing embryos. To obtain a global perspective on these metabolic changes, we performed principal component analysis (PCA) on correlations of metabolite levels, including fatty acids from oil and total protein, in these developing soybean embryos. The first three principal components (PCs) accounted for 83.2% of the total variance among the samples (Figure 2.2). The early, intermediate, and late developmental stages representing different major metabolic processes could be clearly distinguished from each other based on combinations of PC1 and 2 (Figure 2.2A and B). Embryos at days 5, 10, and 15 were metabolically distinct from one another as well as from the rest of the embryos based on PC1. This strongest correlation was driven by the decrease in the levels of the intermediates of the central carbon and nitrogen metabolism (amino acids and monosaccharides) and the accumulation of seed storage compounds (oil and protein) in developing soybean embryos during seed filling (Figure 2.2C).

Embryos at days 20, 25, 30, and 35 clustered together based on both PC1 and 2, while they could be separated from those of days 45 and 55 based on PC2 (Figure 2.2A), suggesting that these oldest embryos (day 45 and 55) are metabolically different from embryos at days 15 – 40. PC2 correlation involved oligosaccharides (raffinose and melibiose) and their sugar alcohols and acids (galactinol, lactobionic acid, and an unknown galactinol-like sugar alcohol), all negatively correlating with oxalate and partially with Glu and Met (Figure 2.2C). Based on these observations, PC2 explains best the metabolic changes that occur during the acquisition of desiccation tolerance. Some developmental stage separation could also be observed based on PC3 and the embryos at days 10 and perhaps 15 appeared metabolically more similar to each other than to the rest of the embryos based on PC3 (Figure 2.2B). In PC3, eigenvectors point to a partial correlation between oligosaccharides and citrate, all negatively, strongly correlated with mannitol, Arg, and Orn and weakly with oxalate and mucic acid (Figure 2.2D), which is

not easily interpreted. No clear data-point separation was observed based on PC4 and the rest of the PCs (not shown). Nevertheless, the first two PCs reflect very clearly the metabolic reprogramming occurring during soybean embryo development.

We also used the SplineCluster tool enabling Bayesian coclustering analyses for time-series (Heard, Holmes, Stephens, Hand, & Dimopoulos, 2005) to obtain a visual representation of the global trends in the profiles of 55 identified metabolites and seed storage compounds (total protein and nine oil-derived fatty acids) during embryo development. This analysis yielded four clusters representing four major temporal trends (Figure 2.3). Cluster 1 mimicked the accumulation of seed storage compounds shown in Figure 2.1, as they all were present in this cluster in addition to citrate. Metabolites present in cluster 2 showed much more variable changes in their levels than did citrate or the seed storage compounds. Specifically, there was an overall increase from day 5 to 10, followed by a decrease from day 10 to 20, a moderate or no increase from day 20 to 35, and an increase from day 35 to 55. The majority of the metabolites present in cluster 2 were oligosaccharides and sugar alcohols, which are known to accumulate in developing seeds during desiccation (Angelovici et al., 2010; S. Baud et al., 2002; Blochl et al., 2005; R. Finkelstein et al., 2008; Gutierrez et al., 2007). Clusters 3 and 4 contained the monosaccharides, amino acids, and carboxylic acids of CCM. The overall trend was similar in these two clusters, with a decrease between day 5 and 20, followed by a region of no or little change in metabolite levels at day 20 to 55. The only difference between these two clusters was the slope of the initial decreases in metabolite levels (Figure 2.3). In agreement with the PCA results, clustering analysis also revealed three major metabolic processes: (i) accumulation of seed storage compounds, (ii) global decrease in the levels of the intermediates of central carbon and nitrogen metabolism, and (iii) the accumulation of desiccation-related oligosaccharides and their alcohols.

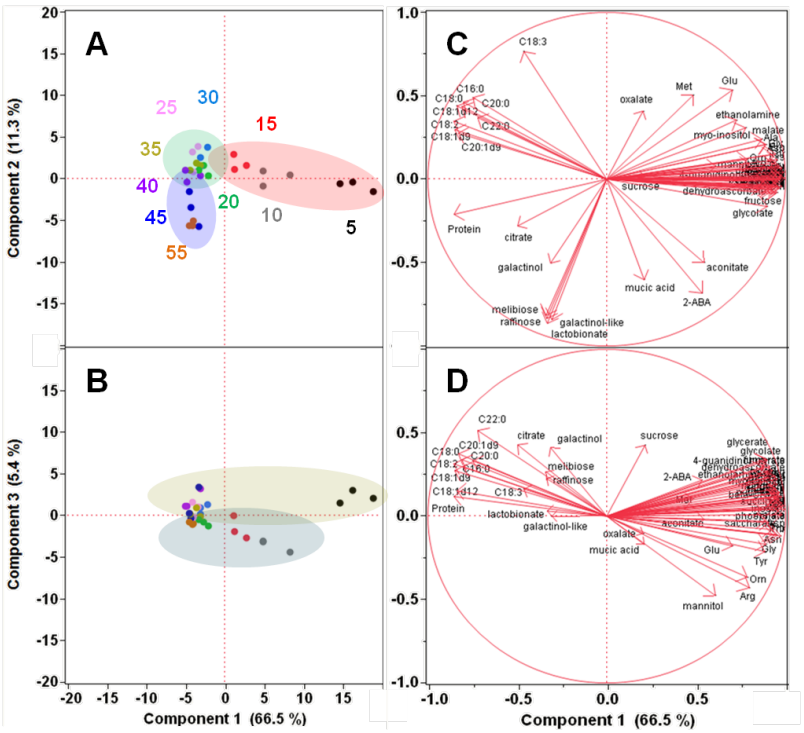

Figure 2.2. Principal Component Analysis on Metabolite Level Correlations. Score plots (A and B) and loading plots (C and D) are shown for combinations of PC1, 2, and 3. PCA was performed on combined non-redundant data involving three replicates of the relative or absolute levels of free metabolites, individual fatty acids from hydrolyzed oil, and total protein in developing soybean embryos by using JMP Pro 10 software (SAS, Cary, NC, USA). The color-coded numbers in A represent the corresponding age of the embryos (days 5 through 55), and each dot of the same color represents replicate samples. The color-coding is retained for B. The ovals highlight the basic clustering/separation of similar/dissimilar samples. In the loading plots C and D, the eigenvectors represented by red arrows show how (the direction) and how much (the length) each metabolite contributes to the individual correlations represented by PC1, 2, and 3.

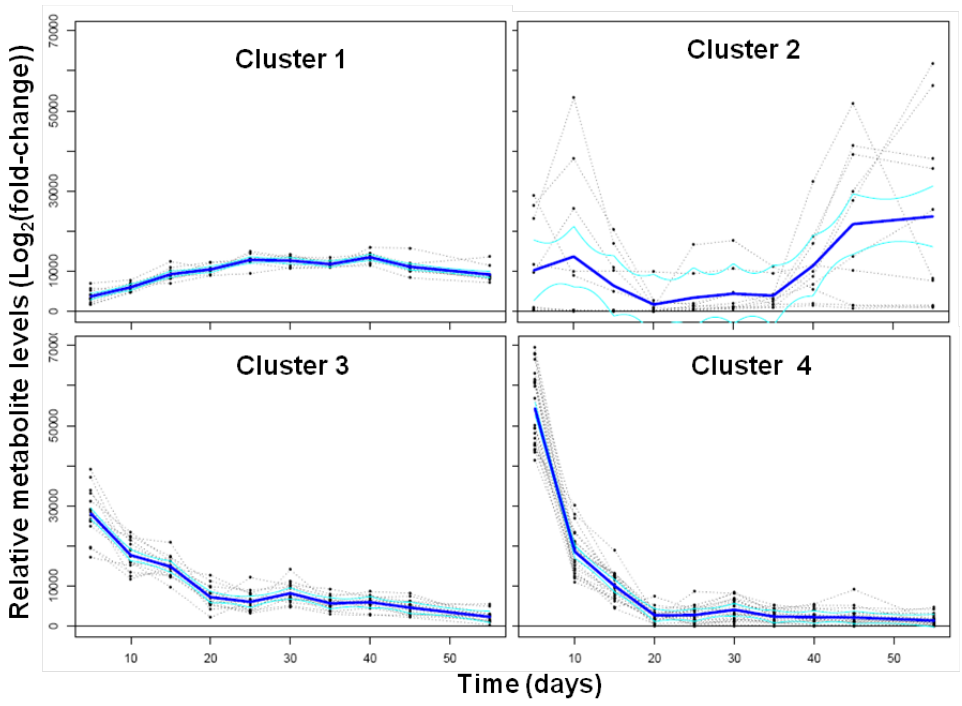

Figure 2.3 Changes in metabolite levels in developing soybean embryos. Relative metabolite levels obtained from GC-MS-FID and UPLC analyses and total protein levels were subjected to Limma computational analysis to obtain trends as described in the Experimental Section. Only metabolites that showed statistically significant changes (*p*-value < 0.05) for at least one time point were subjected to SplineCluster analysis resulting in four clusters (shown in the same scale). The levels of homoserine, aconitate, mucic acid, ornithine, and Met did not change significantly over embryo development. Cluster 1 contained 10 metabolites (citrate and nine fatty acids from oil that were detected by GC-FID) and total protein. Cluster 2 contained 10 metabolites (sucrose, oxalate, mannitol, lactobionate, melibiose, galactinol, galactinol-like, raffinose, Arg, 2-aminobutyrate). Cluster 3 contained 16 metabolites (phosphate, malate, pinitol, inositol, ethanolamine, myo-inositol, maltose, Gly, Asp, Glu, Thr, Ala, g-aminobutyrate, Pro, Lys, Tyr). Cluster 4 contained 23 metabolites (fructose, glucose, glycolate, malonate, succinate, glycerate, fumarate, beta-cyanoAla, pyroGlu, 4-guanidinobutyrate, putrescine, dehydroascorbate, saccharate, His, Asn, Ser, Gln, citrulline, Val, Ileu, Leu, Phe).

2.2.2 Transcriptional Reprogramming in Developing Soybean Embryos

Different stages of oilseed embryo development are characteristic by different metabolic processes predominating specific stages. Metabolic changes in developing embryos are accompanied by the corresponding gene expression changes. To obtain a global perspective on these changes, a detailed transcriptomic time-course was performed. First, we visualized changes in steady-state transcript levels by using metabolic MapMan that was improved from the original version by enabling visualizing time-courses from within the individual bins (Kakumanu et al., 2012; Thimm et al., 2004). Second, we performed a time-course co-expression analysis by using SplineCluster (Heard et al., 2005) as was done for metabolites.

2.2.2.1 RNA Sequencing-Based Transcriptomics

Based on the existing gene expression data and current bioinformatic predictions, *Glycine max* genome contains 54175 protein-encoding loci and 73320 transcripts (www.phytozome.net/soybean). RNA sequencing-based transcriptomics provides a much deeper level of information than microarray-based transcriptomics (Liberman, Sozzani, & Benfey, 2012; L. Wang, Li, & Brutnell, 2010; Z. Wang, Gerstein, & Snyder, 2009) and it was not surprising to see that the majority of the genes present in the soybean genome were expressed at at least in one time point (41619 genes). Most reads were found to map to multiple genes of highly conserved sequences. This observation agrees with the inferred paleopolyploididy of soybean, as its genome was duplicated twice, approximately 59 and 13 million years ago, resulting in the existence of multiple copies (2 – 6) for nearly 75% of the genes in the genome (Gill et al., 2009; Schlueter et al., 2007; Schmutz et al., 2010). These duplicated genes did not have enough time to diversify and most of them are likely to have the same or similar function (Roulin et al.,

2013). From the metabolic engineering perspective, this represents at least two obstacles: (i) genetic redundancy and (ii) unresolved gene function.

Based on the Cuffdiff2 analysis, 10794 genes showed statistically significant differential expression (*p*-value cutoff of 0.05) at at least one time point in developing soybean embryos. This leaves about 30000 genes that showed no statistically significant changes in their steady-state transcript levels during embryo development. We were particularly interested in exploring the genes that showed changes in expression and the trends of the actual changes, as some of these genes could represent targets for metabolic engineering.

2.2.2.2 Analysis and Visualization of Global Transcriptional Changes During Embryo Development

MapMan (Goffard & Weiller, 2006; Kakumanu et al., 2012; Thimm et al., 2004) was used to visualize $\log_2$ of fold-changes in the expression of genes relevant to primary and secondary metabolism in developing soybean embryos. From the global point of view, the major changes in gene expression in terms of the number of genes and the magnitude of the change were observed in later, as opposed to earlier stages of development. At day 10, the majority of metabolic genes were up-regulated relative to those at day 5, while maturing and mature embryos showed down-regulation of many CCM genes that were actively being expressed in young embryos at day 5. At the same time, there was an up-regulation of many CCM-related genes in these older embryos. This did not hold true for all processes, as there were several processes where gene expression was predominantly either up or down regulated during specific stages of embryo development.

Light reaction and Calvin cycle-associated genes showed a progressive down regulation in aging embryos. An exception was one of the small ribulose-1,5-bisphosphate carboxylase (RUBISCO) subunits (Glyma13g00190) and Rubisco activase (Glyma03g12070), which showed gradual and consistent 4- to 8-fold and 4- to 6-fold increases, respectively, in the corresponding transcript levels during seed filling and desiccation stages. RUBISCO is known to function in photoheterotrophic CCM in developing oilseed embryos to improve carbon-use efficiency (D.K. Allen et al., 2009; D.

K. Allen et al., 2007; Lonien & Schwender, 2009; Schwender, Goffman, Ohlrogge, & Shachar-Hill, 2004). Up regulation of a gene encoding RUBISCO and its activase is consistent with an active plastidic RUBISCO bypass in developing soybean embryos.

The expression of genes involved in gluconeogenesis and the peroxisomal glyoxylate cycle was up-regulated in maturing and desiccating embryos, in some instances as early as at day 35. These genes encoded malate synthase (Glyma05g03090, Glyma17g13730), NAD-dependent malate dehydrogenase (Glyma01g40580, Glyma08g06820, Glyma11g04720), phosphoenolpyruvate carboxykinase (Glyma01g02330, Glyma08g36820), citrate synthase (Glyma14g03000, Glyma02g45790), and isocitrate lyase (Glyma06g45950, Glyma12g10780). These enzymes were demonstrated to be active along with those involved in β-oxidation at later, desiccation-related stages of developing rape embryos (Chia et al., 2005). Detailed labeling studies revealed that gluconeogenesis was not occurring in these embryos, instead, CO_2, malate, citrate, Asp, and Glu were formed. The peroxisomal glyoxylate pathway was proposed to provide oxidizable carboxylic acids to the mitochondrial TCA cycle to provide energy and substrates for protein synthesis which occurs also during late desiccation stages (Chia et al., 2005). This is in an agreement with decreased levels of fatty acids in the last two time points, which could be indicative of lipid degradation to provide carbon and energy in maturing embryos with low levels of photosynthesis and maternal sucrose supply.

Raffinose and galactinol-related genes were also up-regulated during seed filling and desiccation (day 20 – 55), including several genes that were nearly identical or highly similar to Arabidopsis genes encoding galactinol-raffinose galactosyltransferase (Glyma19g40550), galactinol-sucrose galactosyltransferase (Glyma05g02510, Glyma06g18890), galactinol synthases 1 and 2 (Glyma19g40680, Glyma20g22700, Glyma03g38080, Glyma10g28610, Glyma19g41550, and Glyma03g38910), and seed imbibition O-glycosyl hydrolases (Glyma04g36410, Glyma03g29440). Up-regulation of these genes, which are involved in the synthesis of raffinose-family oligosaccharides is consistent with the accumulation of raffinose, galactinol, and related compounds observed during seed filling and desiccation phases. These compounds serve as osmoprotectants and antioxidants in leaves and developing seeds (Castillo, Delumen,

Reyes, & Delumen, 1990; X. Li, Zhuo, Jing, Liu, & Wang, 2012; Nishizawa, Yabuta, & Shigeoka, 2008; Tan, Wang, Xiang, Han, & Guo, 2013).

We also aimed to capture the transitions involving the onset and the active synthesis of seed storage compounds as well as the onset of desiccation and dormancy in developing soybean embryos. We hypothesized that capturing these transitions will enable the identification of specific genes encoding proteins involved in these different processes associated with embryo development and metabolism. Co-expression analysis using the $\log_2$ of fold-changes in expression of genes that showed statistically significant differences for at least one time point (10794 genes out of 41619 total expressed genes) generated 105 clusters. Clusters enriched in genes involved in processes of interest could be categorized into five basic gene expression patterns A – E shown in Figure 2.4. Clusters showing similar patterns tended to contain functionally related genes. For instance, careful analysis of clusters 10 – 20 (represented by trend A) revealed a strong enrichment in genes involved in microtubule-based movement and other cell cycle-associated structural and regulatory processes, including DNA replication, chromosome remodeling, mitotic cell division, cell wall remodeling, fatty acid and glycerol biosynthesis, and carbohydrate metabolism.

Genes involved in central carbon and nitrogen metabolism and metabolite transport were present in nearly all clusters, but many showed high transcript levels in the beginning of the time course, with a gradual decrease and no or low expression levels after day 25. Clusters 21 - 26 are representative of these expression patterns (trend B in Figure 2.4). These clusters were particularly enriched in genes related to metabolism and metabolite transport, including photosynthesis, respiration, carbohydrate and starch metabolism, cell wall remodeling, glycolysis, citric acid cycle, pentose phosphate pathway, lipid biosynthesis, and amino acid metabolism. These two groups of clusters contained genes showing similar trends - a gradual decrease in the steady-state transcript levels. On the other hand, in clusters 59 – 62, a corresponding sub-set of metabolism-related genes showed somewhat an opposite trend, as they became expressed after day 15 and showed slowly increasing transcript levels until day 50 (trend C in Figure 2.4).

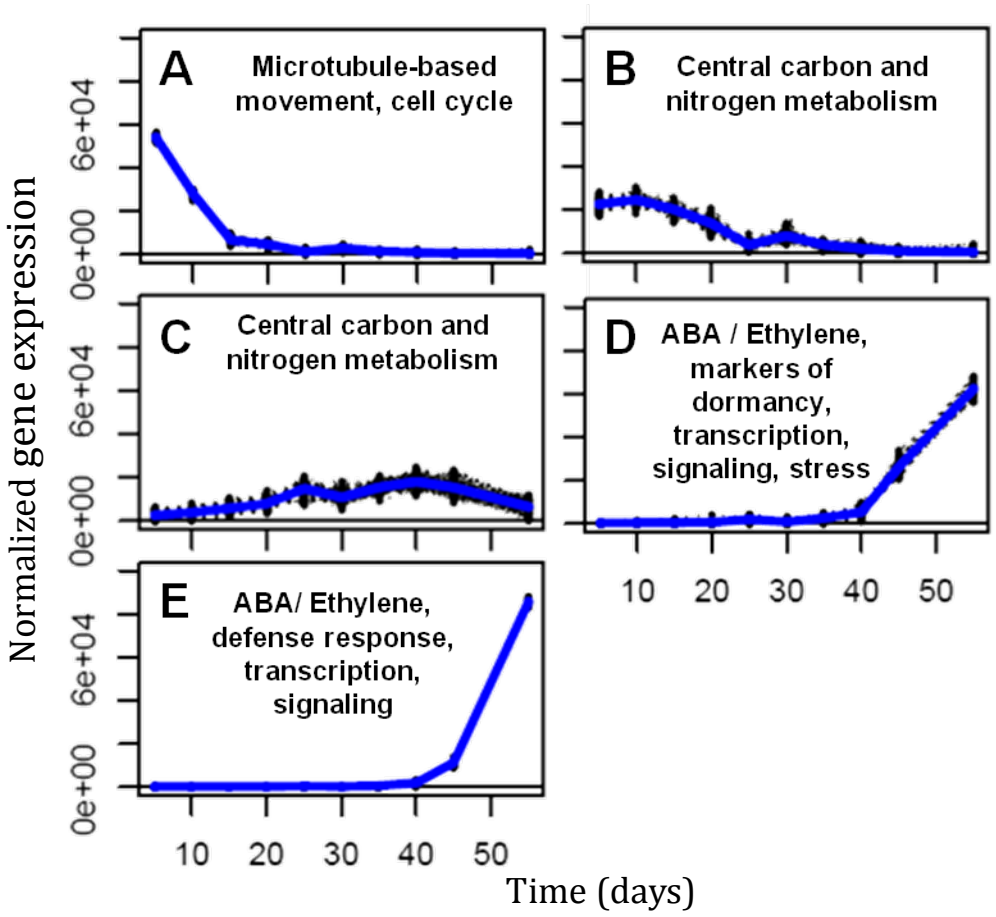

Figure 2.4. Major five gene expression trends (A – E) representing processes of interest observed in developing soybean embryos.

Genes expressed in clusters 91 – 94 and 100 – 104 (trends D and E, respectively, in Figure 2.4) reflected the completion of phases in which metabolism was most active, and the onset of processes associated with stages in the induction of dormancy and the acquisition of desiccation. The information available about the soybean genes that appear in these clusters comes almost exclusively from Arabidopsis databases and hence we have analyzed the contents of clusters 89 – 104 using Gene Ontology (GO) categorization. Many of the genes expressed during this last phase of seed development are associated with transcription and other regulatory processes. The four clusters represented by trend D show a common pattern with modest and constant expression before time point 40 and substantial increases towards the end of the time course, into the seed maturation phase. In contrast, those represented by trend E show essentially no

expression until the last two time points in development, the period in which seed maturation is nearing completion, and dormancy has been fully induced. It is possible, therefore, through study of these clusters, to identify processes that are specific to these maturation phases in seed development.

Specific processes associated with lipid metabolism appeared to have been active in the earlier phase of the seed maturation period. However, as witnessed by the up regulation of abscisic acid- (ABA) and ethylene-related processes (including ethylene and ABA signaling), dormancy induction had already begun at that time. Jasmonic acid (JA)-related signaling genes (33 genes) were also responsive during this period. Marker genes for "seed development ending in dormancy" were also observed in both groups, such as Glyma01g30670, Glyma13g39090, and Glyma11g05540, representing the homologs of AT5G01300, AT2G27300 and AT4G37300, present either in clusters 91 – 94 or clusters 100 – 104. Stress response genes associated with salt, osmotic and responses to pathogens appeared in clusters 91 – 94 (73 genes) and clusters 100 – 104 (46 genes). A group of genes (82 in all) that fell into the GO category "chloroplast" appeared in the 91 – 94 and 100 – 104 groups, including genes associated with amino acid metabolism and tocopherol biosynthesis. A homolog of Glyma13g03390 encoding NPQ4 (AT1G44575 in Arabidopsis), the 22kDA protein associated with non-photochemical quenching from PSII, was part of the 100 – 104 group, in which genes were only up-regulated at the last two time points. Glyma08g23540 and Glyma07g02480, two homologs of a gene associated with unfolded protein binding, AT1G80920, also formed part of the 100 – 104 group. A large set of genes encoding transcription factors were part of both groups (30 in the 91 – 94 group and 27 in the 100 – 104 group), especially those associated with ethylene action (9 genes in 91 – 94 and 11 genes in 100 – 104 cluster group).

The relevant responses underlying desiccation acquisition or drought tolerance were in the past regarded as being purely stress related. In fact, in the case of developing seeds, this genetically programmed developmental process proves to be more complex than a simple acquisition of desiccation tolerance. There is good evidence, from previous high-throughput studies, that desiccation acquisition and dormancy induction in orthodox seeds involve a specific and active transcriptional and metabolomic program (Angelovici et al., 2010; R. Finkelstein et al., 2008), with 30% of genes expressed during Arabidopsis

seed development being expressed during this late and final stage of seed development. That same pattern is reflected in the data reported here, where 22% of all genes expressed over the course of seed development were present in clusters 78 – 105, which showed a common pattern of up regulation at the last time points. Some of the genes that are part of this late phase of seed development are well known as components of stress adaptation in other plant tissues, such as the LEA and heat shock proteins (Fisher, 2008; Gechev, Dinakar, Benina, Toneva, & Bartels, 2012; Illing, Denby, Collett, Shen, & Farrant, 2005). Representatives from these gene families (91 – 94: Glyma13g16510, Glyma01g39260, Glyma07g04860, Glyma16g01440, Glyma03g31380, Glyma08g26100, Glyma19g32920, 100-104: Glyma07g02480, Glyma08g23540, Glyma07g32070, Glyma11g37450, and Glyma16g23750) are present in the late clusters of this data set.

Transcriptional and post-transcriptional events occur in drying and dry seeds (Holdsworth, Bentsink, & Soppe, 2008) and these represent events which have the potential to affect germination and emergence. Further evidence for this is afforded in our data set, in which soybean genes encoding members of several groups of transcription factors (37 in clusters 91 – 94 and 28 in clusters 100 – 104), and genes associated with signal transduction events (28 in clusters 91 – 94 and 21 in clusters 100 – 104), are specifically up regulated at the last two time points of seed development. The acquisition of desiccation tolerance involves the production of stored mRNAs, whose encoded proteins are essential for subsequent germination (Sano et al., 2012). Evidence for this association between transcriptional events occurring during the final phases of seed maturation and those occurring during the early stages of germination is afforded by the expression of several genes associated with germination in Clusters 91 – 94 (7 genes: Glyma04g09380, Glyma11g02040, Glyma06g09520, Glyma13g39090, Glyma17g14650, Glyma17g36080, and Glyma04g40880) and 100 – 104 (Glyma17g09850).

As can be seem from the results presented here, seed maturation and dormancy involve much transcriptional activation and signaling right through the maturation period, especially via the ethylene and ABA pathways. An active role appears to be being played by the chloroplast as a center of metabolic activity. This latter result is somewhat surprising as the chloroplast is losing its photosynthetic capacity during the later stages of seed development, when chlorophyll loss appears to be already taking place.

2.3 Integrated Overview of Transcriptional and Metabolic Changes during Soybean Embryo Development

A systems biology global perspective for viewing the interacting cellular components (genes, transcripts, proteins, and metabolites) and their regulatory networks offers unprecedented insights into their cellular functions at the molecular level (Fukushima, Kusano, Redestig, Arita, & Saito, 2009; Liberman et al., 2012; Urano, Kurihara, Seki, & Shinozaki, 2010). Metabolite and transcript profiling are high-throughput systems biology approaches that enable the assessment of steady-state metabolite and transcript levels, respectively, at a particular moment in the studied system (Fiehn, 2001; Fiehn et al., 2000; L. Wang et al., 2010). Current trends in systems biology high-throughput approaches involve tissue- and cell-specific analyses to obtain a global view of processes specific to the studied system (Dai & Chen, 2012; Kueger, Steinhauser, Willmitzer, & Giavalisco, 2012; Moco, Schneider, & Vervoort, 2009). The major problem in achieving meaningful results from such analyses is the ability to isolate or separate specific cells or tissues of interest without altering the corresponding *in situ* transcriptomes, proteomes, and metabolomes (Klie et al., 2011). Developing soybean embryos represent a unique and highly specific system from this perspective, as the majority of the embryo biomass is represented by cotyledons, with only a limited number of cell types and with the majority of embryonic cells in the cotyledon involved primarily in central carbon and nitrogen metabolism specific to seed filling and desiccation (D. K. Allen et al., 2007; Sriram et al., 2004).

Cells of developing embryos undergo transcriptional and metabolic reprogramming during two main transitions between different types of development and metabolism. First, dividing and differentiating embryonic cells progressively switch their developmental program to cell elongation at the onset of the seed filling phases. This developmental switch is accompanied by gradual metabolic changes from heterotrophic CCM providing substrates and energy for cell division and differentiation to photoheterotrophic CCM during the seed storage reserve accumulation phases. Second, elongating cells at the seed-filling stage turn on seed maturation and desiccation processes to prepare seeds for dormancy. We were able to capture transcriptional and

metabolic changes at the end of the first, and the beginning of the second transition, which enabled the identification of genes potentially connected to developmental, metabolic, and regulatory processes in seed-filling and desiccation phases.

Embryos at early stages of seed filling (days 5 to 15) are already green and accumulating seed storage reserves. From the developmental perspective, these fully differentiated young embryos undergo a combination of cell division and elongation, as numerous mitotic cell-cycle-related structural and regulatory genes, including microtubule-based molecular movement, DNA replication, chromosome remodeling, and epigenetic regulation were still expressed in the beginning of seed filling. However, their relative steady-state transcript levels decreased rapidly within the first 10 days, suggesting that the sole cell elongation starts between day 10 and 15 in the time course (22- to 32-day-old embryos) during seed filling. From the metabolic perspective, these young embryos also accumulated very high levels of the precursors of seed storage compounds including carbohydrates, carboxylic acids, and amino acids. The levels of these CCM intermediates became gradually depleted, which also coincided with a similar decrease in the transcript levels of many metabolic genes involved in various aspects of CCM, including glycolysis, TCA cycle, pentose-phosphate pathway, fatty acid synthesis, amino acid metabolism, and metabolite transport. Clusters 21 – 26 (trend B) were particularly enriched in these metabolic genes. Because their expression coincided with the initial decrease in metabolite levels, we hypothesize that these genes encode CCM enzymes involved in heterotrophic metabolism during early embryogenesis. This type of metabolism remains largely unexplored, as early embryo development has been extensively studied from the developmental, rather than a metabolic perspective (Siobhan A. Braybrook & Harada, 2008; Jeong, Volny, & Lukowitz, 2012; Lau, Slane, Herud, Kong, & Juergens, 2012; Minako Ueda & Laux, 2012). As such, the predicted involvement of these genes in heterotrophic CCM remains to be confirmed experimentally.

A similar set of metabolic genes showed a nearly opposite trend (clusters 59 – 62, trend C), as their expression increased during seed filling, while the precursors of seed storage compounds, especially amino acids, became nearly depleted. Gradual decreases in levels of these CCM intermediates correlated well with the increase in the levels of

total protein and oil-derived fatty acids, suggesting that these intermediates of central carbon and nitrogen metabolism were used for the synthesis of seed storage compounds at early stages of seed filling. Metabolite levels do not reflect metabolic activity of the system and low metabolite levels especially in central carbon and nitrogen metabolism are often indicative of very large metabolic fluxes producing and consuming these metabolites (D. K. Allen et al., 2007; Alonso, Piasecki, Wang, LaClair, & Shachar-Hill, 2010; Schwender & Ohlrogge, 2002; Schwender, Shachar-Hill, & Ohlrogge, 2006; Sriram et al., 2004). Our gene expression data are consistent with very active CCM observed during seed filling in developing soybean embryos (D.K. Allen et al., 2009; D. K. Allen et al., 2007; Iyer et al., 2008; Sriram et al., 2004). Genes present in clusters 59 – 62 represent potential targets for metabolic engineering of seed storage compounds.

2.4 Experimental Section[1]

2.4.1 Plant Growth and Embryo Harvesting

Soybean seeds of *Glycine max* (L.) Merr. *cv*. Williams 82 were potted in ProMix Sunshine #1 in 2-gallon pots and grown under controlled growth chamber conditions, as follows: 11/13 hour day/night photoperiod, 28/22°C day/night temperatures with light intensities between 350 and 450 μE and the relative humidity 70%. Six plants (to provide 3 biological replicates for each time point, with each replicate represented by embryos pooled from two plants) were grown for 30-35 days to achieve the early R5 stage that is characterized by having the first pod containing seeds that are 3 mm long. Seed length was measured with a ruler over the pod on one side while light was shed from the other side to highlight the shape of the seed inside the pod. Every pod that met the 3-mm length criteria was tagged as "day 0" of the time course by color-marking the tip of the pod. The pods were harvested on ice randomly at 5-day intervals (except for the 10-day interval for the last time point) from the day they were tagged throughout the development from the early R5 to the early R7 stages (transition from green to yellow seeds), yielding the embryos belonging to days 5, 10, 15, 20, 25, 30, 35, 40, 45, and 55. Dissected embryos were rinsed in ice-cold water, immediately frozen in liquid nitrogen and stored at -70 °C

[1] Eva Collakova and Yihui Fang performed all the experiments in this chapter.

until embryos from all time points were harvested before grinding and extractions. Frozen embryos were ground to a fine powder. For biomass measurements and metabolite profiling, the powder was lyophilized for 3 days and the measurements were performed on 1.00 ± 0.05 mg of the dry weight. Frozen powdered tissue from the same experiment was used for RNA extractions.

2.4.2 Biomass Measurements

Oil and proteins were extracted from the lyophilized powder with heptanes, diethyl ether, and water of equal volumes (400 µl each) in the presence of 10 µg heptadecanoic acid as an internal standard for fatty acid analysis. *(A) Fatty acid analysis.* Fatty acids were analyzed after acidic hydrolysis as fatty acid methyl esters (FAME) by gas chromatography coupled with flame ionization detection (GC-FID) as described (Y. Lu, Savage, Larson, Wilkerson, & Last, 2011). FAME separation and analysis was achieved on an Agilent 7890A series GC-FID (Agilent Technologies, Santa Clara, CA, USA) equipped with a 30-m DB-23 column (0.25 mm x 0.25 µm, Agilent Technologies). *(B) Total protein analysis.* Total proteins present in the aqueous phase and insoluble interphase were solubilized in the presence of 0.1% (w/v) sodium dodecylsulfate and their relative levels measured using the MGT hydrophobic protein assay kit (Marker Gene Technologies, Eugene, OR, USA) on a BioTek Synergy™ H4 plate reader (BioTek, Winooski, VT, USA) following the manufacturer recommendations.

2.4.3 Metabolite Profiling

Untargeted polar metabolite profiling was performed based as described previously (Collakova et al., 2008; Duran, Yang, Wang, & Sumner, 2003; Goyer et al., 2005). Briefly, the lyophilized powder was extracted with 400 µl each chloroform and water containing norvaline and ribitol (50 µM each) as internal standards. Aqueous phase (100 µl) containing polar metabolites was dried under a stream of nitrogen gas at 55°C and the metabolites were derivatized first with methoxyamine.HCl and then with 2,2,2-trifluoro-n-methyl-n-(trimethylsilyl)-acetamide containing 1% trimethylchlorosilane (Thermo Fisher Scientific, Waltham, MA). Trimethylsilylated metabolite derivatives were separated on an Agilent 7890A series GC equipped with a DB-5MS-DG column (30 m length x 0.25 mm x 0.25 µm with a 10-m pre-column, Agilent Technologies) and

analyzed on an Agilent 5975C series single quadrupole mass spectrometer (MS). The GC temperature program and MS conditions were as described (Kind et al., 2009), except that the m/z scan range was from 100 to 650. Metabolites were identified using the FiehnLib spectral and retention time library (Kind et al., 2009), our own custom-built spectral and retention time library, and the spectral NIST library (National Institute of Standards and Technology, Gaithersburg, MD). Automated Mass Spectrometry Deconvolution and Identification System (AMDIS, NIST) was used to deconvolute signals from the coeluting compounds. We were able to identify and quantify the relative levels of 55 major metabolites in developing soybean embryos by using the Enhanced Mass Selective Detector ChemStation software (Agilent Technologies) in combination with the three above-mentioned libraries. The identities of metabolites and quality of integration were curated manually on an individual basis by using the QEdit function of the ChemStation software after automated peak area integration. Relative areas of the internal standards were used to correct for recovery in quantitation of each metabolite and all samples were standardized in respect to the dry weight of the tissue used for extractions.

Most amino acids in embryos at or after day 20 could not be detected by the GC-MS-based metabolite profiling. As such, their absolute levels in all collected embryos were determined by using an H-class Acquity UPLC coupled to a fluorescent detector (FLD) (Waters, Milford, MA). Amino acids were extracted with equal volumes (100 µl) of chloroform and 10 mM HCl containing 20 µM norvaline as an internal standard. Five and ten µl of the aqueous phase from first two and the remaining time points, respectively, were taken to determine the levels of free amino acids in a 0.5-µl injection (50-µl total reaction volume) by using the AccQ•Tag Ultra Amino Acid kit (Waters) according to the manufacturer's recommendations. Two different UPLC gradients were used to enable the separation of (i) Gln and Asn and (ii) Glu and citrulline. The first gradient method used was developed by Waters for analysis of free amino acids in cell cultures and was implemented without any modifications. The second method was based on another Waters method originally intended for analyzing amino acids originating from protein hydrolysates, but with modifications to the UPLC gradient to at least partially separate Glu and citrulline. The following gradient with Waters Eluent A (5% in water by volume)

and undiluted Eluent B at a flow rate of 0.7 ml min^{-1} were used (% given for A and curve was 6 unless otherwise stated): 0 - 0.54 min isocratic 99.9%; 5.74 min 90.9% (curve 7); 7.74 min 78.8%; 8.04 min 40.4%; 8.05 – 8.64 min isocratic 10%; 8.65 – 9.00 min isocratic 0%; and 9.20 – 12.00 min equilibration to 99.9%. Standard curves and Empower 3 software (Waters) were used to obtain absolute amino acid levels (corrected for recoveries and dry weight) in developing soybean embryos.

2.4.4 Transcriptomics

2.4.4.1 RNA Isolation, cDNA Library Preparation, and Illumina RNA Sequencing

RNA sequencing was performed to investigate changes in the soybean transcriptome during seed development. First, total RNA was isolated from the frozen ground embryos by using an RNeasy Plant RNA Purification Mini Kit (Qiagen, Germantown, MD) according to the manufacturer's recommendations, with a minor modification. The samples were centrifuged at 16,000 g at 4°C for 2 min after the addition of the RLT buffer to obtain a clear aqueous phase for the subsequent column purification steps. The RNA concentration and quality was measured by using a Bioanalyzer 2100 (Agilent Technologies). A minimum of 400 ng ml^{-1} of RNA and an RNA integrity number (RIN) greater than 8.0 was obtained for reliable RNA sequencing analyses. Library preparation and RNA sequencing analysis were performed at the Génome Québec Innovation Centre (Montréal, Canada). The cDNA libraries were generated using TruSeq RNA sample preparation kit (Illumina, San Diego, CA) and paired-end 100-bp long RNA-seq reads were generated on a HiSeq 2000 sequencing system (Illumina), with each lane containing multiplexed cDNA libraries pooled from the three biological replicates of the individual time-points.

2.4.4.2 RNA Sequencing Data Analysis Pipeline

Our pipeline (Figure 2.5) consists of the following steps: First, each sequence read was evaluated for the presence of a recognizable, library-specific 5' barcode sequence. Raw reads with length ≤ 32 and the quality score ≤ 30 were filtered out using Illumina purity filter and further evaluated based on the distribution of phred-like quality scores at

each cycle. The demultiplexed reads were then mapped to the reference genome using Tophat, which internally uses Bowtie as its high-throughput read alignment tool. Tophat can also find splice junctions between exons with good accuracy (C. Trapnell et al., 2009). The resulting reads were mapped to version 1.0 of the *G. max* reference genome (Schmutz et al., 2010) provided in Phytozome v8.0 (http://www.phytozome.net/soybean.php) using TopHat v2.0.4 (C. Trapnell et al., 2010) in conjunction with Bowtie v0.12.8 (Langmead, Trapnell, Pop, & Salzberg, 2009) using all default parameters except for the distance between mated pairs set to 60 bp.

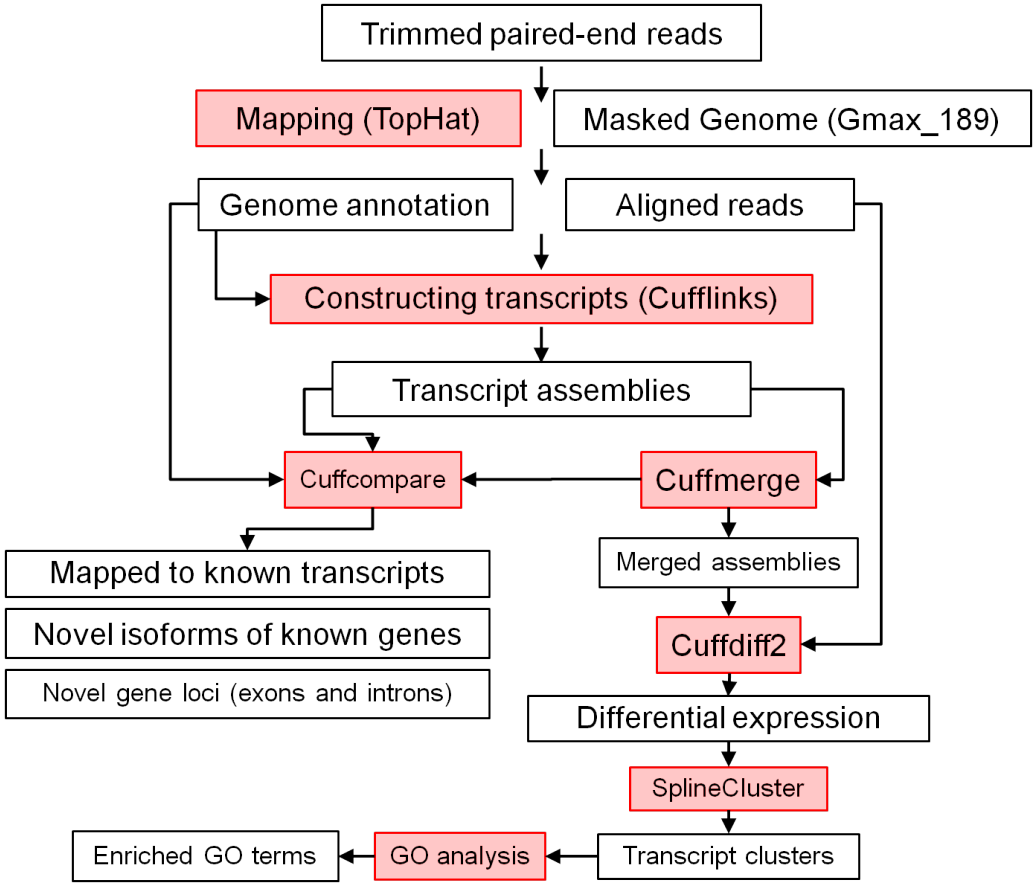

Figure 2.5 RNA-Seq analysis pipeline

Second, the reads were concatenated using Cufflinks (C. Trapnell et al., 2010). The RABT (Reference Annotation Based Transcript assembly) technique was used for this purpose (Roberts et al., 2011), which results in a good accuracy for finding novel genes and isoforms of genes when a high quality sequence reference exists for that genome. Cufflinks results were merged using the Cuffmerge tool (C. Trapnell, Roberts, et al., 2012). The Cufflinks and Cuffmerge results were compared with the reference genome

using Cuffcompare to find known genes, novel genes, novel splice variants, and transcripts expressed from intergenic regions (C. Trapnell, Roberts, et al., 2012).

Third, the reads from Tophat and merged assemblies from Cuffmerge were fed to Cuffdiff2 (C. Trapnell, Hendrickson, et al., 2012) for testing differential expression in a time-series manner, including splice variants. This allows testing for the presence of differentially-expressed genes between two consecutive time points and the identification of gene sets showing statistically significant change (p-value < 0.05) at least in one of the time points (10794 genes). These differentially-expressed genes were subjected to the SplineCluster clustering software enabling gene coexpression analysis in a time-series-dependent manner (Heard et al., 2005). Since we were interested in common patterns of behavior for genes across the time course (not the actual expression values), the mean of each gene expression value was moved to a fixed value before the clustering analysis. The lowest gene expression value was 4.71126E-11 and all zeroes were replaced with this value to prevent the infinity problem (dividing by zero) to enable differential gene expression analysis. They were considered not to be expressed. The Arabidopsis homologs of the clustered soybean genes were associated with GO annotations with an in-house perl program using the Gene Ontology OBO v1.2 specification and The Arabidopsis Information Resource (TAIR) GO annotations (both downloaded form their respective website on March 11, 2013).

2.4.5 MapMan

Modified MapMan that enables the visualization of time-courses within specific bins (Goffard & Weiller, 2006; Kakumanu et al., 2012; Thimm et al., 2004) was used to visualize possible temporal changes in gene expression. To generate the MapMan input file, FPKM gene expression values obtained from Cuffmerge output for every gene at every time point were compared to the corresponding values in the first time point to obtain fold-change in gene expression. We used $\log_2$(fold-change) relative to the first time point and only genes present in the MapMan Gmax_109_peptide mapping file and showing a statistically significant change ($P \leq 0.05$) for at least one time point were included in the MapMan input file for visualization purposes. Soybean gene IDs did not contain alternate splicing information and as such, the MapMan Gmax_109_peptide

mapping file was altered to remove this information as well as all redundant gene ID entries.

2.5 Conclusions

We generated high quality transcriptomic and metabolomic data relevant to seed filling and desiccation acquisition in developing soybean embryos. Computational analyses of these large data sets enabled a systems biology view of global transcriptional and metabolic changes during transitions from cell division to elongation and from seed filling to desiccation processes. From the metabolic engineering perspective, there appears to be a set of metabolism-related genes specific to heterotrophic metabolism at early stages of soybean embryo development and another set of similar genes expressed during seed filling. These genes represent potential targets for future metabolic engineering of seed composition. However, there is a high level of genetic and metabolic redundancy in developing soybean embryos. In addition, CCM is highly regulated and there are several levels of regulation between transcription and the final product/process targeted for metabolic engineering.

Detailed coexpression analysis in combination with metabolomic data presented in this study also provide a valuable resource of potential regulators associated with metabolic and desiccation tolerance genes in individual clusters. Every cluster contained a number of transcription factors, protein kinases and phosphatases, components of the ubiquitin-associated protein degradation system, various proteases, and genes encoding other protein-modifying enzymes such as farnesyl and acyl transferases. Besides regulators, our data sets enabled the identification of splice variants and at least 3400 novel genes containing exons and introns that map to the soybean genome for future studies.

The relevant responses underlying desiccation acquisition or drought tolerance have historically been regarded as stress related. In the case of developing seeds, this genetically programmed developmental process proves to be more complex than a simple acquisition of desiccation tolerance and accumulation of raffinose-related oligosaccharides. As can be seen from the results presented here, seed maturation and dormancy involve much transcriptional activation and signaling right through the

maturation period, especially via the ethylene pathway with an active role being played by the chloroplast as a center of metabolic activity. This latter result is somewhat surprising as the chloroplast is losing its photosynthetic capacity during the later stages of seed development. As such, understanding the underlying mechanisms for these responses is potentially useful in engineering drought and desiccation tolerant plants.

3 Changes in RNA Splicing in Developing Soybean (*Glycine max*) Embryos[1]

This chapter is produced from (Aghamirzaie et al., 2013), which is an open access journal.

Aghamirzaie, D., Nabiyouni, M., Fang, Y., Klumas, C., Heath, L. S., Grene, R., & Collakova, E. (2013). Changes in RNA splicing in developing soybean (*Glycine max*) embryos. *Biology, 2*, 1311-1337. doi:10.3390/biology2041311

Abstract: Oilseeds represent an important source of food, chemicals, and biofuel in the form of oils, proteins, and carbohydrates. These molecules, also referred to as seed storage compounds, are synthesized in developing seeds and provide resources and energy during seed germination. The accumulation of storage compounds in developing seeds is highly regulated at multiple tiers, including at the transcriptional and post-transcriptional levels. RNA sequencing was used to provide comprehensive information about transcriptional and post-transcriptional events that take place in developing soybean embryos. Bioinformatic analyses lead to the identification of different classes of alternatively spliced isoforms and corresponding changes in their levels on a global scale during soybean embryo development. Alternative splicing was associated with transcripts involved in various metabolic and developmental processes, including central carbon and nitrogen metabolism, induction of maturation and dormancy, and splicing itself. Detailed examination of selected RNA isoforms revealed alterations in individual domains that could result in changes in subcellular localization of the resulting proteins, protein-protein and enzyme-substrate interactions, and regulation of protein activities. Different isoforms may play an important role in regulating developmental and metabolic processes occurring in developing oilseed embryos.

[1] Delasa Aghamirzaie performed all the computational data analysis. Mahdi Nabiyouni was involved in the visualization of splicing graphs. Curtis Klumas helped in data analysis section with guidance of Lenwood Heath. Yihui Fang performed experimental validation including PCR. Ruth Grene and Eva Collakova performed all the biological data mining of the results.

Keywords: abscisic acid; alternative splicing; auxin; central carbon and nitrogen metabolism; desiccation tolerance; dormancy induction; post-transcriptional regulation; seed and embryo development; soybean

3.1 Introduction

Seeds are essential for sexual propagation of plants and serve as an important source of proteins, oils, and carbohydrates for food and industrial purposes (Durrett, Benning, & Ohlrogge, 2008; Weselake et al., 2009). These compounds represent the seed storage compounds that are synthesized through pathways of central carbon and nitrogen metabolism (CCNM) and accumulate during the seed filling and maturation phases of embryo development (D.K. Allen et al., 2009; D. K. Allen et al., 2007; Alonso et al., 2010; Bates et al., 2009; S. Baud, J.-P. Boutin, M. Miquel, L. Lepiniec, & C. Rochat, 2002; Borisjuk et al., 2005; Iyer et al., 2008; Schwender & Ohlrogge, 2002; Sriram et al., 2004). Prior to seed filling, embryonic cells divide and differentiate into different cell types to form the asymmetric apical-basal embryo axis giving rise to a fully differentiated embryo (D. K. Allen et al., 2007; Jeong et al., 2012; Lau et al., 2012; Minako Ueda & Laux, 2012). As these embryos become photosynthetically competent in some plant species, such as soybean, they also start accumulating various seed storage compounds that will serve as substrates and provide energy during germination (Eastmond & Graham, 2001; Graham, 2008; Penfield et al., 2005). Drying seeds initiate molecular and physiological responses leading to dormancy, and the acquisition of desiccation tolerance, ensuring seed viability during storage preceeding germination (Angelovici et al., 2010; Blochl et al., 2005; R. Finkelstein et al., 2008; Gutierrez et al., 2007). The processes that occur during the various stages of embryo and seed development and maturation are highly regulated at various levels. Post-transcriptional regulation represents one of the many tiers of complex regulatory events accompanying embryo and seed development and metabolism.

Alternative splicing (AS) is a post-transcriptional regulatory process contributing to transcriptome and proteome diversities by enabling the production of multiple mRNA and protein molecules from a single gene (Barbazuk, Fu, & McGinnis, 2008; Marden,

2008). Different splice isoforms originating from the same gene may contain or lack specific sequences, including functional, regulatory, and interaction domains as well as organelle localization sequences. As such, the resulting mRNA and protein molecules may be affected in terms of stability, subcellular localization, structure, protein-molecule interactions, regulation, and function. However, AS does not seem to diversify plant proteomes as much as transcriptomes because novel domains are not often introduced to give rise to new proteins or proteins with additional or altered functions (E. I. Severing, van Dijk, Stiekema, & van Ham, 2009). Nevertheless, the presence or absence of the domains has the potential to greatly affect protein-protein interactions (E. I. Severing et al., 2012; Syed, Kalyna, Marquez, Barta, & Brown, 2012).

The extent of AS in plants was underestimated prior to the availability of numerous whole-genome sequences and high-throughput RNA sequencing (RNA-seq) technologies (Akhunov et al., 2013; Campbell, Haas, Hamilton, Mount, & Buell, 2006; Filichkin et al., 2010; Lister, Gregory, & Ecker, 2009). Recent global AS analyses revealed that between 20 to 60 % of plant genes are subjected to AS in different plant species and under different conditions (Akhunov et al., 2013; Filichkin et al., 2010; James et al., 2012; S. E. Sanchez et al., 2010). It appears that in plants, AS is a wide-spread phenomenon regulating expression of genes involved in growth, development, metabolism, and abiotic and biotic stresses (D. K. Allen et al., 2007; Hirayama & Shinozaki, 2010; James et al., 2012; F. Jia & Rock, 2013; Jones et al., 2012; Lightfoot, Malone, Timmis, & Orford, 2008; Martin-Trillo et al., 2011; Mastrangelo, Marone, Laido, De Leonardis, & De Vita, 2012; Matsukura et al., 2010; Matsumura et al., 2009; Mazzucotelli et al., 2008; Peremyslov et al., 2011; Rosloski et al., 2013; S. E. Sanchez et al., 2010; Staiger, Korneli, Lummer, & Navarro, 2013). In their genomes, plants contain numerous non-consensus splice sites that tend to lead to transcripts with premature protein synthesis termination (Filichkin et al., 2010; B. B. Wang & Brendel, 2006). Some truncated mRNA molecules are subjected to nonsense-mediated decay (NMD), a mechanism proposed to regulate transcript abundance (James et al., 2012; Kalyna et al., 2012; Palusa & Reddy, 2010). In *Arabidopsis thaliana*, approximately 13% of genes containing introns are templates for truncated mRNA molecules degraded by NMD and were shown to play major roles in development, regulation, and stress responses (Kalyna et al., 2012).

AS takes place on a large complex called the spliceosome that is predicted to be composed of numerous snRNA and protein molecules (B.-B. Wang & Brendel, 2004). The individual AS components themselves are subjected to AS and intricate regulation (Day et al., 2012; Stauffer, Westermann, Wagner, & Wachter, 2010). In Arabidopsis, Ser/Arg-rich proteins regulate AS of other proteins and themselves and some of the resulting transcripts contain stop codons and are likely to undergo NMD (Day et al., 2012; Palusa & Reddy, 2010; Thomas et al., 2012). Some isoforms of the Ser/Arg-rich proteins have also distinct functions in plant development (X. N. Zhang & Mount, 2009). The consensus splice sequences do not by themselves possess sufficient specificity, hence the existence of additional *cis* and *trans* elements including exonic and intronic splicing enhancers and silencers and regulatory proteins is inferred (M. Pertea, Mount, & Salzberg, 2007). This results in a diverse specificity of AS targets regarding individual processes. Some proteins regulate the AS of a small number of genes, while others target genes associated with entire processes or a group of related processes. SMU1 and 2 are examples of AS regulators of development in maize and Arabidopsis that have specific targets (Chung, Wang, Kim, Yadegari, & Larkins, 2009). MOS12 is an Arg-rich splicing factor important for AS of plant resistance transcripts (F. Xu, Xu, Wiermer, Zhang, & Li, 2012). In contrast, polypyrimidine tract binding protein homologs regulate AS in hundreds of genes, many belonging to stress- and desiccation-related processes as well as flowering involving phytochrome- and abscisic acid (ABA)-mediated processes (Ruhl et al., 2012). Similarly, Gly-rich RNA-binding proteins are able to bind RNA molecules and regulate AS of multiple targets in Arabidopsis (Streitner et al., 2012).

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) activase was the first protein shown to be a subject of AS regulation in plants (Werneke, Chatfield, & Ogren, 1989). Truncated transcripts of RuBisCO activase produced during heat stress were more stable than the long isoform (DeRidder, Shybut, Dyle, Kremling, & Shapiro, 2012). In plant metabolism, AS is also known to regulate enzyme activity (Brummell et al., 2011; Hasse, Mikkat, Hagemann, & Bauwe, 2009), sub-cellular localization of the enzymes (Dixon, Hawkins, Hussey, & Edwards, 2009; Kriechbaumer, Wang, Hawes, & Abell, 2012; Lamberto, Percudani, Gatti, Folli, & Petrucco, 2010; Puyaubert, Denis, & Alban, 2008; Wiszniewski, Smith, & Bussell, 2012) and metabolic responses to stress (Roman et

al., 2012). Expression of oleosin genes in the moss *Physcomitrella patens* is also subjected to tissue-specific AS (C. Y. Huang, Chung, Lin, Hsing, & Huang, 2009). Some isoforms are expressed only in spores, where the resulting oleosins function in the formation of oil bodies similar to those observed in seeds and pollen. As in seeds and pollen, oil stored in these bodies is mobilized during spore germination, which is crucial for sexual reproduction of mosses (C. Y. Huang et al., 2009).

AS plays an important role during early and late phases of seed and embryo development. In maize, the *ROUGH ENDOSPERM 3* gene is a source of 19 alternatively spliced variants, one of which localizes to the nucleolus and encodes an RNA splicing factor required for cell differentiation of maize endosperm during embryo development (Fouquet et al., 2011). In Arabidopsis, the C2H2-domain proteins are nuclear proteins represented by nine isoforms that are most likely involved in nuclear division in the endosperm essential for normal seed development (X. D. Lu et al., 2012). Corresponding mutants are lethal, as embryo development is arrested between the globular and heart stage transition, while the endosperm nuclei contain multiple nucleoli (X. D. Lu et al., 2012). AS of the transcriptional regulator of seed development in Arabidopsis ABI3 (ABA INSENSITIVE 3) is regulated by the splicing factor SUA (SUPPRESSOR OF ABI3) (Sugliani, Brambilla, Clerkx, Koornneef, & Soppe, 2010). In Arabidopsis, the phytochrome-interacting factor PIF6 is a transcription factor (TF) existing in two splice isoforms (Penfield, Josse, & Halliday, 2010). The full length variant has a potential DNA-binding domain that is spliced out of the short variant, resulting in the introduction of a stop codon. Overexpression of the short, but not the long isoform reduces seed dormancy (Penfield et al., 2010). Acquisition of desiccation tolerance in maturing seeds shares similarities with abiotic stress responses in plants (Angelovici et al., 2010; Buitink et al., 2006; E. Collakova et al., 2013; Gechev et al., 2012; Grene, Vasquez-Robinet, & Bohnert, 2011). It is reasonable to expect that some AS events regulating seed maturation will be conserved between these processes.

Here we present results from a comprehensive global AS analysis of the transcriptome in developing soybean embryos during seed filling and maturation. Our results indicate that AS is a widely spread phenomenon in both metabolic and hormone-mediated signaling processes during seed filling and acquisition of dormancy and

desiccation tolerance in developing soybean embryos. We discuss these findings from a global perspective as well as by focusing on selected examples of AS-derived protein isoforms involved in various aspects of CCNM and ABA- and auxin-mediated signaling relevant to the maturation and desiccation processes.

3.2 Experimental Section

3.2.1 RNA-Sequencing-Based Transcriptomics

In the previous study, a detailed time-course of soybean embryo development involving ten time points with three replicates each was performed (Eva Collakova et al., 2013). Reads were mapped to the Gmax_109 version of soybean *G. max* (cv. Williams 82) genome, which was recently sequenced (Schmutz et al., 2010) and subjected to a RNA-seq and differential gene expression pipeline as described (Eva Collakova et al., 2013). The resulting data sets are available in the Gene Expression Omnibus database (GEO accession number GSE46153). For this study, the newest available version of the genome from Phytozome (Gmax_189 (Goodstein et al., 2012)) was used for analyzing transcripts reported here. Briefly, the RNA-seq analysis pipeline consists of the following steps. First, the reads are mapped to the reference genome using Tophat (C. Trapnell et al., 2009). Second, the reads are concatenated using Cufflinks (C. Trapnell et al., 2010) and the RABT (Reference Annotation Based Transcript) assembly technique is used for this purpose (Roberts et al., 2011). This results in a good accuracy for finding novel genes and splice isoforms when a high quality sequence reference exists for that genome. The assembled transcripts from all samples are merged using Cuffmerge and are compared with the reference genome using the Cuffcompare tool to find known and novel genes, known and novel isoforms, and transcripts expressed from intergenic regions. Third, the reads from Tophat and merged assemblies from Cuffmerge were used as an input for Cuffdiff2 (C. Trapnell, Hendrickson, et al., 2012). Cuffdiff2 in the time course mode was used for differential expression analysis of individual transcripts within the RNA-seq data and the bioinformatic analysis pipeline is presented in Figure 3.1.

Cuffdiff2 is an excellent isoform-based differential expression analysis tool (C. Trapnell, Hendrickson, et al., 2012). However, we also explored two other leading tools for AS analysis. SpliceGrapher is also an isoform-based AS analysis tool claimed to be

superior to Cuffdiff2 as it minimizes the identification of false positives (Rogers, Thomas, Reddy, & Ben-Hur, 2012). However, closer examination revealed that SpliceGrapher does not consider the non-canonical splice sites that are frequently found in plants (Filichkin et al., 2010; B. B. Wang & Brendel, 2006) and the resulting isoforms are considered false positives. As such, SpliceGrapher is limited to the identification of known plant transcripts without the potential of retrieving novel transcripts originating from non-canonical splicing. Unlike Cuffdiff2 and SpliceGrapher, DEXSeq is an exon-based tool for AS analysis and it was not used because it is intended for differential expression of individual exons and introns rather than whole transcripts (Anders et al., 2012). Accordingly, Cuffdiff2 (C. Trapnell, Hendrickson, et al., 2012) was further used for differential expression analysis of transcripts in developing soybean embryos, while SpliceGrapher (Rogers et al., 2012) was used to visualize selected isoforms based on existing gene models and Cuffdiff2 data because of its superior graphical isoform representation.

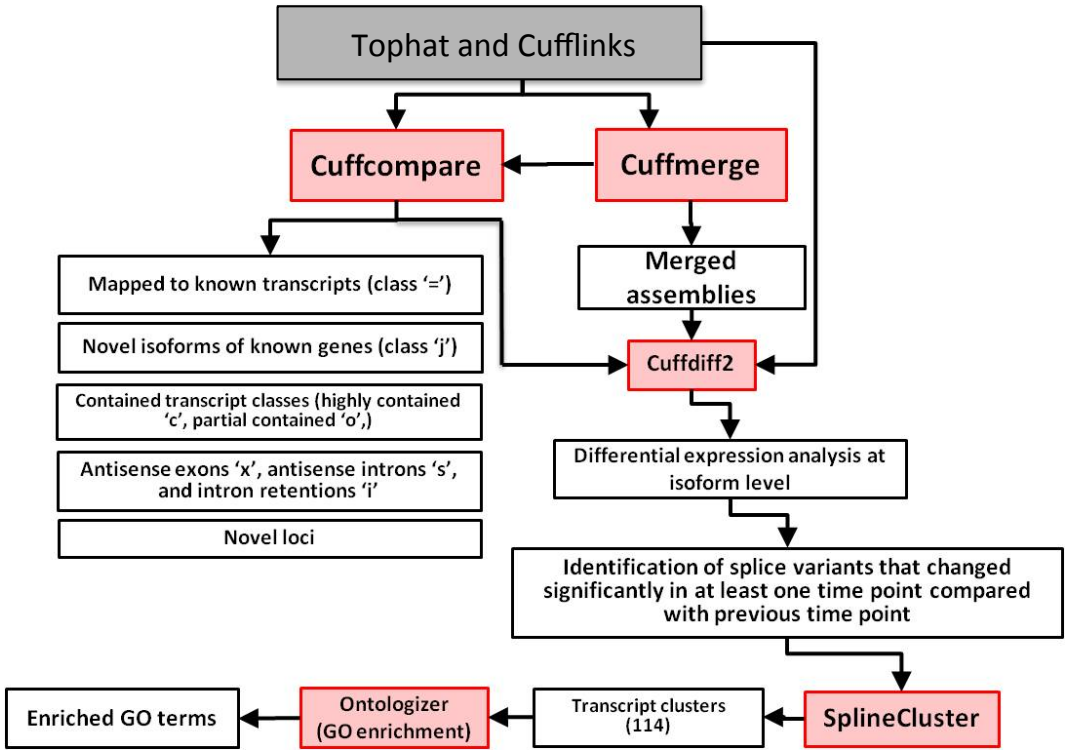

Figure 3.1. Flowchart of bioinformatic analyses used for differential expression of splice isoforms and subsequent data mining. Tools are shown in red boxes.

Cuffdiff uses a set of 12 class codes assigned in Cuffcompare to categorize assembled transcripts obtained from Cufflinks (C. Trapnell et al., 2010). Briefly, these class codes serve as a basis for information about the structure of the various assembled transcripts with reference to transcripts with well-characterized splicing patterns (class '='). It is noteworthy that assigning class codes to the transcripts is prioritized, such that when an isoform has a novel splice junction it is classified as class 'j', although its structure may fall into other lower priority classes as well. Class 'j' transcripts are potentially novel isoforms, in that they have at least one novel splice junction and at least one splice junction shared with the reference transcript. Class 'o' transcripts are assembled transcripts that show exonic overlap with the reference transcript, but do not fall into other higher priority class such as 'c' or 'j'. Class 'c' stands for contained and is used when a transcript has a high exonic overlap with a known transcript. Class 'c' was not observed in our significantly differentially expressed transcripts, but has a high priority among class codes. Transcripts in classes 'x' and 's' have exonic and intronic overlap, respectively, with the reference transcript on the opposite (antisense) strand. Class 'i'

transcripts are those, for which some sequence falls entirely within a reference intron. These transcripts are representative of intron retention events, in which the transcript did not fall into other higher priority classes. We named new assembled transcripts based on the number of knows transcripts for each gene based on the Gmax_189 version of the soybean reference genome. For example, if the number of known transcripts for the gene X is 2 (X.1 and X.2) then the first novel transcript will be designated X.N3.

3.2.2 AS and Clustering Analyses

For further investigation of the 13253 differentially expressed transcripts, we clustered them using SplineCluster (Heard et al., 2005). SplineCluster is a model-based hierarchical co-clustering tool for analyzing gene expression data fitting statistical regression models to a gene expression time series data. Therefore, SplineCluster can be used to group the transcripts with a common expression pattern over time (Heard et al., 2005). Because we were interested in the changing behavior of transcript AS over time during soybean embryo development rather than actual expression values, the FPKM values originating from Cuffdiff2 were scaled to have an arbitrary mean of 10000. The resulting expression values were clustered using SplineCluster, which yielded 114 clusters with the prior precision of 10^{-10} and reallocation sweeps of 10^{13}. In order to further group the genes in each cluster based on their functional description, we carried out Gene Ontology (GO) enrichment analysis using Ontologizer (Bauer, Grossmann, Vingron, & Robinson, 2008). The "Topology Elim" option in conjunction with "Westfall-Young Single Step" were used for GO enrichment analysis with reallocation sweeps of 10000. GO enriched terms with the p-value < 0.05 were identified as statistically significant and were subjected to further data mining.

3.2.3 Quantitative Real-Time PCR (qPCR) Validation of Gene and Isoform Expression[1]

Thirteen genes and eight splice isoforms representing four genes were selected to validate the observed RNA-seq-based gene and isoform expression changes by qPCR at

[1] Eva Collakova and Yihui Fang performed experiments in this chapter.

time points of interest, specifically, those at which transcript abundances changed significantly compared to the previous time point. Selection of genes and isoforms for validation was based on the trends of expression changes, relative expression strength, and diverse functions in metabolism and acquisition of desiccation tolerance. Two genes (Glyma14g00360 and Glyma13g16500) that showed stable expression, but different abundances (high and low) were used as controls. One of them (Glyma14g00360) was used as an internal control for qPCR normalization. Specific primer pairs were designed to amplify selected genes as well as unique regions of each selected isoform by using SeqBuilder in LaserGene Core Suite Application (v.10) (Burland, 2000). The melting temperatures of these primers were between 57°C and 62°C.

Total RNA was isolated as described previously (E. Collakova et al., 2013). cDNA was prepared as follows: The thermal cycling program involved an initial 10-minute incubation at 25°C, followed by a 30-minute reverse transcription step at 48°C, and, then, a 5-minute enzyme inactivation step at 95°C. 200 ng of cDNA was used as a template in a 20-µL reaction using SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's recommendation and qPCR was performed using an ABI 7500 Series RT-PCR System (Applied Biosystems). Three biological replicate reactions along with one non-template control were performed for each sample. The 2^{-DDCT} method was used to determine the relative levels of gene and isoform expression (Livak & Schmittgen, 2001). Briefly, gene and isoform expression was first normalized to that of the internal control in every sample. Normalized gene and isoform expression in every sample was then compared to the sample involving the first selected time point to obtain relative amount of gene and isoform expression. In general, trends obtained with RNA-seq were consistent with qPCR, with the exception of Glyma08g24420 (WRINKLED1) that showed a very low or non-existent expression with qPCR (C_T > 35 in a 40-cycle amplification program) probably due to sub-optimal properties of primer pairs used. In Arabidopsis, the *WRINKLED1* gene is known to be expressed prior to oil accumulation, followed by a decrease during the early stages of seed filling (Schmid et al., 2005), which is consistent with our RNA-seq data.

3.3 Results and Discussion

3.3.1 Global Assessment of AS in Developing Soybean Embryos

The soybean genome contains 54175 genes and 73320 known transcripts, suggesting that, on average, a single gene will encode about 1.35 transcripts (Goodstein et al., 2012; D. Grant, Nelson, Cannon, & Shoemaker, 2010; Schmutz et al., 2010). It is clear that some genes will only encode a single, while others multiple transcripts and we first aimed to assess the extent of AS in developing soybean embryos. Overall, 47331 genes were expressed in developing soybean embryos, giving rise to 217371 total transcripts (many of them novel) and, on average, over 4.6 transcripts per gene. This is a larger number than anticipated based on the existing soybean genome data. However, we need to take into consideration that part of soybean embryo development involves desiccation-related processes similar to drought and salt stress and that AS is induced during stress (He et al., 2007; Hirayama & Shinozaki, 2010; Mastrangelo et al., 2012; Matsukura et al., 2010; Mazzucotelli et al., 2008). In fact, the majority of AS events took place during the late maturation and desiccation stages of soybean embryo development as nearly 50% of differentially expressed transcripts showed an increase in expression in the last 2 – 3 time points. Most importantly, the frequency of 1395 differentially expressed genes that had two or more transcripts showing changes in expression at least in one time point was the highest in the last time point (Figure 3.2), suggesting that AS in general was induced during embryo maturation, dormancy and desiccation tolerance acquisition.

The total number of isoforms (encoded by 15368 genes) that showed differential expression during embryo development was 16,915 (p-value < 0.05 and FDR < 5%). Most genes that exhibited changes in their expression (13,973) in developing soybean embryos had at least one changing isoform and only one gene had six transcripts that showed statistically significant ($p < 0.05$) changes in expression in at least one time point. Among 13,973 genes that had only a single differentially expressed transcript, 10311 genes were also represented by at least one other isoform that was expressed in developing soybean seeds in a stable manner (Table 3.1). Because one isoform was changing its expression, while the other one was not, these isoforms showed different patterns of expression and were considered as differentially expressed relative to each other. Together with 1395 genes that encoded at least two differentially expressed transcripts, 11706 genes were considered to be alternatively spliced. Only 3,662

differentially expressed genes were represented by single transcripts and most these transcripts were a result of conventional splicing and not AS, as only 15 belonged to one of the other classes (j, i, and o; Table 3.1).

Table 3.1. Statistics for 16915 transcripts that showed significant expression changes in at least one time point compared to the previous time point (*p*-value < 0.05 and FDR < 5%) in developing soybean embryos. Transcripts that were expressed, but did not show differential expression during embryo development are not included among 10311 and 2942 (total 13253 transcripts originating from 11706 genes) transcripts listed in this table. Classification of known, novel, and other classes of splice isoforms was based on class definitions assigned by Cuffcompare (C. Trapnell et al., 2010).

		One transcript changed and at least one other did not change	Single transcript	At least two isoforms changed
	Genes	10311	3662	1395
Codes		**Corresponding isoforms that showed differential expression**		
=	Known transcript	8466	3647	1421
j	Novel transcript	1525	12	1313
i	Intron retention	3	1	0
o	Partial overlap with known reference transcript	183	2	68
x	Antisense exon	128	0	116
s	Antisense intron	6	0	24
	Total	10311	3662	2942

Our significantly differentially expressed transcripts fall into 6 Cuffcompare classes (=, j, o, x, s, i) (C. Trapnell et al., 2010) and the majority (75%) of these transcripts were previously known (class =). The remaining 25% are divided among these other classes (Table 3.1). Detailed analysis of these transcripts revealed that 10311 genes that had one isoform changing significantly along with at least one stably expressed isoform were templates mostly for known transcripts (82%), while only 15% of novel isoforms were observed (Table 3.1). Only 3% of these transcripts belonged to one of the other four

Cuffcompare classes represented in our data set (i, o, x, and s). For 1395 genes that had more than one differentially expressed transcripts, known and novel isoforms were about equally frequent and represented the majority of transcripts, while other classes had only 7% representation (Table 3.1). This is not unexpected, as only one or two known isoforms have been reported for most soybean genes . As such, genes encoding more than one isoform are by definition enriched in this group. Collectively, these observations suggest that the majority of genes that were differentially expressed in developing soybean embryos produce multiple transcripts due to AS.

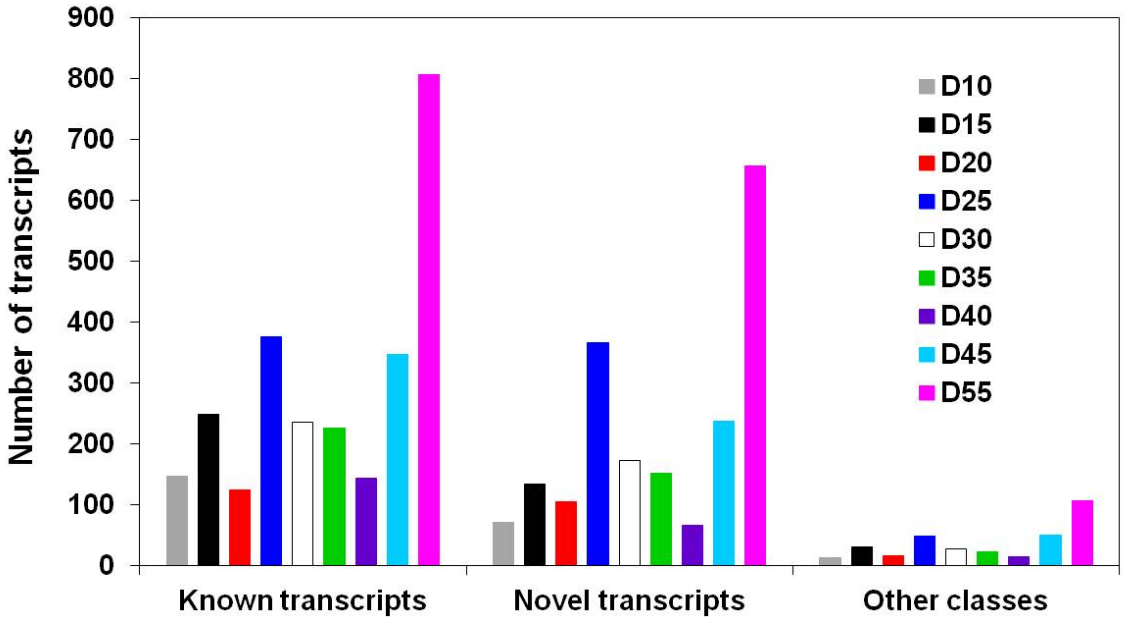

Figure 3.2 Distribution of known, novel, and other class transcripts during soybean embryo development.

3.3.2 Differential Expression of Related Isoforms Involved in CCNM and Maturation during Soybean Embryo Development

Co-expression analysis was performed on 15368 transcripts showing differential expression in a time-series-dependent manner using the Bayesian co-clustering SplineCluster tool (Heard et al., 2005), resulting in 114 clusters. As mentioned above, most genes (13973) were represented by a single isoform showing differential expression and each transcript was present only in one of 114 clusters, while 1395 genes were represented by 2942 isoforms. Some of these isoforms were co-expressed, while others

showing distinct expression patterns. Embryo development is characterized by three main phases: (i) embryo axis development (Jeong et al., 2012; Lau et al., 2012; Brandon H. Le et al., 2007; Minako Ueda & Laux, 2012), (ii) accumulation of seed storage compounds (D.K. Allen et al., 2009; D. K. Allen et al., 2007; Bates et al., 2009; Sébastien Baud et al., 2002; Borisjuk et al., 2005; Hill & Rawsthorne, 2000; Iyer et al., 2008; Schwender & Ohlrogge, 2002), and (iii) maturation and acquisition of desiccation tolerance (Angelovici et al., 2010; Blochl et al., 2005; R. Finkelstein et al., 2008; Gutierrez et al., 2007). These processes are essential for successful sexual propagation of plants. Therefore, genes represented by multiple transcripts involved in soybean CCNM and acquisition of dormancy and desiccation tolerance were closely examined. Since ABA-related processes are crucial for dormancy acquisition (Shu et al., 2013), as well as the action of other hormones, such as auxin (X. Liu et al., 2013), splicing events in those categories were also queried. Genes associated with splicing itself were examined, since there is evidence that these transcripts are often the product of stress-mediated splicing themselves (Duque, 2011), giving rise to specific isoforms that have the potential for roles in regulation.

3.3.3 Central Carbon and Nitrogen Metabolism

The majority of isoforms showing changes in expression were associated with various largely unknown cellular and regulatory processes and only a small proportion of these transcripts were encoded by metabolic genes. This is in agreement with the observation that genes encoding CCNM enzymes represent only a small proportion of the total genes in plant genomes (Eva Collakova et al., 2013; Dal'Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010; Saha, Suthers, & Maranas, 2011). We previously identified two different sets of homologous CCNM genes involved in two different types of metabolism in developing soybean embryos; metabolism supporting: (i) cell division in differentiating and young embryos and (ii) cell elongation and accumulation of seed storage compounds during seed filling (E. Collakova et al., 2013). The question is whether the pre-mRNA molecules originating from these different sets of genes are subjected to AS and whether the resulting isoforms also show distinct expression patterns.

Every cluster contained some metabolic genes, but only some clusters displaying specific expression patterns were enriched with many alternatively spliced transcripts encoding different enzymes and metabolite transporters involved in CCNM pathways including photosynthesis, photorespiration, respiration, glycolysis, gluconeogenesis, tricarboxylic acid (TCA) cycle, pentose phosphate pathway, and amino acid and lipid metabolism. Results regarding GO enrichment in specific metabolic processes are summarized in brief in Table 3.2. Clusters 1 – 26 (black in Table 3.2) were consistent with the gene expression patterns of CCNM genes predicted to support cell division and differentiation in young embryos. This trend is characterized by an initial high gene expression at day 5 followed by decreases of variable slopes until day 15 or 25 and a subsequent stable very low or no expression for the remainder of the developmental time course. Isoforms present in cluster 14 also showed a moderate increase in expression from day 40 to day 55, suggesting that they are also involved in CCNM supporting maturing and desiccating embryos.

We identified other clusters showing trends that resembled the expression pattern of cluster 14, but exhibited subtle initial decreases and quite considerable final increases in expression patterns (clusters 55 – 59, 72, 73, and 99 – 101; shown in gray in Table 3.2). These clusters (including cluster 14) were enriched in transcripts encoding enzymes and transporters involved in photosynthesis, photorespiration, gluconeogenesis, glycolysis, TCA cycle, respiration, and amino acid and lipid metabolism. Many of these processes, particularly gluconeogenesis, glycolysis, TCA cycle, respiration, and amino acid metabolism are coupled with oil degradation occurring during the maturation of oilseeds. These processes provide substrates and energy for growth and continuing seed storage protein accumulation in drying seeds when the connection between the maternal tissue and the seed is gradually severed (Sébastien Baud et al., 2002; Chia et al., 2005). However, it is also possible that some of these alternatively spliced transcripts accumulate and are stored in RNA-processing bodies (P-bodies) or stress granules (Bogamuwa & Jang, 2013; Davies, Stankovic, Vian, & Wood, 2012; J. Xu & Chua, 2009) in drying embryos to provide carbon and energy sources during seed germination, which involves the same types of metabolic processes (Sébastien Baud et al., 2002; Chia et al., 2005; Eastmond & Graham, 2001; Graham, 2008; Penfield et al., 2005). In

addition, photosynthetic activity is non-existent in yellow drying embryos and photosynthesis-related transcripts accumulating in these embryos could be pre-made and stored for germination. Such surprising, but not entirely unexpected, potential multiple functions in different types of CCNM for these isoforms remain to be confirmed. Nevertheless, the use of the same transcripts for CCNM at completely different stages of plant development (early and late stages of seed development as well as seed germination) suggests existence of common regulatory components of AS and subsequently CCNM during these different developmental phases.

Table 3.2. Major metabolic processes of CCNM and relevant clusters enriched in transcripts encoding a variety of proteins involved in these processes. GO enrichment analysis was performed on isoforms in all 114 clusters (*p*-value < 0.05). Only clusters displaying three basic trends consistent with early (initial decrease followed by stable low or no expression; black), seed filling (moderate increase followed by a moderate decrease; red), and maturation (stable low or no initial expression followed by a final increase; blue) CCNM are shown. Clusters showing conceptually similar trends based on visual assessment have the same color. Bold numbers were used to distinguish subtle differences in similar trends, while using the same color. Abbreviations: AA, amino acid, PS, photosystem, TCA, tricarboxylic acid

Processes	Clusters
Light harvesting	18
PS I	18, **20**
PS II	**11**, 18, **20**, 104
Light reactions	**12**, **16**, 17, 18, **57**, **73**
Electron transport	**11**, 17 - 19, **20**, **27**, **56**, **100**
PS assembly and stabilization	**11**, 17 - 19, **20**, **27**, **72**, **73**, 76, 77, 101
Calvin cycle	**5**, **11**, 17
Photorespiration	**22** - **24**, **73**
Respiration	**22**, **28**, **57**
Gluconeogenesis	**14**, **16**, **23**, **24**, **55**, **59**, **84**
Glycolysis	**16**, 17, **21** - **24**, **57**
TCA cycle	**14**, **21**

Pentose phosphate pathway	**16**, 17 - 19, **20**, **57**, **73**, **100**
AA metabolism and transport	**1**, **5**, **7**, **10** - **12**, **14**, **15**, **16**, 17 - 19, **20** - **24**, **37**, **38**, **43**, **45**, **52**, **56**, **57**, **72**, **73**, 76, **80**, **83**, **84**, **87**, **88**, 92, 93, 96, **99**, **100**, 102, 105, 110
Lipid metabolism and transport	**3**, **10** - **13**, 17, 19, **20** - **24**, **28**, **43**, **45**, **47**, **58**, **59**, **79**, 90, 93, 96, 102, 104, 105, 112

Clusters 27, 28, 37, 38, 43, 45, 47, and 52 (red in Table 3.3) share a common overall trend of an initial moderate increase from day 5 to 20 or 25, followed by a gradual moderate decrease in isoform expression. This trend is consistent with the trend of CCNM genes potentially involved in seed filling (E. Collakova et al., 2013), but only specific processes and fewer genes are represented here than in the case of the potential cell division CCNM genes. In the case of photosynthesis and respiration, transcripts encoding several light-harvesting complexes and electron carriers are present in these clusters. In contrast, amino acid and lipid metabolism and transport GO categories dominate these clusters from the perspective of metabolism, however, they are under-represented relative to cell division-related CCNM (Table 3.3).

Interestingly, ACT domain-containing proteins represented by the transcripts Glyma10g42580.N6, Glyma18g52120.1, Glyma13g09310.4, and Glyma15g00560.N7 were also identified among AS-derived isoforms potentially regulated by amino acids and/or involved in amino acid metabolism. ACT (Asp kinase, Chorismate mutase, and TyrA) domain enables binding of amino acids to proteins and subsequently allosteric regulation of enzyme activity (Chipman & Shaanan, 2001; G. A. Grant, 2006; Liberles, Thorolfson, & Martinez, 2005). The Arabidopsis genome contains at least 12 copies of these proteins that can have functions as sensors or in the regulation of metabolism (Heo & Sung, 2011; Hsieh & Goodman, 2002). Glyma10g42580 (ACR12 in Arabidopsis) was expressed as two isoforms (5 and N6) exhibiting completely different expression patterns (clusters 23 and 37, respectively), while the rest of these ACR genes had at least one other stably expressed transcript.

Table 3.3. Isoforms present in clusters 27, 28, 37, 38, 43, 45, 47, and 52 representing CCNM supporting seed filling. Amino acid and lipid metabolism is dominating these clusters.

Process	Isoforms	Annotation
Photosynthesis	Glyma05g24660.N2; Glyma18g03220.3; Glyma02g47560.2; Glyma20g35530.1;	Light-harvesting complexes
	Glyma07g21150.N2; Glyma11g08230.1; Glyma05g03730.1; Glyma12g32680.1	Electron carriers
	Glyma15g40450.1	O_2-evolving complex
Respiration	Glyma10g08480.1	Kinesin-like protein 1
	Glyma08g00880.N4	NADPH/respiratory burst oxidase protein D
Amino acid metabolism	Glyma19g27500.1; Glyma11g38130.1; Glyma18g06840.N6; Glyma06g13280.N3; Glyma14g32500.1; Glyma19g28770.1	Asp- and Glu-family enzymes
	Glyma15g05630.1	Ser decarboxylase
	Glyma10g35580.N13; Glyma12g07720.N10	Aromatic amino acid enzymes
	Glyma04g42190.1; Glyma07g30500.N8; Glyma19g29880.N2	Branched-chain amino acid enzymes
	Glyma13g24580.N25; Glyma02g42800.1; Glyma09g37270.N2	Amino acid transporters
Lipid metabolism	Glyma05g24650.N3; Glyma17g07480.N2; Glyma05g36910.1; Glyma12g13020.1; Glyma07g18370.N4; Glyma06g17640.1; Glyma17g03070.N3	Acyl-CoA-related enzymes
	Glyma18g06950.1; Glyma14g37350.N3	Fatty acid desaturases

Numerous clusters showing a low, but stable, or no expression from day 5 to day 35 - 45, followed by a substantial gene expression increase of varying slopes were enriched in a large number of isoforms involved primarily in CCNM and transport involving amino acid and lipid metabolism (Table 3.2). This is expected, as during late developmental stages, embryos degrade their chlorophyll through ABA-mediated signaling and photosynthesis ceases to function (Delmas et al., 2013; Nakajima, Ito, Tanaka, & Tanaka, 2012). This means that these embryos rely on nutrients of maternal origin as well as on internal lipid degradation to provide carbon and energy sources for ongoing seed storage protein accumulation and metabolism relevant to maturation and acquisition of desiccation tolerance (Sébastien Baud et al., 2002; Chia et al., 2005; Munier-Jolain et al., 1998). Because of a large number of transcripts related to amino acid and lipid metabolism and transport that were present in these clusters, gene expression relevant to this type of CCNM and metabolite transport appears to be regulated by AS. GO categories involving other metabolic processes were not highly enriched, suggesting that, similarly to seed filling-related CCNM, these processes are also not globally regulated by AS.

AS has the potential to introduce or remove additional sequences to or from proteins that could result in truncated proteins or changes in protein localization, stability, interactions with other molecules, regulation, and/or biological function of the resulting proteins (Barbazuk et al., 2008; Marden, 2008; E. I. Severing et al., 2012; Syed et al., 2012). When considering CCNM and metabolite transport, only a few RNA isoforms showed completely different expression patterns, while the majority belonged to the same or similar clusters. Our focus here is on specific examples of isoforms showing different expression trends and representing diverse metabolic processes. As expected, many novel transcripts that displayed a strong differential expression when compared to the known transcripts contained a premature stop codon, and could encode only truncated peptides, most of which would lack functional and regulatory domains. The above-mentioned ACR12 gene encodes a plastidic amino acid transporter 1-like protein and it was represented by two isoforms Glyma10g42580.5 and N6. While the first of these transcripts encodes a protein containing two ACT domains, the latter contains a stop codon. Glycolytic 2,3-biphosphoglycerate-independent phosphoglycerate mutase-related

isoform Glyma08g28530.1 (cluster 25) contains a metalloenzyme superfamily domain proposed to have function in metal binding and catalysis. AS lead to the appearance of a premature stop codon in Glyma08g28530.N2 (cluster 45) and a substantial truncation of this domain. These two isoforms show completely different expression profiles. While Glyma08g28530.1 exhibits a gradual decrease throughout the time course, in contrast, the novel isoform shows an initial increase from day 5 to 20, followed by a decrease from day 30 to the end. In another example, Glyma04g14650.1 encodes a short acyl-CoA-binding protein containing overlapping acyl-CoA-binding and CoA-binding sites, both composed of a number of amino acids spread throughout the protein. This isoform displays a gradual slow decrease in expression (cluster 24). However, AS yielded a novel isoform Glyma04g14650.N2 (cluster 46) that does not appear to produce meaningful protein products, but peaks at day 30. Premature stop codons in transcripts lead to premature protein synthesis termination and subsequent peptide/protein degradation and some of these transcripts could undergo NMD (Filichkin et al., 2010; James et al., 2012; Kalyna et al., 2012; Palusa & Reddy, 2010; B. B. Wang & Brendel, 2006). However, this remains to be confirmed on the individual basis along with the biological functions of these transcripts and resulting peptides.

However, in some instances, AS could lead to the production of large truncated proteins. Glyma20g32260.1 (cluster 56) encodes a full-length vacuolar amino acid transporter family protein with an intact amino acid permease domain SdaC, while this domain is truncated in Glyma20g32260.N3 (cluster 112) encoding the first half of the full-length protein. Only Glyma20g32260.1 shows an initial decrease in expression from day 5 to 25 and both transcripts are expressed during the late maturation phases of soybean embryo development. With truncated proteins such as these, it is difficult to predict their function, but they have a potential to sequester substrates and cofactors or interact with their full-length counterparts or other proteins and influence their activities.

Three differentially expressed isoforms Glyma19g05120.2, N3, and N4 (clusters 104, 20, and 28; Figure 3.3) are predicted to encode a full-length 6-phosphogluconate dehydrogenase based on our results. Phosphogluconate dehydrogenase is a key enzyme in the oxidative pentose phosphate pathway and is active in the chloroplast stroma during seed filling in soybean (D.K. Allen et al., 2009; D. K. Allen et al., 2007; Bates et al.,

2009; Borisjuk et al., 2005; Iyer et al., 2008; Sriram et al., 2004). In this case, AS resulted in introduction of alternative start codons and subsequently potentially distinct sub-cellular localization of this enzyme. Based on the SoyBase database, the respective proteins are found in plastids, peroxisomes, and cytosol, though only the existence of the Glyma19g05120.2 isoform was known previously (D. Grant et al., 2010). In the yeast *Candida albicans*, AS is also responsible for dual targeting of this enzyme to the cytosol and the peroxisomes (Strijbis, van den Burg, Visser, van den Berg, & Distel, 2012). The new isoforms N3 and N4 could encode these other differentially localized isoforms, with Glyma19g05120.N4 being cytosolic, as it lacks any transit peptide. In contrast, the subcellular location of the other two protein isoforms is unclear. Clusters 20 and 28 have somewhat similar trends, showing a moderate initial increase or steady expression during seed filling, followed by a subsequent decrease and low or no expression after day 35 (Figure 3.3). Based on this expression pattern, it is tempting to speculate that N3 is the plastidic isoform as it may be needed to provide ribulose-5P for the Rubisco bypass during seed filling (D.K. Allen et al., 2009; D. K. Allen et al., 2007; Schwender et al., 2004). Glyma19g05120.2 is expressed only during the late maturation and desiccation phases, starting at day 40 and could encode the peroxisomal isoform to generate NADPH for lipid degradation and antioxidant regeneration. 6-Phosphogluconate dehydrogenase was also found in pea peroxisomes and proposed to provide reductant for peroxisomal metabolism and recycling of oxidized ascorbate and glutathione (Corpas et al., 1998) and this soybean isoform could have similar functions.

Phosphogluconate dehydrogenase is not the only example of an enzyme targeted to different compartments and originating from AS, but the roles of the resulting transcript isoforms are not apparent. Glyma02g47560.1-encoded LHCB2.1 (cluster 18) functions in PSII antennae and has a chloroplast targeting sequence. AS also yielded Glyma02g47560.2 (cluster 28) lacking the chloroplast targeting sequence. However, both proteins have an identical chlorophyll a/b binding domain, but only Glyma02g47560.1 is targeted to chloroplasts. Similarly, putative 5,10-methylene tetrahydrofolate dehydrogenase/cyclohydrolase predicted to function in folate interconversions originates from two isoforms having different transit peptides, Glyma09g39790.1 and Glyma09g39790.N2 (clusters 15 and 50, respectively). The first transcript encodes a

plastidic isoform of this enzyme, while the localization of the second one remains to be determined. Regardless of the localization, these isoforms show completely opposite expression profiles. Glyma09g39790.1 displays an initial decrease in expression from day 5 to 15, then a very low stable expression as opposed to Glyma09g39790.N2, which is only expressed during days 35 through 55. Subcellular locations and functions of the proteins encoded by these new isoforms remain to be elucidated.

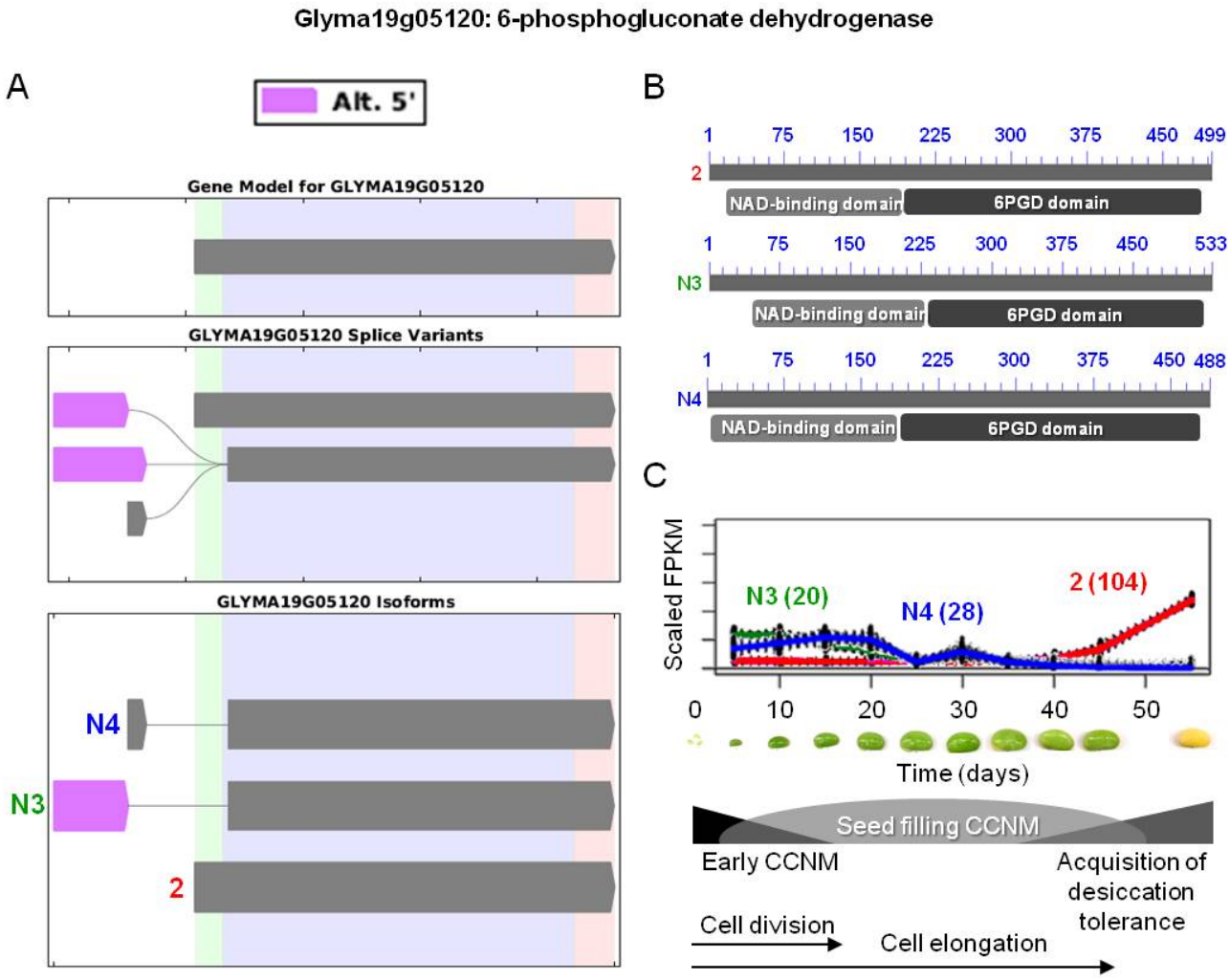

Figure 3.3. AS of 6-phosphogluconate dehydrogenase involved in oxidative pentose phosphate pathway. (a) SpliceGrapher representations of individual isoforms detected during soybean embryo development. Only Glyma19g05120.2, N3, and N4 (corresponding to 2, N3, and N4) were differentially expressed. Alternative 5' splicing yielded these novel isoforms. (b) Proteins resulting from these three transcripts with the respective positions of NAD-binding domains (pfam03446) and 6-phosphogluconate dehydrogenase (6PGD) C-terminal domains (pfam00393). Domain-related information was obtained from NCBI. Numbers in blue represent amino acid residues in the individual proteins. (c) Overlayed expression profiles of Glyma19g05120.2, N3, and N4 obtained from SplineCluster.

Numbers shown in parenthesis represent the isoform clusters. Representative developing soybean embryos are shown for each time point along with a temporal representation of processes occurring during embryo development.

The most intriguing and rare scenario is when the resulting proteins contain or lack domains that have protein-protein interaction, catalytic, and/or regulatory functions, as this could lead to proteins with potentially novel functions and regulatory capabilities. Such protein isoforms could have similar or completely different expression patterns, which, together with systematic detailed domain analysis, can assist in dissecting their functions in specific processes. For example, if such protein isoforms have mutually exclusive patterns they will not be able to physically interact. We were not able to identify any novel isoforms that would encode proteins with additional novel domains. However, random sampling of metabolic isoforms present in different clusters that showed similar trends revealed several genes where domains were modified such that they could have distinct catalytic, regulatory, or protein-protein interaction capabilities.

In soybean, 3-ketoacyl-CoA thiolase 3 is a fatty acid b-oxidation enzyme encoded by isoforms Glyma10g24590.1, 2, and 3. Isoforms Glyma10g24590.1 and 3 encode these enzymes with alternate C-termini possessing the acetyl-CoA C-acyltransferase multidomain containing specific sites (Figure 3.4). Glyma10g24590.1 (cluster 93) has all expected residues in the thiolase active and dimer interface sites. In contrast, Glyma10g24590.3 (cluster 90) is missing one of three of the amino acid residues found in thiolases and two of twenty of the residues that make up the dimer interface domain. In Arabidopsis, the closest homolog of Glyma10g24590 is AT2G33150, a peroxisomal 3-ketoacyl-CoA thiolase, but another close Arabidopsis homolog AT5G48880 is also subjected to AS yielding both a peroxisomal and cytosolic isoforms (Wiszniewski et al., 2012). The AT2G33150-encoded enzyme is involved in seed dormancy and germination, as well as turnover of fatty acids during natural and dark-induced senescence (Kunz et al., 2009; Li-Beisson et al., 2013; Z. Yang & Ohlrogge, 2009). This enzyme also plays an important role in positive regulation of ABA signaling by acting downstream of a WRKY TF involved in ABA-mediated signaling, thus promoting embryonic nature of developing embryos and suppressing germination-related processes in Arabidopsis (Jiang, Zhang, Wang, & Zhang, 2011). This potential and unexpected dual role for this enzyme as a component of ABA signaling provides a connection between CCNM and acquisition of

desiccation tolerance processes during late maturation phases of oilseed embryo development.

3.3.4 AS of Splicing-Associated Transcripts, Acquisition of Dormancy and Desiccation Tolerance, ABA, and Other Phytohormone-Related Events

In Figure 3.5, a summary of GO enrichment results for categories related to dormancy acquisition and desiccation tolerance are shown, together with an indication of the clusters responsible for that enrichment. Interestingly, clusters reflecting splicing events early on in the time course of seed development contained isoforms of genes associated with seed dormancy. In particular, in clusters 9, 10, 15 and 21, isoforms of Arabidopsis homologs of regulatory genes (Glyma08g19650.2, soybean homolog of separase) were differentially expressed during this first phase of seed development, corresponding to the cell division phase, as shown in Figure 3.5. Isoforms of four transcripts whose homologs in Arabidopsis are EMB genes ("embryo lethal", EMB2656, 1968, and 2271) appeared in the clusters corresponding to early events in seed development. Clusters 60 and 85, corresponding to the late phase of seed development involving acquisition of dormancy and desiccation tolerance, were also enriched for the GO category "seed dormancy process", albeit with fewer genes overall. Mutant studies using the specific isoforms, which were differentially expressed over the time course, are needed to determine if, and how, these isoforms influence dormancy acquisition. In the case of, for example, Glyma05g35280.3 whose Arabidopsis homolog is AT3G24440 (VRN7) encoding a fibronectin, type III protein associated with chromatin modification and epigenetic effects, it would be interesting to identify its interacting partners, and whether only the third known isoform influences subsequent developmental processes. The same holds for Glyma01g02350.3 whose Arabidopsis homolog is IAA8, a gene associated with the regulation of organ formation. It is important to note that the information provided by GO comes from transcriptional studies, and no extensive reservoir of information on isoform expression is yet available. Thus, rigorous associations between developmental processes and splicing events have not yet been made, in the vast majority of cases. Nonetheless, the appearance, early on in seed development, of transcripts corresponding to genes

encoding regulators of organ formation, suggests a particular role for those isoforms in pattern embryo formation.

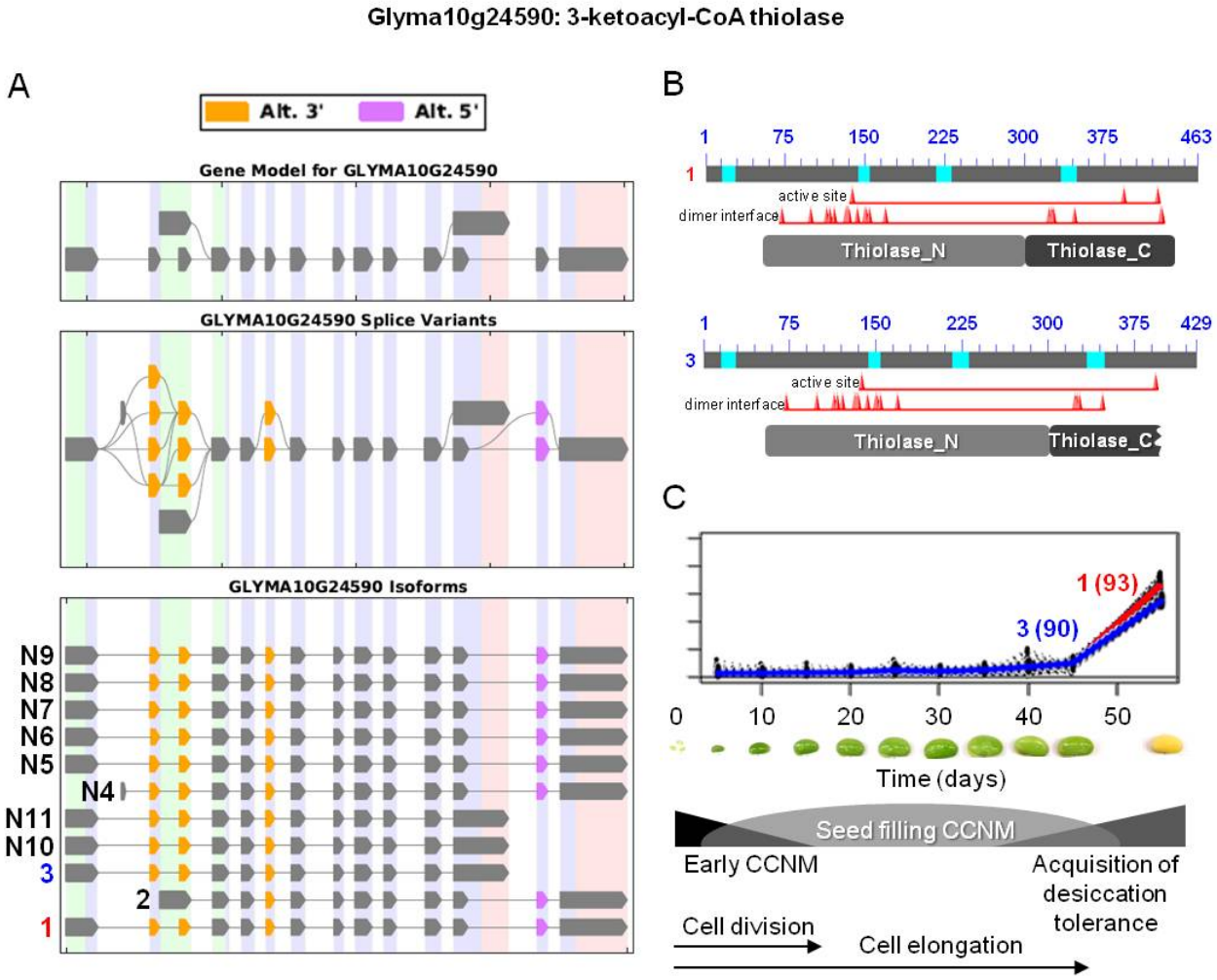

Figure 3.4. AS of 3-ketoacyl-CoA thiolase involved in fatty acid degradation during late maturation stages of soybean embryo development. (a) SpliceGrapher representations of individual isoforms detected during soybean embryo development. Alternative 3' and/or 5' splicing of many novel isoforms were identified. Glyma10g24590.1 and 3 correspond to 1 and 3 and were the only two transcripts displaying differential expression. (b) Proteins resulting from two isoforms of interest with the respective positions of active and dimer interface sites in the N- and C-terminal thiolase domains (pfam00108 and pfam02803, respectively), showing the positions of individual amino acid residues (red triangles). Bright blue sections of the protein show the sequences that were not included in domain analysis due to an amino acid composition-related bias. (c) Overlaid expression profiles of Glyma10g24590.1 and 3 obtained from SplineCluster.

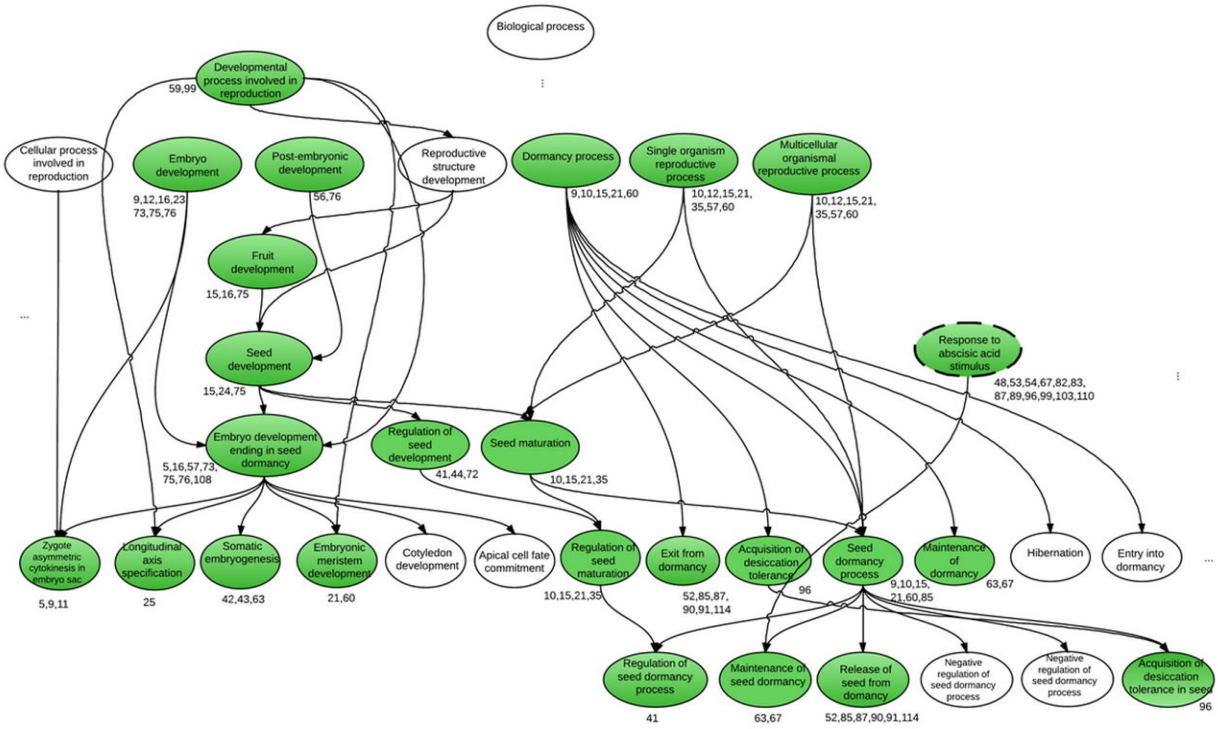

Figure 3.5. GO enrichment tree depicting processes associated with dormancy. Filtered enriched GO categories were searched for keyword 'dormancy' and the GO database (Ashburner et al., 2000) was used to identify the parents and children for each GO term. GO terms that were enriched significantly in genes involved in dormancy are shown in green. The clusters containing these dormancy-related genes are shown below the corresponding category.

GO categories associated with response to ABA were enriched in multiple clusters, from #48 onwards. All clusters that were enriched for ABA responses show a common pattern, with increases in differential expression towards the later part of seed development. This is, perhaps, to be expected, since ABA is well known as the phytohormone that confers dormancy on maturing seeds (Angelovici et al., 2010; Blochl et al., 2005; R. Finkelstein et al., 2008; Gutierrez et al., 2007). However, it is noteworthy that this response develops, not during the final stages of maturation, but much earlier, during seed filling stages. This result suggests that the dormancy induction process may be initiated on a different time scale than was previously thought. In keeping with this inference, the GO category "Maintenance of Dormancy" was enriched for differential expression of isoforms in clusters 63 and 67, rather than the clusters displaying a sharp rise at the far end of development, when dormancy acquisition has already occurred.

Two unexpected GO categories "Release of Seed from Dormancy" and "Exit from Dormancy" were enriched for differential splicing in a number of clusters from #52

onwards. Genes encoding ABA 8'-hydroxlyase, a member of the cytochrome P450 gene family and an ABA-degrading enzyme, and ABA 8'-hydroxylase-like proteins were over-represented in these clusters. On inspection, it appears that several isoforms among the group corresponded to full-length P450 proteins, and hence it would be expected that they are enzymatically active. However, some isoforms lacked a portion of the P450 domain due to a premature stop codon (Figure 3.6). Since ABA levels are maintained as dormancy is acquired, it seems unlikely that ABA degradation takes place during that time period. It is possible that the group of transcripts encoding these ABA-degrading enzymes are stored in either P-bodies or stress granules as we proposed for photosynthesis-related transcripts, to be translated upon imbibition and the initiation of germination.

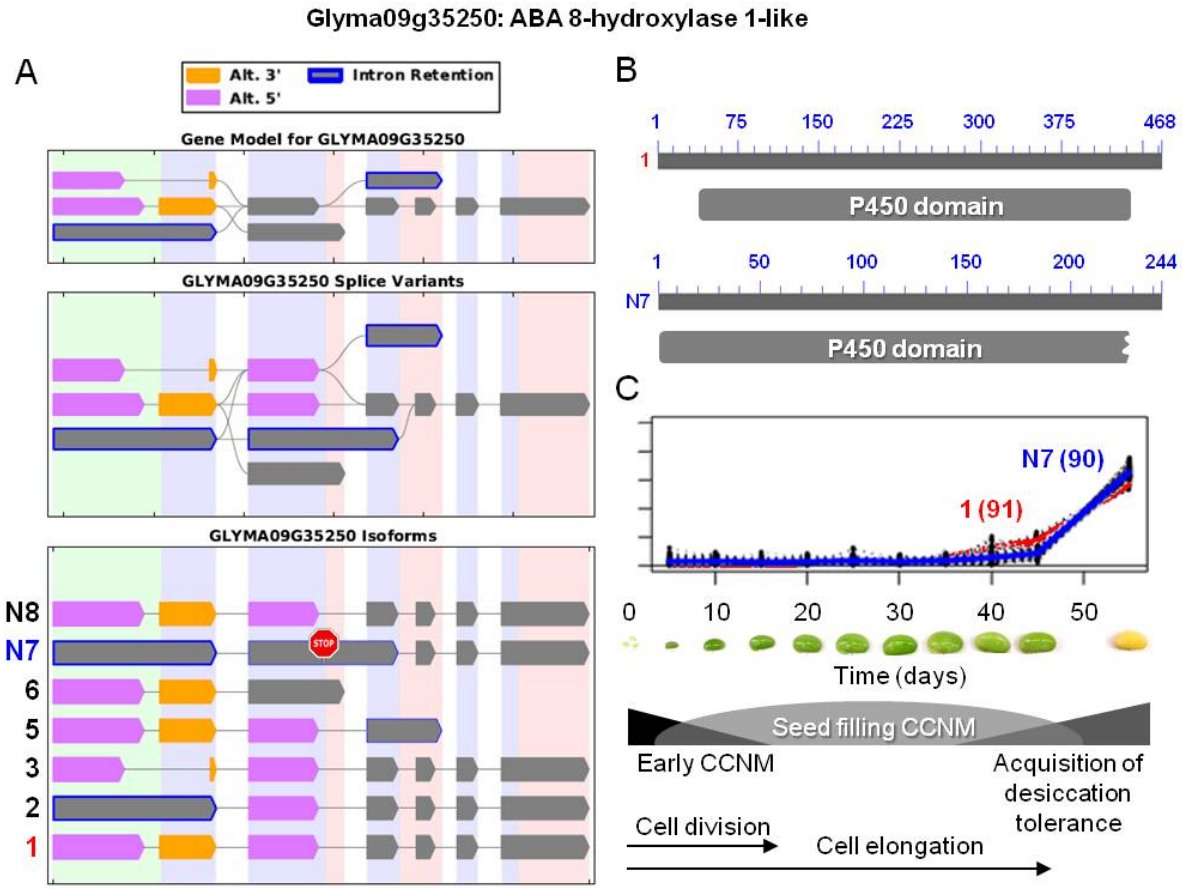

Figure 3.6. AS of ABA 8'-hydroxylase 1-like protein possibly involved in ABA degradation. (a) SpliceGrapher representations of individual isoforms detected during soybean embryo development. Alternative 3' and/or 5' splicing, and/or intron retention was observed in these isoforms. Glyma10g24590.1 and N7 corresponded to isoforms 1 and N7. Stop sign in N7 indicates a stop codon. (b) Proteins resulting from two isoforms of interest with the respective positions of the p450 domains (pfam00067). Isoform N7 contains a truncated p450 domain. (c) Overlayed expression profiles of Glyma10g24590.1 and N7 obtained from SplineCluster.

Figure 3.7 shows a GO tree corresponding to enrichment of terms related to splicing. All clusters that showed enrichment for splicing-associated processes for mRNA fell between clusters 55 and 82, corresponding, as in the case of the ABA-associated clusters, to the late phase of seed development. It is known that the splicing process itself is affected by abiotic stress (W. F. Li, Lin, Ray, Lan, & Schmidt, 2013; Palusa, Ali, & Reddy, 2007), and the acquisition of desiccation tolerance in seeds shares many gene expression events with stress responses reported in vegetative tissue (Verdier et al., 2013). Much AS associated with splicing factors occurred in cluster 70, including the appearance of specific isoforms of the soybean homologs of SR34a (a Ser/Arg-rich protein known to be involved in splicing, Glyma06g14060.2). RRC1, an RNA recognition motif (RRM)-containing protein, Glyma08g34030.4, as well as

Glyma04g07300.3, a homolog of AT4G32420 encoding a cyclophilin-like peptidyl-prolyl *cis-trans* isomerase family protein known to interact with SR proteins in pre-splicing events at the spliceosome (Lorkovic, Lopato, Pexa, Lehner, & Barta, 2004), were also present in cluster 70.

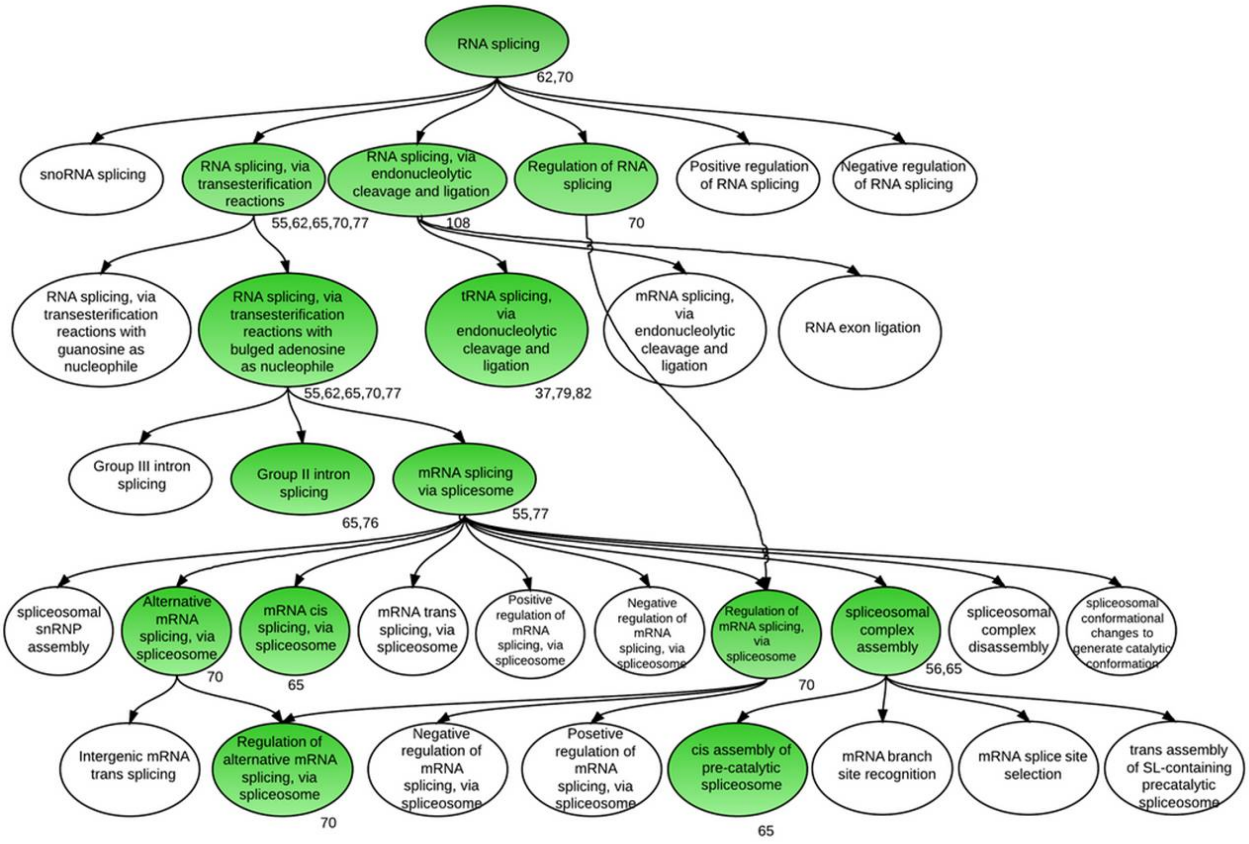

Figure 3.7. Tree displaying GO-enriched processes associated with RNA splicing.

3.4 Conclusions

Seed filling and acquisition of desiccation tolerance in maturing oilseed embryos represent important developmental stages crucial for acquiring and maintaining seed viability. AS plays roles in post-transcriptional regulation of developmental, metabolic, and stress-related processes in plants. Bioinformatic analyses enabled identification and global analyses of various AS events in developing soybean embryos. As expected, many pre-mRNAs encoding enzymes and proteins involved in diverse aspects of CCNM and signaling pathways of desiccation-related processes as well as components of the actual splicing machinery were subjected to AS. Late maturation stages of soybean embryo development were characterized as having a larger proportion of AS-derived isoforms

than other stages, which can be explained by induction of splicing as part of maturation and desiccation tolerance acquisition pathways. These pathways involve stress-like hormonal responses and signaling pathways known to be regulated at the AS level. Detailed analysis of selected isoforms involved in CCNM and ABA-related metabolism revealed possible roles for AS in regulating activities, subcellular localization, and protein-protein interactions of the resulting proteins. These are just first steps of analysis of the large data sets generated in this study that will provide a vast resource for further data mining and testable hypothesis generation.

4 CodeWise[1]

This chapter is partly extracted from (Aghamirzaie et al., 2015), which is an open access journal.

Aghamirzaie, D., Batra, D., Heath, L. S., Schneider, A., Grene, R., & Collakova, E. (2015). Transcriptome-wide functional characterization reveals novel relationships among differentially expressed transcripts in developing soybean embryos. *BMC Genomics, 16*(1), 928. doi:10.1186/s12864-015-2108-x

4.1 Introduction

I developed a classifier called CodeWise for accurate assessment of the coding potential of plant transcripts. CodeWise integrates several tools that aid in the categorization of coding versus noncoding transcripts.

CodeWise development pipeline is shown in Figure 4.1.

[1] Delasa Aghamirzaie developed CodeWise and performed all the data analysis. Dhruv Batra advised D.A. in development of CodeWise. Lenwood Heath advised in the data analysis sections. Andrew Schneider performed PCR for experimental validation. Ruth Grene and Eva Collakova performed all the biological data mining of the results.

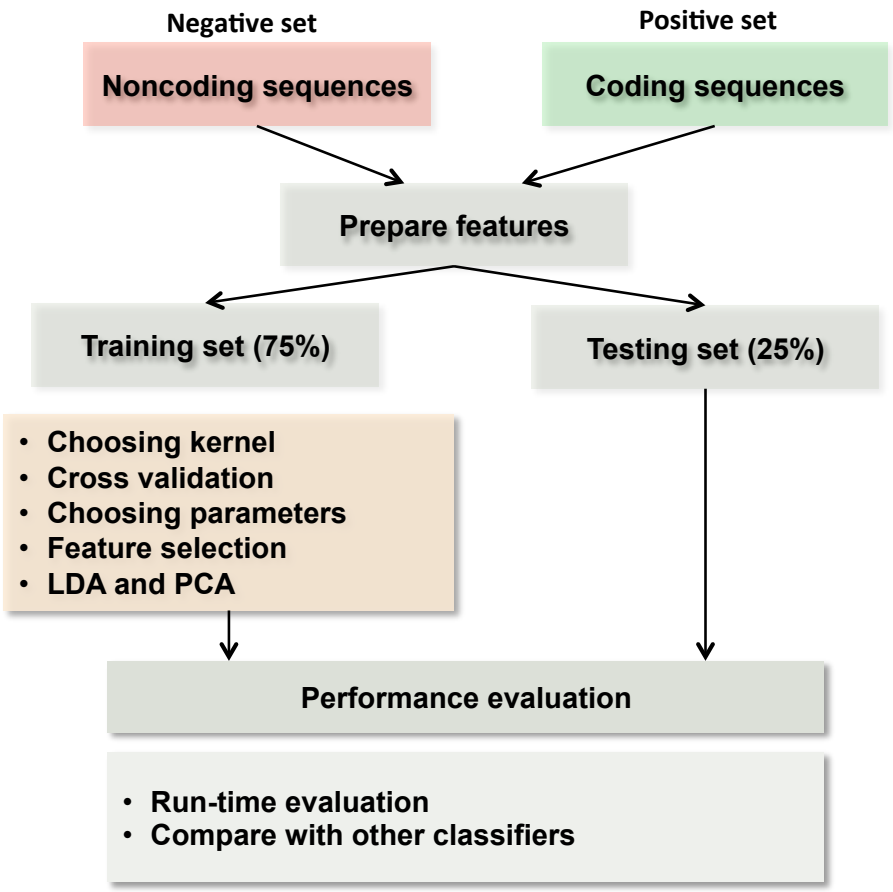

Figure 4.1 CodeWise development pipeline

Features. CodeWise features are shown in Figure 4.2 include: (i) sequence length, potential ORF ratio, UTR ratio, and potential protein length, (ii) sequence content (GC content, and T/A and G/C ratios), (iii) conserved domain information (number of conserved domains and extent of domain truncation), (iv) RNAfold-based minimum free energy of RNA secondary structure, (v) protein sequence similarity and functional annotation (presence of transcripts within the MapMan bins), and (vi) CPC scores. The Batch CD-Search tool was used as described above to identify conserved domains in each amino acid sequence. The extent of domain truncation is reflected in the "truncation ratio" defined as (the number of truncated domains)/ (total number of domains).

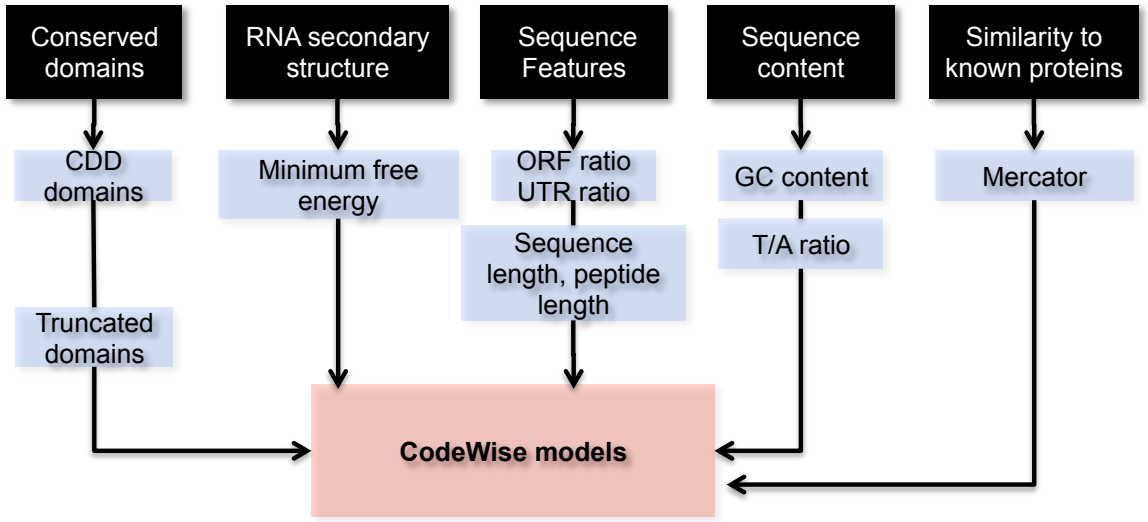

Figure 4.2 CodeWise features

Training and testing. The features described above were compiled for a total set of 115,000 unique *Arabidopsis thaliana* transcripts from The Arabidopsis Information Resource (TAIR) 10 database (Lamesch et al., 2012) and *Glycine max* (version 189) transcripts (Goodstein et al., 2012), including coding (positive class) and noncoding transcripts (negative class). The LibSVM package was used for implementation of the SVM classifier (Chang & Lin, 2011). The positive training set included 35,000 Arabidopsis transcripts and 50,000 soybean coding transcripts. The negative training set included non-redundant known Arabidopsis noncoding transcripts from 3 resources: (i) 25,000 from the plant long non-coding RNA database PLncDB (Jin, Liu, Wang, Wong, & Chua, 2013), (ii) 3,800 from the NONCODE version 4 database (Xie et al., 2014), and (iii) 278 from TAIR10 (Lamesch et al., 2012). There is currently no available source for noncoding soybean transcripts. Due to the low number of available noncoding transcripts, weighted SVM training (-wi weight in LibSVM) with a 3 to 1 ratio was used to prevent unbalanced training. 75% of the data were randomly selected for training, and the remainder of the data were used for testing, keeping the existing 3 to 1 ratio between the coding and noncoding transcript classes. The training and testing samples were normalized between -1 and +1 prior to training using the *svm-scale* program available in LibSVM. Several kernels (radial basis functional (RBF), polynomial, and linear) were used to select the best model. The linear kernel showed the best accuracy of 96%, followed by the RBF kernel (94%). Accuracy was determined as the ratio of correct

predictions to the total number of transcripts. SAS JMP Pro 11 was used for feature assessment in CodeWise using Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA).

4.2 CodeWise development

An important and challenging step for functional predictions is determining the coding potential of transcripts, which is a measure of how likely a transcript is to encode a protein. Noncoding transcripts can potentially interfere with, or otherwise affect, gene expression, which makes them candidates as important transcriptional and post-transcriptional regulators (Guttman et al., 2013; Ponjavic et al., 2007; Jeremy E. Wilusz et al., 2009; L. Yang et al., 2014). Integration of as many features as possible improves the accuracy of coding potential prediction tools. We gathered a large compendium of data related to sequence, RNA structure, conserved domains, sequence similarity, and functional annotation of Arabidopsis and soybean transcripts. We used binary SVM classification, which is a supervised learning approach known to yield high accuracy in high dimensional input data such as genomics data (Bhasin & Raghava, 2004; X. Zhang et al., 2006). A large spectrum of features was selected for evaluating the coding potential of each transcript in CodeWise to distinguish coding from ncRNAs: (i) sequence length, (ii) sequence content, (iii) presence and truncation of conserved domains, (iv) free energy of RNA secondary structure, (v) protein sequence similarity, and (vi) CPC score.

These features were selected based on the current state of knowledge about the characteristics of coding and noncoding transcripts. First, features related to the nucleotide and protein sequence length were shown to be necessary, but insufficient, for the separation of coding from noncoding transcripts (Amor et al., 2009; Z. J. Lu et al., 2011; Torarinsson, Sawera, Havgaard, Fredholm, & Gorodkin, 2006). These features include ORF and 5'-UTR ratios and potential protein lengths. Second, noncoding transcripts were shown to have higher GC content and T/A ratio than protein-coding transcripts (Amor et al., 2009; Crawford & Yanofsky, 2011). Third, protein-coding transcripts are likely to have conserved domains. Truncation of domains in either the C- or the N-terminus can affect protein function. To obtain information on the presence, absence, and/or truncation of functional domains, transcripts (including ncRNAs) were subjected to computational batch translation and batch CD-search. In the case of

ncRNAs, putative start and stop codons and potential peptides can still be identified computationally. Fourth, protein-coding transcripts have more stable secondary RNA structures than noncoding transcripts, which is reflected in their minimum free energy (Bánfai et al., 2012; Crawford & Yanofsky, 2011). This parameter was predicted by using RNAfold (Hofacker, 2003) for all coding and noncoding transcripts. Fifth, protein sequence similarity and functional annotation can be important to distinguish coding from noncoding transcripts. Mercator (Lohse et al., 2014) was used to assign transcripts into MapMan (Thimm et al., 2004) ontology bins based on protein sequence similarities. Sixth, CPC is used to assess the coding potential of transcripts (Kong et al., 2007). Incorporation of the CPC score as a feature in CodeWise was evaluated as well. These features were tested together and in different combinations to assess their importance for the overall accuracy of CodeWise.

4.3 CodeWise performance evaluation

CodeWise classified transcripts into coding and noncoding groups with the area under the receiver operating characteristic curve (AUC) > 0.98 on an independent test set, when all features were used for training (Figure 4.3). For assessing the coding potential of transcripts in CodeWise, no predetermined cutoff was used for distinguishing coding from noncoding transcripts with respect to protein length and sequence similarity. Instead, the classifier learned the cutoffs and patterns that exist between coding and noncoding classes among the training features. CodeWise assigned both coding and noncoding probabilities to each transcript. CodeWise outperformed CPC by a higher number of true predictions and a lower number of the false predictions (Figure 4.3B). Because the other available tools, specifically iSeeRNA, PhyloCSF, and CPAT do not include plants as a model system, evaluating their performance relative to CodeWise was irrelevant.

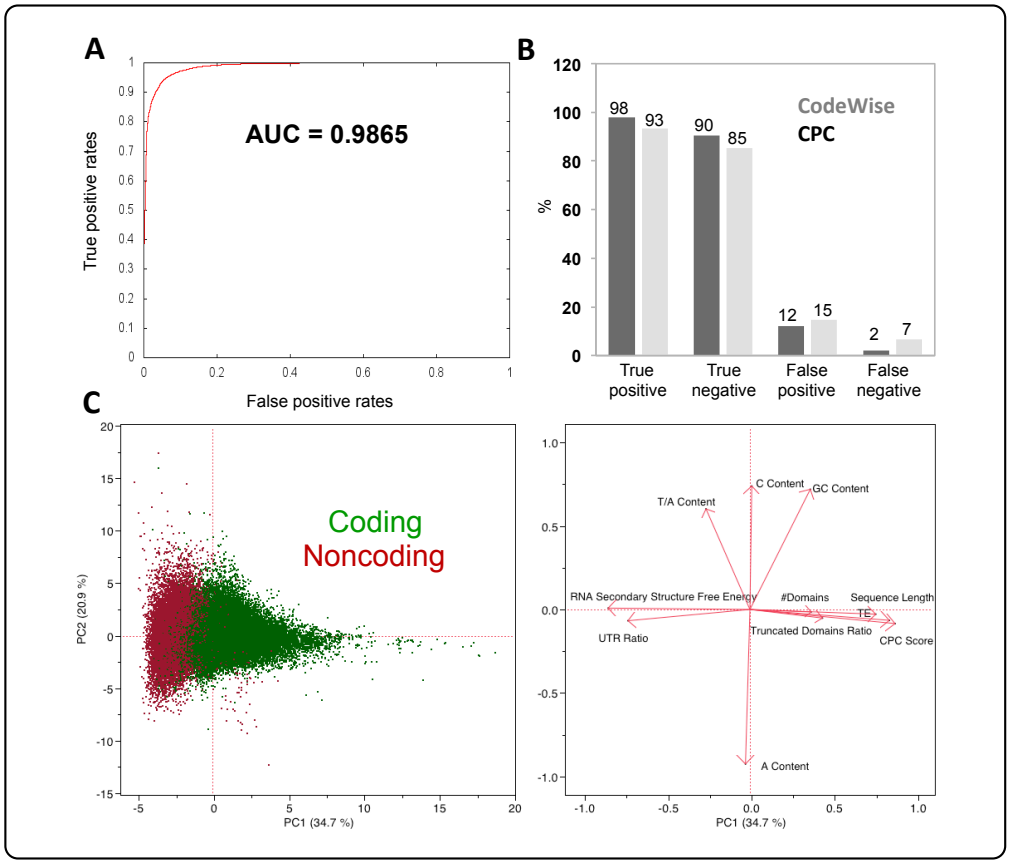

Figure 4.3. CodeWise performance evaluation. (A) ROC curve on the test set. (B) Comparison of CodeWise and CPC prediction power on the same set of known coding and noncoding transcripts. True positive: a coding transcript correctly predicted to be coding, true negative: a noncoding transcript correctly predicted to be noncoding, false positive: a noncoding transcript incorrectly classified as coding, false negative: a coding transcript incorrectly classified as noncoding (C) PCA results showing correlations of all features used to train the classifier. Score and loading plots for the first two principal components are shown on the left and right, respectively. Eigenvectors are shown as red arrows in the loading plot.

We used three methods to assess the contribution of different features within CodeWise. First, the CodeWise classifier was trained and tested with all combinations of six feature groups (127 combinations). Second, principal component analysis (PCA) was performed to evaluate how different features contributed to the variance between coding and noncoding classes (Figure 4.3C). PCA results revealed that sufficient separation (with only a small proportion of outliers) between the coding and noncoding transcripts was achieved solely through principal component 1 (PC1), which accounted for 34.3% of the variance. PCA also revealed the positive and negative correlations among the specific individual features (depicted as eigenvectors aligning in the same and opposite directions, respectively, along the PC1 axis in the loading plot of Figure 4.3C). Third, LDA was

used to find linear combination of features with the highest covariate scores for separation of coding and noncoding transcripts. LDA resulted in 93.94% correct classification of coding and noncoding transcripts with AUC of 0.9797 for both coding and noncoding classes. These three evaluation techniques revealed that at least three specific feature groups are required for separation of coding from noncoding transcripts: (i) the free energy associated with RNA secondary structure, (ii) the presence of conserved domains, and (iii) sequence features (5'-UTR ratio, potential ORF ratio, and protein length). Because CPC scores are highly correlated with ORF ratio and protein length, this feature does not affect CodeWise predictions (Figure 4.3B). No significant differences were observed in the nucleotide content, specifically between the GC content or the T/A ratio, of coding transcripts and noncoding transcripts (Figure 4.3B).

Noncoding transcripts had significantly higher minimum free energy of RNA secondary structure and tended to be shorter, with higher 5'-UTR ratio, lower potential ORF ratio, shorter predicted protein lengths, no conserved domains, and lower CPC scores than coding transcripts (Figure 4.4). The average minimum free energy of RNA secondary structure of coding and noncoding sequences was -371 and -134 cal mol^{-1}, respectively (Figure 4.4F).

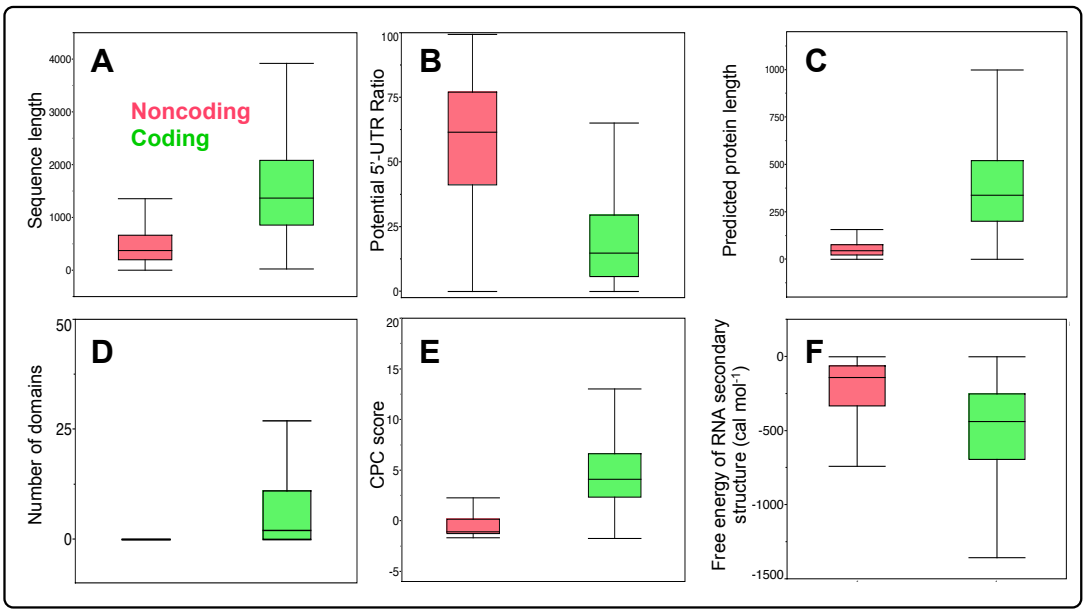

Figure 4.4 Box-and-whisker plot comparisons of coding and noncoding transcripts predicted by CodeWise with respect to individual features. (A) Sequence length. (B) Potential 5'-UTR ratio. (C) Potential protein length. (D) Number of conserved domains. (E) CPC score. (F) Minimum free energy of RNA secondary structure.

5 Transcriptome-wide functional characterization reveals novel relationships among differentially expressed transcripts in developing soybean embryos[1]

This chapter is produced from (Aghamirzaie et al., 2015), which is an open access journal.

Aghamirzaie, D., Batra, D., Heath, L. S., Schneider, A., Grene, R., & Collakova, E. (2015). Transcriptome-wide functional characterization reveals novel relationships among differentially expressed transcripts in developing soybean embryos. *BMC Genomics, 16*(1), 928. doi:10.1186/s12864-015-2108-x

Abstract

Background. Transcriptomics reveals the existence of transcripts of different coding potential and strand orientation. Alternative splicing (AS) can yield proteins with altered number and types of functional domains, suggesting the global occurrence of transcriptional and post-transcriptional events. Many biological processes, including seed maturation and desiccation, are regulated post-transcriptionally (e.g., by AS), leading to the production of more than one coding or noncoding sense transcript from a single locus.

Results. We present an integrated computational framework to predict isoform-specific functions of plant transcripts. This framework includes a novel plant-specific weighted support vector machine classifier called CodeWise, which predicts the coding potential of transcripts with over 96% accuracy, and several other tools enabling global sequence similarity, functional domain, and co-expression network analyses. First, this framework was applied to all detected transcripts (103,106), out of which 13% was predicted by CodeWise to be noncoding RNAs in developing soybean embryos. Second, to investigate the role of AS during soybean embryo development, a population of 2,938 alternatively spliced and differentially expressed splice variants was analyzed and mined

[1] Delasa Aghamirzaie performed all the computational data analysis. Dhruv Batra aided in development of CodeWise. Lenwood Heath advised in data analysis sections. Andrew Schneider performed PCR for experimental validation. Ruth Grene and Eva Collakova performed all the biological data mining of the results.

with respect to timing of expression. Conserved domain analyses revealed that AS resulted in global changes in the number, types, and extent of truncation of functional domains in protein variants. Isoform-specific co-expression network analysis using ArrayMining and clustering analyses revealed specific sub-networks and potential interactions among the components of selected signaling pathways related to seed maturation and the acquisition of desiccation tolerance. These signaling pathways involved abscisic acid- and FUSCA3-related transcripts, several of which were classified as noncoding and/or antisense transcripts and were co-expressed with corresponding coding transcripts. Noncoding and antisense transcripts likely play important regulatory roles in seed maturation- and desiccation-related signaling in soybean.

Conclusions. This work demonstrates how our integrated framework can be implemented to make experimentally testable predictions regarding the coding potential, co-expression, co-regulation, and function of transcripts and proteins related to a biological process of interest.

5.1 Background

Seed maturation and induction of dormancy represent essential stages in soybean seed development that are triggered through highly coordinated signaling and metabolic pathways within the seed maturation and desiccation programs. LEAFY COTYLEDON (LEC) 1 and transcription factors (TFs) containing the B3 DNA-binding domain, namely LEC2, ABSCISIC ACID INSENSITIVE (ABI) 3, and FUSCA (FUS) 3, are key regulators of seed filling, commonly called “the B3 regulatory network” (R. Finkelstein, 2013; Santos-Mendoza et al., 2008). Their mutual interactions and interactions with their targets and components of phytohormone-mediated signaling connect these TFs within the well-studied B3 regulatory network to developmental and metabolic processes leading to the synthesis and accumulation of seed storage compounds. During late seed filling, maturing seeds acquire desiccation tolerance (DT) and dormancy, as the water content decreases, primarily through abscisic acid (ABA)-mediated signaling (Angelovici et al., 2010; R. Finkelstein, 2013; R. Finkelstein et al., 2008; Gutierrez et al., 2007). Seed filling-, desiccation-, and dormancy-related processes are regulated at both the transcriptional and post-transcriptional levels. To gain a basic understanding of these regulatory processes, it is important to identify additional regulatory molecules, e.g.,

proteins and RNA, involved in the seed maturation developmental program, which can be achieved through transcriptomics in conjunction with bioinformatics analyses.

High-throughput RNA sequencing (RNA-Seq) reveals high transcriptional activity in unannotated and annotated regions of genomes in various organisms, resulting in the discovery of many previously unknown transcripts (Ponjavic et al., 2007; L. Yang et al., 2014). Alternative splicing (AS) is a major source of this transcript diversity, as a single gene can encode multiple transcripts. These transcripts can be coding or noncoding, genic or intergenic, and sense or antisense. Coding transcripts are translated into proteins or regulatory peptides that can contain, or lack known domains important for function, regulation, interaction with other molecules, and subcellular localization (Andrews & Rothnagel, 2014; Lindsey et al., 2002). In contrast, noncoding transcripts, including long noncoding RNAs (lncRNA) and long intergenic noncoding RNAs (lincRNA) can act directly as regulators (Guttman et al., 2013; Ponjavic et al., 2007; Jeremy E. Wilusz et al., 2009; L. Yang et al., 2014). These noncoding RNAs (ncRNAs) perform their regulatory functions through transcriptional interference, sense and antisense hybridization, interactions with RNA-binding proteins, and/or serving as precursors for small regulatory RNAs (Jeremy E. Wilusz et al., 2009; L. Yang et al., 2014). Noncoding transcripts have been reported to be involved in the regulation of development and in responses to stress in plants (Amor et al., 2009; Boerner & McGinnis, 2012; W. Zhang et al., 2014). To date, specific plant lncRNAs have been implicated in the regulation of flowering, response to cold, root meristem development, and modulation of AS (Bardou et al., 2014).

High-throughput experimental testing to predict the functions of newly identified transcripts is not possible. As a first step towards future experimentation, the functions of coding and noncoding transcripts can be inferred computationally by integrating several approaches, such as functional annotation based on sequence similarity, global functional domain analyses, determining the coding potential of transcripts, co-expression analyses, and the construction of hypothetical regulatory networks. Because transcripts can be coding or noncoding, determining their coding potential is a necessary step towards their functional characterization. Identification of conserved domains within newly identified coding sequences using well-established tools such as InterPro (Apweiler et al., 2001)

and Batch Conserved Domain (CD) Search (Marchler-Bauer et al., 2011) is important for *in silico* function prediction.

Several tools have been developed for predicting the coding potential of individual transcripts primarily in animal systems, including Coding Potential Calculator (CPC) (Kong et al., 2007), Coding Potential Assessment Tool (CPAT) (Jin et al., 2013), PhyloCSF (Lin et al., 2011), and iSeeRNA (Sun et al., 2013). These tools rely upon sequence similarity and open reading frame (ORF) length to distinguish between coding and noncoding transcripts (Chew et al., 2013; Guttman et al., 2013). However sequence similarity and ORF length alone lack sufficient power to accurately distinguish between coding and noncoding RNAs (ncRNA). Additional features, such as the presence of conserved functional domains, GC content, and the free energy of RNA secondary structure, are needed to improve the detection accuracy of ncRNAs (Amor et al., 2009; Z. J. Lu et al., 2011; Torarinsson et al., 2006). To our knowledge, there are currently no comparable tools available to globally characterize coding and ncRNAs specifically in plants.

Here we present the development and implementation of a transcriptome-wide computational framework that combines high-throughput information with bioinformatics tools to predict potential functions and novel associations among transcripts and inferred proteins. Co-expression-related guilt-by-associations, timing of expression, sequence similarity, presence of functional domains in protein variants, and coding potential of transcripts were each used to infer possible function. The framework includes (i) a pipeline for global analysis of functional domains in proteins, and (ii) CodeWise, an accurate support vector machine (SVM) classifier that uses several features to predict the coding potential of transcripts. This framework was applied to an existing data set related to seed filling and early desiccation stages in developing soybean embryos (Aghamirzaie et al., 2013; E. Collakova et al., 2013). We mined this data set extensively in the context of AS events and (i) the coding potential of transcripts, (ii) the presence or absence of functional domains, (iii) similarity to Arabidopsis proteins, and (iv) timing and patterns of expression, including co-expression network analysis, during soybean embryo development. Highly connected nodes within the co-expression network (hubs) connecting the majority of transcripts expressed during the desiccation phase were

identified. Hypothetical ABA- and FUS3-related signaling pathways focusing specifically on signaling components subjected to AS and related to soybean seed filling and acquisition of DT are also presented and discussed.

5.2 Methods

5.2.1 Definition of terms

Common terms used in this study are defined in Table 5.1.

Table 5.1 Glossary of common terms that were used in this study

Term	Definition
degree of connectivity	Identification of how well a node is connected in a network. For example, if a network has 10 nodes and a node is connected to 5 nodes, it's degree of connectivty is 0.5.
desiccation tolerance (DT) phase	The last phase in developing embryos characterized by loss of water, a sharp increase of desiccation-related metabolites and transcripts, and acquisition of DT in yellow embryos at day 55.
differentially expressed transcript	A transcript that was significantly differentially expressed at at least one time point compared with previous time point (FDR < 0.05).
domain categorization	The domain composition of SV-pairs were compared and categorized into similar domains, no known domain, and disparate domains.
early maturation phase	This first phase in seed filling is characterized by an initial decrease in metabolites and cell-division-related transcript levels and the onset of accumulation of seed storage compounds.
expressed transcripts	A transcript was defined as expressed if the sum of its FPKM values across the time course was greater than one.
hub	Highly connected nodes in a network (nodes with the highest degree of connectivity)
mid-to-late maturation phase	This phase in embryo development is characterized by a stable accumulation of seed storage compounds.
nearest neighbors	Nodes directly connected through individual edges to a single node of interest.
regulon	A group of transcripts known to be targets of a common TF. For example, a group of transcripts known to be targets of ABI3, is called the ABI3 regulon.
soybean developmental stages	Three major developmental stages (early maturation, mid-to-late maturation, and DT) defined in this report on the basis of changes in the levels of relevant metabolites, seed storage compounds, and transcripts in developing soybean embryos.
splice variant (SV)	Transcripts that are products of the same precursor mRNA.
sub-network	A group of nodes and edges that are part of a larger network. For example, a node with all its nearest neighbors comprising the members of the FUS3 regulon is a sub-

	network.
super-cluster	Clusters of transcripts with similar expression profiles grouped according to predefined soybean developmental stages.
SV group	A group containing at least two SVs of a gene both of which were significantly differentially expressed during embryo development. For example, gene X has three SVs (X.1, X.2, X.3), and X.1 and X.3 showed changes in transcript levels and X.2 was stably expressed in developing embryos. X.1 and X.3 belong to the same SV group representing gene X.

5.2.2 Analysis of RNA-Seq data and identification of differentially expressed transcripts

Our RNA-Seq data set (GEO accession number GSE46153) includes ten time points with three biological replicates per time point, representing the phases of soybean embryo development from the onset of seed filling to the onset of seed desiccation. Read mapping, transcriptome assembly, and differential expression analyses were done using Tophat2, Cufflinks, and Cuffdiff2 available in the Tuxedo Suite (C. Trapnell, Hendrickson, et al., 2012) RNA-Seq pipeline (Aghamirzaie et al., 2013; E. Collakova et al., 2013). The *Glycine max* reference genome (version 189) was used to guide transcriptome assembly, which yielded 39,191 known and 64,005 novel expressed transcripts. A transcript was defined as expressed if the sum of its FPKM (fragments per kilobase of exon per million fragments mapped) values across the time course was greater than 1. Based on the Cuffdiff2 results and temporal differential expression analysis at the isoform level, 17,181 transcripts were significantly differentially expressed during at least one time point when compared to the previous time point (false discovery rate (FDR) < 0.05). Based on further categorization, 2,938 out of 17,181 transcripts were also alternatively spliced and originated from 1,393 genes, meaning that for each of these genes, at least two differentially expressed splice variants (SVs) were identified. Nucleotide sequences of newly assembled transcripts were extracted and assembled using an in-house Python program to parse the transcriptome reference output by Cuffmerge (merged.gtf) from the soybean genome. Class codes used are a set of 12 Cuffcompare transcript codes proposed by (C. Trapnell et al., 2010). Novel transcripts appeared as novel SVs of known genes (transcript classes “j”, “o”, and “c”), as well as in intergenic

(transcript classes "-" and "u") and antisense (transcript classes "x" and "s") classes. The term "transcript" is used as a general term and includes all types of detected transcripts as opposed to SVs that are defined as transcripts produced from the same premature messenger RNA (pre-mRNA). The nomenclature for novel SVs was adapted from (Aghamirzaie et al., 2013). For example, if a gene had two known SVs, two novel SVs were designated N3 and N4.

5.2.3 Transcriptome-wide computational framework

We devised an extensive framework to obtain isoform-specific information for all expressed transcripts using Batch CD-Search (Marchler-Bauer et al., 2011), Mercator (Lohse et al., 2014), RNAfold (Hofacker, 2003), CPC (Kong et al., 2007), and CodeWise. The results obtained from the application of each tool were mined separately and also in conjunction with the other tools to enable functional inference for selected known and novel transcripts. All parameters in the publicly available tools were set to their default values unless otherwise stated. In the following sections, implementation details of each tool are described.

Batch translation and Batch-CD Search. First, the in-house Python program BatchTranslator.py was used to find the longest protein sequence in each nucleotide sequence in batch mode. This program evaluates all ORFs of a sequence, starting with the AUG start codon and ending with any of the three stop codons, returning only the longest protein sequence. The program produces two separate output files: (i) FASTA protein sequence and (ii) information on translation statistics including the length of the 5' untranslated region (5'-UTR), potential ORF length, potential ORF ratio (length of potential ORF)/(length of transcript), 5'-UTR ratio (length of 5'-UTR)/(length of transcript), and protein length. The in-house program accepts a FASTA file and returns the most likely ORF of a transcript with accuracies of 99% in *Arabidopsis thaliana*, 96% in *Medicago truncatula*, and 95% in *Glycine max,* (all data were obtained from Phytozome v9). Second, Batch CD-Search, which accepts up to 100,000 protein sequences at a time, was used to identify conserved domains (Marchler-Bauer et al., 2011).

Mercator (Lohse et al., 2014) is a sequence similarity-based functional annotation tool that uses the Basic Local Alignment Search Tool (BLAST) algorithm to identify sequences that resemble the query sequence (above a specified threshold) from several reference sequence databases. Collectively, these databases contain information on all Arabidopsis proteins, proteins from the SwissProt Plant Protein Annotation Program (6,000 plant proteins), 57,000 rice proteins, 17,000 *Chlamydomonas reinhardtii* protein models, and 2,169 domains from InterPro, conserved domain database (CDD), and Eukaryotic Orthologous Groups (KOG) databases. Mercator assigns each transcript to a MapMan ontology bin. The presence of a given transcript in a known MapMan bin helps to predict functionality of that transcript. All parameters were utilized in the Mercator web server and the BLAST cutoff parameter was set to 50.

RNAfold, available in the Vienna package (Hofacker, 2003), was used to predict RNA secondary structure and the minimum free energy of all transcripts by using the command-line version of RNAfold in batch mode. The **CPC** web server (Kong et al., 2007) was used to assess the coding potential of transcripts. CPC uses an SVM classifier trained with respect to sequence similarity (using BLAST) and length (using FrameFinder). Coding potential is predicted with reference to known protein sequences in the UniProt database (UniProt, 2015).

5.2.4 Clustering and correlation analyses

GeneCluster 3.0 (de Hoon, Imoto, Nolan, & Miyano, 2004) was used for centering, normalizing, and clustering of SVs into 5, 10, 15, 25, and 30 clusters based on their FPKM values with 500 iterations using the k-means algorithm. Distinct expression patterns within the transcript population were detected in 25 clusters. Therefore, k=25 was selected for further visualization in Java Tree View (Saldanha, 2004) and for data mining. Pearson correlation analysis of sense and antisense transcripts was calculated using an in-house Python program. Sense and antisense transcripts that showed significant correlation of expression over the time course of soybean embryo development were identified (p-value < 0.05).

5.2.5 Co-expression network analysis

ArrayMining was used to construct a co-expression network for the set of 2,938 differentially expressed and alternatively spliced transcripts. ArrayMining yields a weighted gene co-expression network of significantly correlated genes that have similar expression patterns within a user-defined threshold (Glaab, Garibaldi, & Krasnogor, 2009). The Fruchterman-Reingold method was used for network visualization, the edge-adjacency threshold was set to 0.9, and the resulting network was visualized using an organic layout in Cytoscape 3.1 (Shannon et al., 2003). We define two nodes as nearest neighbors in a network if there is a direct edge connecting those two nodes. If a node (*x)* is connected to *m* nodes and *n* is the total number of nodes in a super-cluster *sc*, the degree of connectivity of *x* in *sc* can be defined as:

$$Degree\ of\ connectivity_{x,sc} = {}^{m}/_{n}$$

The degree of connectivity for super-cluster *sc* did not follow a normal distribution and median was chosen to represent this distribution. The degree of connectivity for a super-cluster *sc* was defined as:

$$Degree\ of\ connectivity_{sc} = median\ (Degree\ of\ connectivity_{all\ nodes,sc})$$

5.2.6 Signaling Pathway Visualization

The in-house tool Beacon was used for the visualization of signaling pathways (Kakumanu et al., 2012). The Beacon Pathway Editor consists of a tool designed to draw pathways encoded in the Systems Biology Graphical Notation Activity Flow language that is a standard for describing pathways in terms of perturbations, influences, activities, logical operators, and phenotypes (Le Novere et al., 2009).

5.2.7 Quantitative real-time PCR

Quantitative Polymerase Chain Reaction (qPCR) was performed on selected sense and antisense transcripts, including LEC1-Like (L1L), two ETHYLENE RESPONSE FACTOR/APETALA 2 (ERF/AP2) TFs, gibberellin 2 (GA2) oxidase, and phytochrome-

interacting basic helix-loop-helix 5 (PIL5), using samples from several time points to further validate the changes in transcript levels obtained from RNA-Seq as described (Aghamirzaie et al., 2013).

5.3 Results

5.3.1 Overview of the transcriptome-wide computational framework

The data used in this study were taken from an existing transcriptomics data set pertaining to seed filling and early desiccation stages of soybean embryo development (E. Collakova et al., 2013). Differential expression analysis of this data set yielded 17,181 transcripts (many of which were previously unidentified) that showed significant changes in their levels over time (FDR < 0.05) (Aghamirzaie et al., 2013). Some of the newly identified transcripts were novel SVs, intergenic, and/or antisense and of different coding potentials. These types of transcripts, although important in regulating various aspects of cell development (Amor et al., 2009; Andrews & Rothnagel, 2014; Guttman et al., 2013; Lindsey et al., 2002), have been largely neglected in analyses of transcriptomics studies to date.

Our framework (Figure 1) involves new and existing tools that were applied (i) globally to all identified known and novel transcripts and (ii) to a set of 2,938 transcripts originating from 1,393 genes. Each of these genes was defined as having at least two significantly differentially expressed SVs in developing soybean embryos. These 2,938 transcripts do not include transcripts that showed stable, non-changing expression levels. While the entire analysis was performed at the transcriptome-wide level, detailed mining of splicing events and function predictions was only performed on the smaller data set of 2,938 transcripts.

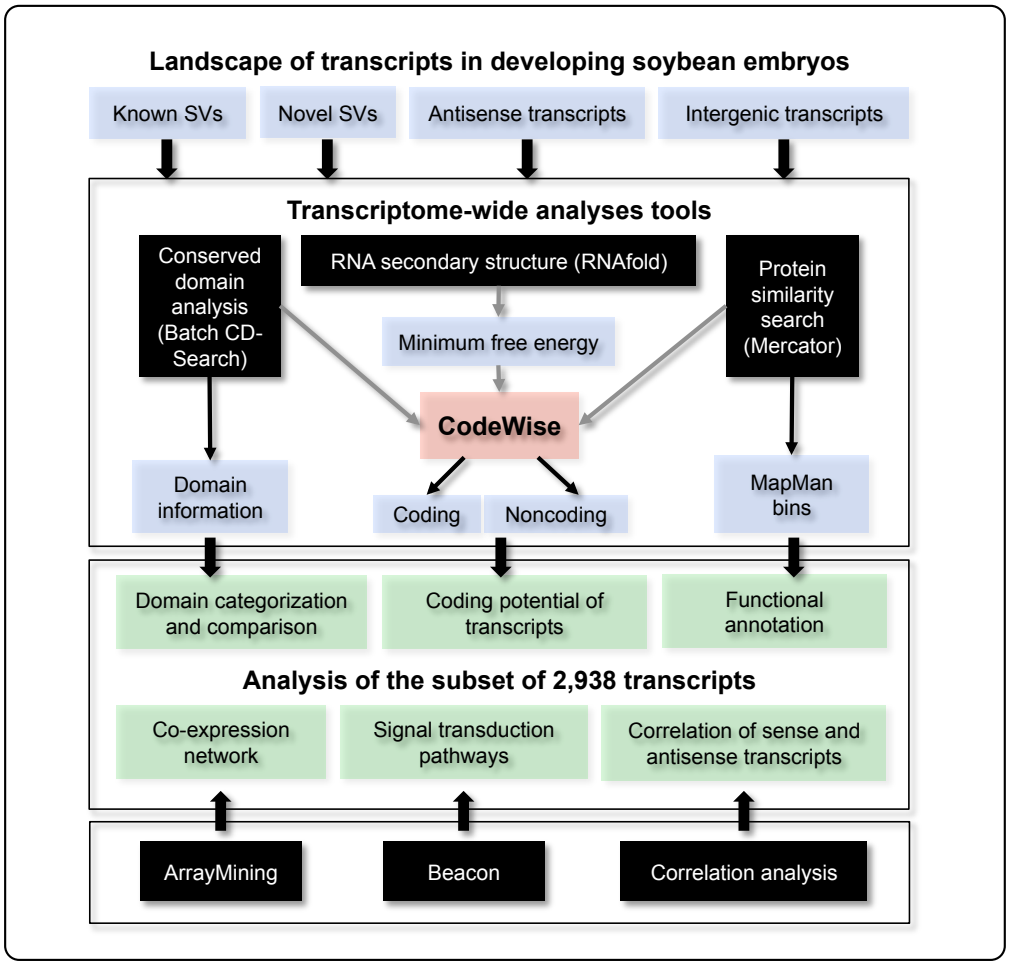

Figure 5.1 Computational framework. Transcriptome-wide analysis tools comprised large-scale conserved domain search using Batch CD-Search, RNA secondary structure prediction using RNAfold, functional annotation by Mercator, and coding potential assessment using CodeWise. These tools, in conjunction with co-expression network analysis by ArrayMining and signal transduction pathway analysis in Beacon, were used for detailed data analyses of a set of 2,938 transcripts. This population afforded the opportunity to identify candidate SVs transcribed from the same gene with potentially distinct functions in different stages of soybean embryo development. Each of the 1,395 genes had more than one transcript significantly differentially expressed during time-course of soybean embryo development (FDR < 0.05), leading to the detection of 2,938 transcripts. Tools are shown in black, inputs in blue, and outputs in green.

The first steps in the analysis included (i) large-scale functional domain analysis by batch CD-Search (Marchler-Bauer et al., 2011), (ii) predictions of RNA secondary structures and their minimum free energy by RNAfold (Hofacker, 2003), and (iii) functional predictions and annotations by Mercator (Lohse et al., 2014). These tools were used independently and also in conjunction with the in-house SVM classifier CodeWise to assess the coding potential of transcripts. Second, additional tools were applied to the set of 2,938 transcripts, including (i) co-expression network analyses by ArrayMining (Glaab et al., 2009) and visualization in Cytoscape (Shannon et al., 2003), (ii) the

depiction of inferred signal transduction pathways in the Beacon Pathway Editor (Kakumanu et al., 2012) based on prior knowledge combined with our data, and (iii) Pearson correlation analysis of sense and antisense transcripts by an in-house python program, all in the context of AS, and timing and patterns of transcript expression. In the following sections, different modules of this framework are explained in detail.

5.3.2 Transcriptome-wide domain analysis of protein variants

Detection of the presence or absence of functional domains can aid in predicting interactions and therefore, functions of protein isoforms. For example, a novel protein isoform possessing a new domain known to facilitate interactions with signaling proteins of known function can be inferred to potentially interact with these other proteins and function in those signal transduction pathways. Hence, there was a need to obtain global information concerning the presence, absence, and/or truncation of functional domains. All known and novel expressed transcripts were translated *in silico* to identify the longest amino acid sequence and then subjected to Batch CD-Search (Marchler-Bauer et al., 2011) to identify conserved domains in the set of protein sequences. This transcriptome-wide domain analysis led to the extraction of specific domain information for all expressed transcripts.

5.3.3 Transcriptome-wide analysis of coding and noncoding transcripts in developing soybean embryos using CodeWise

Transcriptomics analysis revealed the expression of 39,101 known and 64,005 novel transcripts in developing soybean embryos (Table 5.2). The transcript population included transcripts from genic and intergenic regions. Based on the Cuffdiff2 analysis, 17,181 out of 103,106 transcripts showed significant differential expression. On average, a soybean gene produced three transcripts and, for the most part, the previously known SVs showed significantly higher expression than the novel transcripts ($p < 0.0001$, t-test). Known transcripts had significantly higher average FPKM values than antisense, intergenic, novel genic sense, and overlapped transcripts ($p < 0.0001$, t-test). Using the CodeWise classifier, we identified 13,652 lncRNAs, including 10,023 genic lncRNA, 1,064 lincRNAs, and 2,295 noncoding antisense transcripts. Based on CodeWise test results on coding and known noncoding plant transcripts compiled from existing

databases (Goodstein et al., 2012; Jin et al., 2013; Lamesch et al., 2012; Xie et al., 2014), we estimated that about 96% (AUC > 0.98) of these lncRNAs were correctly rejected as coding (true negatives). Long noncoding transcripts showed significantly lower expression than coding transcripts ($p < 0.0001$, t-test), which is consistent with other studies (Derrien et al., 2012; Kung, Colognori, & Lee, 2013; Z. J. Lu et al., 2011).

Table 5.2 Transcript distribution among different classes of significantly changed transcripts. Cuffdiff2 was used for time-course differential expression analysis at isoform level. Transcripts that were differentially expressed at least at one time point compared with previous time point with FDR < 0.05 were defined as significantly differentially expressed. Out of 103,106 transcripts detected in developing soybean embryos, 17,181 transcripts were significantly changed.

Transcript classes	Transcript number	Significantly changed
Known (=)	39,101	13,398
Novel splice junction (j)	57,376	2,840
Overlapped (o)	1,689	252
Antisense exon (x)	2,266	242
Antisense intron (s)	599	30
Intergenic	2,075	419
Total	**103,106**	**17,181**

5.4 Bioinformatics analyses of AS events

5.4.1 Identification of alternatively spliced and significantly differentially expressed transcripts

We previously identified 1,393 genes, each with more than one significantly differentially expressed transcript, resulting in a population of 2,938 known and novel SVs and antisense transcripts (Aghamirzaie et al., 2013). This population afforded the opportunity to identify candidate SVs transcribed from the same gene with potentially distinct functions in different stages of soybean embryo development. Transcriptome-wide analysis of this relatively small subset revealed several interesting phenomena in the

context of embryo development. For example, this data set includes (i) SVs with different splicing patterns covering major developmental stages of developing soybean embryos, (ii) coding and noncoding transcripts such as lncRNAs and antisense transcripts, (iii) SVs with different number and types of conserved domains with the same and/or different expression profiles.

To illustrate the use of our framework for biological data mining and function inference, this set of transcripts was further analyzed in several ways. The k-means clustering algorithm was used to group these 2,938 transcripts into 25 clusters that represented major trends in seed filling and early desiccation-related processes. Some of the clusters displayed similar trends and were therefore merged into six super-clusters (Figure 5.2A), based on prior knowledge obtained from the same data set concerning the timing of metabolite and metabolism-related transcript accumulation (E. Collakova et al., 2013). Three basic trends reflecting changes in metabolite and transcript levels (Figure 5.2B) included: (i) early maturation - initial decrease until day 15 – 20, followed by stable low levels (green trend), (ii) mid-to-late maturation - initial increase, followed by stable high levels (blue trend), and (iii) desiccation (DT) - appearance of metabolites and transcripts in yellow embryos at day 55 (red trend). Overall, the transcripts were not evenly distributed among the six super-clusters. The majority of AS events were observed in the DT super-cluster, followed by the early and mid-to-late super-clusters. Known and novel splice junction SVs dominated all super-clusters, but a small number of transcripts belonging to other classes (exon skipping and antisense) were also observed in nearly all super-clusters.

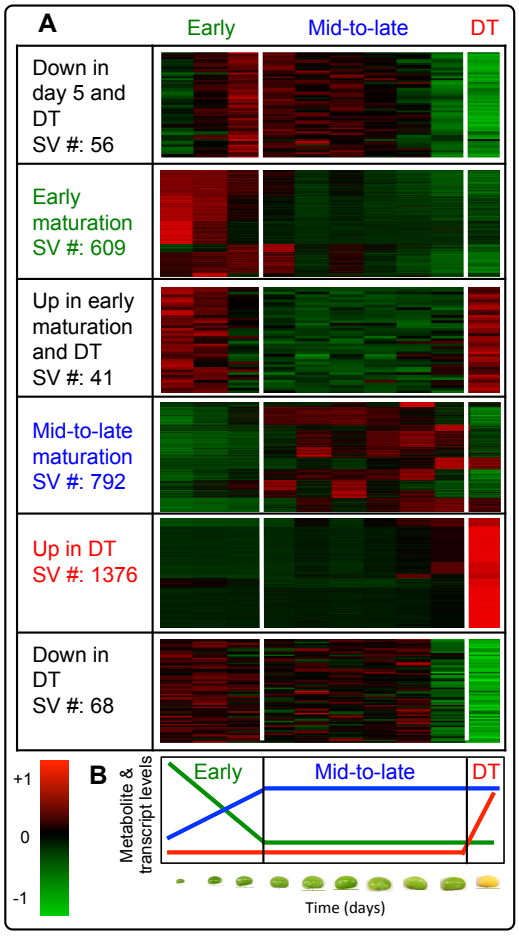

Figure 5.2 Assembly of super-clusters (A) Normalized FPKM values of the set of 2,938 transcripts were clustered into 25 groups using the k-means algorithm. Clusters with similar expression profiles across developmental stages were grouped into six super-clusters. Three major developmental stages (early maturation (green), mid-to-late maturation (blue), and desiccation (red)) containing 94.5% of the transcript population were identified. (B) Trends involving changes in metabolites and transcript levels [28] were grouped by developmental stages and color coded as corresponding three major super-clusters.

The three temporal patterns provided the basis for further mining of the population of 2,938 SVs, which included (i) expression patterns, (ii) presence or absence of conserved domains, (iii) functional annotation based on protein similarity by Mercator, (iv) CPC- and CodeWise-derived coding potential predictions, (v) potential ORF ratio and 5'-UTR length, and (vi) GC content. Classification of 2,605 transcripts was consistent between CodeWise and CPC, while 333 transcripts were reclassified by CodeWise. Based on the testing results presented above, 96% of these predictions are estimated to be correct.

5.4.2 Conserved domain analysis of potential protein variants

We define an "SV group" as those isoforms in the population of 2,938 transcripts that were spliced from the same pre-mRNA. Members of each SV group were divided into three categories in terms of differences in their conserved domains. This categorization was done by performing pairwise domain comparisons of isoforms within each SV group on Batch-CD Search results (Figure 5.3), with the focus on (i) disparate domains, defined as SVs differing in at least one conserved domain, (ii) similar domains, defined as SVs having the same types of domains, but the number of domains can be different, and (iii) no known domains, defined as at least one of the SVs lacked any conserved domains. This domain categorization allowed the exploration of differences among SVs with respect both to their functional domains and timing of expression, which can facilitate prediction of possible functional roles of different SV pairs. SVs having different expression profiles (reflected in their presence in different super-clusters) with different number and types of conserved domains may play distinct roles in developing embryos. Domain comparisons among SVs present within the same super-cluster were also performed to obtain information about SVs that had similar expression profiles.

Interestingly, the majority of SVs (80%) originating from the same gene co-expressed and belonged to the same super-cluster. Others were expressed at different times (different super-clusters, e.g., "DT, early maturation"). The majority of SV groups either had similar domains (48%) or one of the SVs in the group lacked any known domain, regardless of super-cluster comparisons (37%) (Figure 5.3). The group of SVs with no domains included both sense lncRNAs (12%) and transcripts encoding peptides or proteins with no known domains (14%). For example, 777 SV pairs were up-regulated at the DT stage (both SVs belonged to the DT super-cluster). Among these SV pairs, the majority had similar domains (399 SV pairs), while only 72 SV pairs had disparate domains. The remainder of the SV pairs contained one partner SV with no known domain. Eighty SV pairs belonged to the mid-to-late and early maturation super-clusters. While these SV pairs had completely different expression profiles, 38 and 42 SVs had similar and different, respectively, conserved domains.

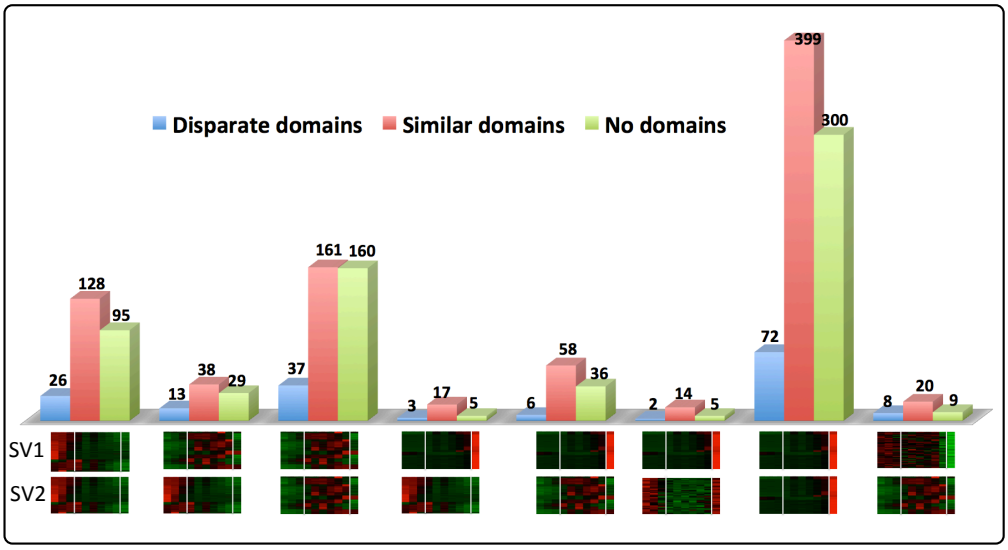

Figure 5.3 Pairwise comparisons of transcripts with respect to their domains and expression patterns across the population of 2,938 transcripts. Batch CD-Search results were mined to identify transcripts encoding proteins and peptides with disparate, similar, or no known domains. Each gene in this set had at least two transcripts that were differentially expressed. Thus, transcripts originating from a gene may belong to the same (hence same super-cluster comparisons) or different super-clusters (hence different super-cluster comparisons). SV pairs belonging to the same or different super-cluster pairs are shown below for each comparison.

5.4.3 Sense and antisense transcript pair analysis

Emerging studies provide evidence that natural antisense transcripts play an important role in regulating gene expression (Jabnoune et al., 2013; H. Wang et al., 2014). RNA-Seq analysis enabled the identification of 167 novel sense and antisense transcript pairs that showed changes in expression during soybean embryo development. A plausible hypothesis is that if a corresponding sense and antisense transcript pair shows positively or negatively correlating expression patterns, then the stability of the sense transcript will be affected by the antisense transcript. For sense and antisense transcript pairs, potential correlations were investigated using Pearson correlation analysis. The majority of sense-antisense pairs (155 out of 167 pairs) had significantly correlated expression profiles during soybean embryo development. Specific examples of potential antisense regulation will be discussed in detail in section 5.4.8 in relation to ABA and/or FUS3 action and timing of their expression.

5.4.4 AS events related to ABA and/or FUS3 action

FUS3 plays a key role in the regulation of seed development [51], as does the phytohormone ABA [1]. It was therefore of great interest to understand the relationship of RNA splicing and antisense regulation to ABA- and FUS3-related events in developing soybean embryos and to search for possible clues to as yet unknown and/or partially understood regulatory mechanisms. Therefore, we mined the set of the 2,938 transcripts for potential ABA- and FUS3-related targets. ABA-related Arabidopsis genes were extracted from (R. Finkelstein, 2013) and included proteins involved in ABA metabolism and signaling, as well as those associated with interactions of ABA with other hormone-mediated pathways. Similarly, the identity of the genes in the FUS3 regulon in Arabidopsis was obtained from (F. Wang & Perry, 2013). The Arabidopsis gene IDs associated with the corresponding soybean genes encoding these 2,938 differentially expressed transcripts were cross-referenced with the Arabidopsis ABA-related and FUS3-regulated genes to obtain ABA-related and FUS3-regulated potential homologs in soybean. This mining led to the detection of 318 transcripts. These transcripts were carefully examined with respect to conserved domains, coding potential, and functional annotation. The majority of ABA-related transcripts were expressed during the mid-to-late and DT phases of soybean embryo development (89%). FUS3 is encoded by two genes in soybean, each producing one transcript (Glyma16g05480.2 and Glyma19g27336.1) in developing embryos, both genes showing similar and stable expression until day 55 when their levels dropped significantly (E. Collakova et al., 2013). The FUS3 regulon (F. Wang & Perry, 2013) contained 181 transcripts, some of which are also related to ABA signaling, showed differential expression during soybean embryo development.

5.4.5 AS events related to ABA and/or FUS3 action during early maturation

The early super-cluster is relevant to young, fully differentiated, green embryos that expressed genes associated with various aspects of cell division but already had started to accumulate seed storage compounds (E. Collakova et al., 2013). The FUS3-related SVs belonging to the early super-cluster included: (i) L1L, (ii) receptor protein kinase barely any meristem (BAM) 2 and calcium-dependent protein kinase (CPK) 11, and (iii) a

component of 26S proteasome-mediated protein degradation radiation sensitive (RAD) 23. A number of Auxin response factors (ARFs) and regulatory proteins involved in flower development-related cell division and differentiation, some of which are connected to regulating seed development (Pillitteri, Bemis, Shpak, & Torii, 2007; L. Wang et al., 2011; Wu, Tian, & Reed, 2006), were also identified.

Some soybean SVs that were expressed during the mid-to-late phase showed differences with respect to their respective functional domains (ARF2 and 6, CPK11, and RAD23). For example, a novel CPK11 SV was missing the EF-hand (EFh) domain present in the canonical SV (Figure 5.4). The novel RAD23 SV lacked a ubiquitin (UBQ) superfamily domain. AS can also change the protein sequence, so that the SVs resemble different, but related proteins, which was observed for ARF6 and 8 (Glyma02g45100.N2 and 1). Interestingly, the novel ARF6 SV also lacked two domains, but had a new PB1 superfamily domain found in dimer-forming protein kinases (Diaz-Meco & Moscat, 2001; Ito, Matsui, Ago, Ota, & Sumimoto, 2001; Terasawa et al., 2001).

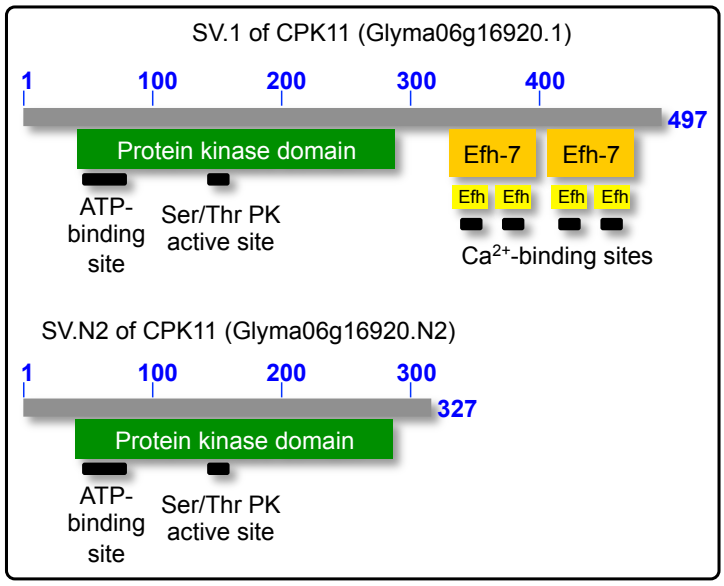

Figure 5.4 Graphical representations of functional domains present in known (Glyma06g16920.1) and novel (Glyma06g16920.N2) CPK11 protein isoforms. Resulting protein variants are shown in gray with length intervals and the actual lengths indicated by blue numbers. The corresponding domains are shown below each protein variant. Active and molecule-binding sites are shown below the relevant functional domains. The sizes and arrangements of these domains and sites reflect the reality. The novel CPK11 protein variant retained the protein kinase domain, but lost the EF-hand Ca^{2+}-binding domain. Efh, EF-hand domain; PK, protein kinase.

5.4.6 AS events related to ABA and/or FUS3 action during mid-to-late maturation

The mid-to-late super-clusters included SVs that showed increased transcript levels during the stages of steady-state seed storage compound accumulation (E. Collakova et al., 2013). Only three alternatively spliced FUS3 targets connected to the B3 network, belonging to the mid-to-late super-cluster were identified: (i) the transcriptional regulators HAP2A nuclear factor YA (NF-YA) 1 and bZIP66 (ABA-responsive element binding protein AREB3) and (ii) the E3 UBQ ligase DREB2A-interacting protein (DRIP) 2. While the HAP2A protein was not affected by AS, bZIP66 Glyma03g00580.N2 had ten additional amino acid residues at the C terminus and an extended 5'-UTR. The DRIP2 SV Glyma02g15980.N2 lacked the C3HC4-type Really Interesting Gene (RING) finger domain important for protein-protein interactions of UBQ ligases (Metzger, Hristova, & Weissman, 2012) and belonged to the early super-cluster.

5.4.7 AS events related to ABA and/or FUS3 action during DT

The DT super-cluster contains predominantly ABA-related SVs that showed basal transcript levels during all seed filling phases and high expression in yellow embryos at day 55 (E. Collakova et al., 2013). ABA-related SVs included transcripts encoding proteins similar to: (i) the epigenetic regulator histone deacetylase (HDAC) 6 associated with chromatin remodeling, (ii) two regulatory components of ABA- and G-protein related receptors REGULATORY COMPONENTS OF ABA RECEPTOR 3 (RCAR3) and G-PROTEIN COUPLED RECEPTOR 1 (GCR1), respectively, (iii) the transcriptional regulators of ABA signaling no apical meristem (NAM) TF (ATAF1) and ABI5-binding proteins (AFPs), (iv) sucrose nonfermenting related kinase protein (SnRK2.6), and (v) signaling-related phospholipase D delta (PLDdelta). ATAF1 was also identified in the FUS3 regulon. The only other FUS3-regulated SVs related to seed maturation and expressed in DT were those encoding saposin-like Asp proteases.

Differences with respect to protein length, number of domains (e.g., PLD delta, ATAF1, saposin-like Asp proteases), the absence of any known domains, (AFP4), and length of either the 5' or the 3' UTR (GCR1) occurred among different SV pairs. Variants of the same protein that showed similarity to different, but related, Arabidopsis proteins were also identified. While the SV Glyma04g38560.1 was similar to ATAF1 (At1g01720), Glyma04g38560.N2 was more related to At5g63790, a NAM domain-containing protein (NAC) 102, than to ATAF1. Similarly, Glyma17g01500.1 was similar to the saposin-like Asp protease At1g62290, but N3 resembled a different vacuolar protease (At1g11910). Interestingly, AS did not affect the structure of the G protein-coupled receptor domain in the novel, slightly shorter GCR1 protein, instead, the two SVs differed with respect to their 3'-UTRs. Differential expression was also observed in other SV pairs. The novel PLDdelta and HDAC6 SVs belonged to the early super-cluster, and the novel GCR1 SV to the mid-to-late super-cluster.

5.4.8 Antisense events related to ABA and/or FUS3 action

SV groups of ABA- and FUS3-related genes that were co-expressed during the same developmental phase showed multiple differences. AS lead to the production of lncRNAs and/or differences in protein sequence, number and types of functional domains, in 5'-

and 3'-UTR length and sequence, and expression patterns. While these changes were detected in all phases of soybean embryo development, the occurrence of antisense transcripts among the ABA- and/or FUS3-related transcripts was confined to the early and DT phases. Antisense transcripts expressed at the early phase include those associated with genes encoding ABA glucosylase, L1L, BAM2, and genome-uncoupled (GUN5) (a putative ABA receptor at the chloroplast envelope) (Dubrovina, Kiselev, & Zhuravlev, 2012; Mochizuki, Brusslan, Larkin, Nagatani, & Chory, 2001). The only exception was an antisense transcript associated with a putative cytokinin transporter *PUP1* gene. This antisense transcript co-expressed with its sense transcript during the mid-to-late phases.

Several antisense transcripts were also detected at day 55 of embryo development (DT) and appear to be connected to processes involving interactions of ABA with other phytohormones. Overall, 23 transcript pairs, regardless of any relation to ABA signaling, in which one member of each pair was antisense, together with seven single antisense transcripts without an accompanying sense transcript were detected at day 55. Among this population, several transcripts encoding proteins related to GA or ethylene signaling were present in both antisense and sense orientations and included GA2 oxidase, several ERFs, and PIL5 protein.

5.5 Generation and analysis of co-expression network

Co-expression networks have been used to infer potential gene interactions and functions (Y. X. Wang & Huang, 2014; S. Zhang, Jin, Zhang, & Chen, 2007). However, the majority of these networks have been limited to genes due to lack of isoform-specific transcript information. Here, ArrayMining (Glaab et al., 2009) was used to obtain an isoform-specific co-expression network for the set of 2,938 transcripts (Figure 5.5A). In the resulting network, each node represents a transcript and is colored according to its respective super-cluster. To reveal possible specific relationships among transcripts belonging to ABA- and FUS3-related events, the 318 transcripts were identified within the co-expression network. These ABA- and/or FUS3-related transcripts were used to generate a sub-network, reflecting temporal expression in the context of six super-clusters

Figure 5.5B). Of the 318 total transcripts encoded by target genes associated with FUS3- and/or ABA–related function, 311 transcripts were located within the three super-clusters corresponding to the three major phases of soybean embryo development.

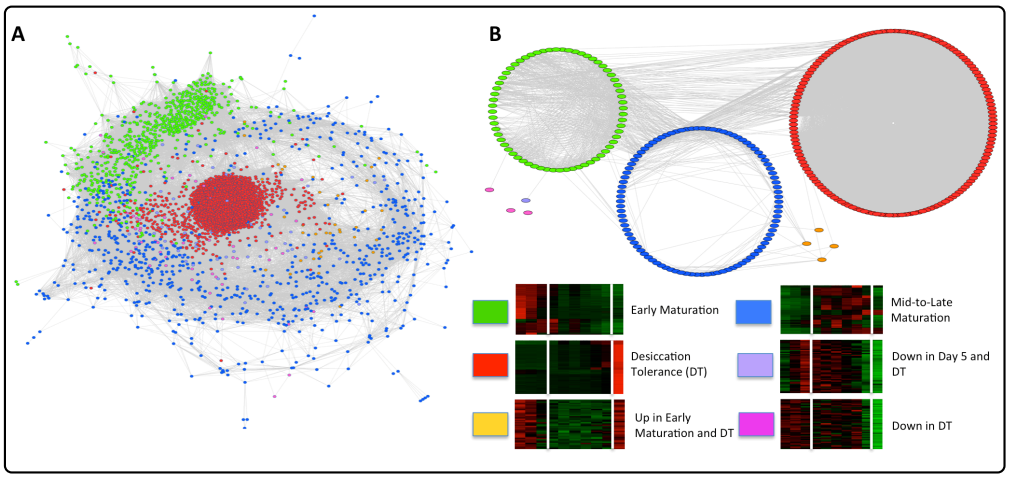

Figure 5.5 Isoform-specific co-expression network for the set of 2,938 transcripts. ArrayMining was used to construct the co-expression network. The resulting network was visualized in Cytoscape. Transcripts (nodes) are color-coded based on their respective super-cluster. (A) Co-expression network visualization of the 2,938 transcripts in the organic layout. (B) Co-expression sub-network for all 318 ABA-related and FUS3-regulated targets (see Results Section 4.4) grouped by super-cluster. The relative density of edges (in gray) reflects different degrees of connectivity among super-clusters. The majority of ABA- and FUS3-related transcripts (310) were located in the three major super-clusters.

5.5.1 Identification of the hubs

Although the co-expression network separates different super-clusters, interpretation of connections among the nodes remained intractable due to the large size of this network. To address this problem, we identified the most highly connected nodes (hubs) within each super-cluster. Hubs are the key network properties that reduce network complexity to the major connectors. In all cases, only one transcript derived from the same soybean gene was found to be a hub. Available functional information for hubs is presented in Figure 5.6. Approximately 50%, 13%, and 86% of transcripts were significantly connected to the hubs represented by transcripts of diverse functions in the early, mid-to-late, and DT super-clusters, respectively. Five hubs with a large number of associated nodes were identified in the case of DT. Among the DT-associated hubs were transcripts encoding soybean proteins similar to Arabidopsis peroxin 19 targeted to the peroxisome (Hadden et al., 2006), PATATIN-like protein 6 involved in lipid and auxin

signaling (Labusch et al., 2013), redox-related GST PH9 protein implicated in JA signaling, an F-box protein associated with an E3 UBQ ligase complex (Kuroda, Yanagawa, Takahashi, Horii, & Matsui, 2012), and a lncRNA transcribed from a homolog of At1g60940, SnRK 2.10 involved in ABA signaling (Umezawa et al., 2013).

Hub SV	Degree of connectivity	Annotation and function	Non-hub SV	SV differences
Glyma06g02120.1	50.3%	Glyma06g02120.1/ unknown	Glyma06g02120.N2	5'- and 3'-UTRs, protein length
Glyma09g35061.1	51.6%	RING/U-box superfamily protein/ chloroplast/ unknown	Glyma09g35061.N3	3'-UTR, protein length, domains
Glyma16g02650.2	50.3%	auxin response factor 9 (ARF9)/ affects petal growth, cell division and expansion	Glyma16g02650.N3	5'- and 3'-UTRs, protein length, domains
Glyma06g35330.1	50.3%	basic helix-loop-helix (bHLH) DNA binding/ unknown	Glyma06g35330.N2*	coding potential
Glyma19g32070.1	50.3%	magnesium chelatase/ GUN5 /ABA signaling/ retrograde signaling from the plastid to nucleus	Glyma19g32070.N2*	coding potential
Glyma08g45210.N2	13.3%	alpha-glucan phosphorylase/ drought stress	Glyma08g45210.1	5'- and 3'-UTRs, protein length, super-cluster
Glyma11g06290.3	12.6%	SKU5 similar 17 (SKS17) response to karrikin	Glyma11g06290.2	5'-UTR, super-cluster
Glyma12g36260.N12	12.9%	light-regulated zinc finger protein 1 (LZF1)/ regulates chloroplast biogenesis, and TRX and ferredoxin expression	Glyma12g36260.2	5'- and 3'-UTRs, protein length, domains, super-cluster
Glyma08g02840.1	12.9%	ER Stress Osmotic Stress/ Development and Cell Death domain	Glyma08g02840.3	3'-UTR, protein length, super-cluster
Glyma17g15080.2	85.6%	peroxin 19-2 (PEX19-2)/ targeting to peroxisome	Glyma17g15080.1	3'-UTR, protein length
Glyma10g35550.1	85.8%	PATATIN-like protein 6/ PLA III alpha/ auxin signaling/ lipid and auxin signaling	Glyma10g35550.N2	5'-UTR
Glyma14g03470.1	85.8%	glutathione-S-transferase PH9/ JA signaling	Glyma14g03470.N2**	sense coding/ antisense noncoding
Glyma18g38120.N2	85.6%	F-box protein/ part of specific E3 ligase complexes/ function unknown	Glyma18g38120.1	3'-UTR
Glyma04g38270.N2*	86.1%	Glyma04g38270.N2/ SnRK2-10/ ABA signaling	Glyma04g38270.1	coding potential

Figure 5.6 Hubs associated with the three major super-clusters. An in-house python program was used to identify the most highly connected nodes (transcripts) within each super-cluster. The corresponding non-hub SV is also included in the table. Structural or functional differences among these SV-pairs are also indicated. Super-clusters are color-coded as follows: green for early maturation, blue for mid-to-late maturation, and red for DT.

Degree of connectivity reflects how well transcripts are connected to each hub (number of connections/ total number of transcripts). *lncRNA, **antisense lncRNA.

5.5.2 Identification of the nearest neighbors of GCR1 and CPK11

GCR1 and CPK11 are two putative regulators of seed development belonging to the group of 318 ABA- and/or FUS3-related transcripts. *GCR1* and *CPK11* pre-mRNAs were alternatively spliced in developing soybean embryos and the resulting SVs were present in different super-clusters. The domain composition of each GCR1 and CPK11 SVs was confirmed by using the InterPro database (Apweiler et al., 2001). In Arabidopsis, GCR1 (At1g48270) is an ABA-responsive, G-protein-related receptor component distinct from the well-studied RCAR group of receptors (Warpeha et al., 2007). CPK11 (At1g35670) is a protein kinase acting as a positive regulator of ABA/FUS3-mediated responses during seed filling (Zhu et al., 2007). Guilt-by-association of GCR1 and CPK11 SVs with transcripts of known functions can yield improved understanding of their regulation and function.

To further elucidate isoform-specific functions of these two important regulators, the nearest neighbors of GCR1 and CPK11 were identified in the corresponding sub-networks (Figure 5.7) originating from the ABA/FUS3-related co-expression network (Figure 5.5B). The two GCR1 SVs were expressed at the DT and mid-to-late stages, respectively, and were associated with two distinct groups of transcripts. The nearest neighbor group comprising 38 nodes representing SVs expressed during DT was associated with Glyma17g33480.1-encoded GCR1. Transcripts associated with Glyma17g33480.N3 were differentially expressed during the early and mid-to-late phases of embryo development.

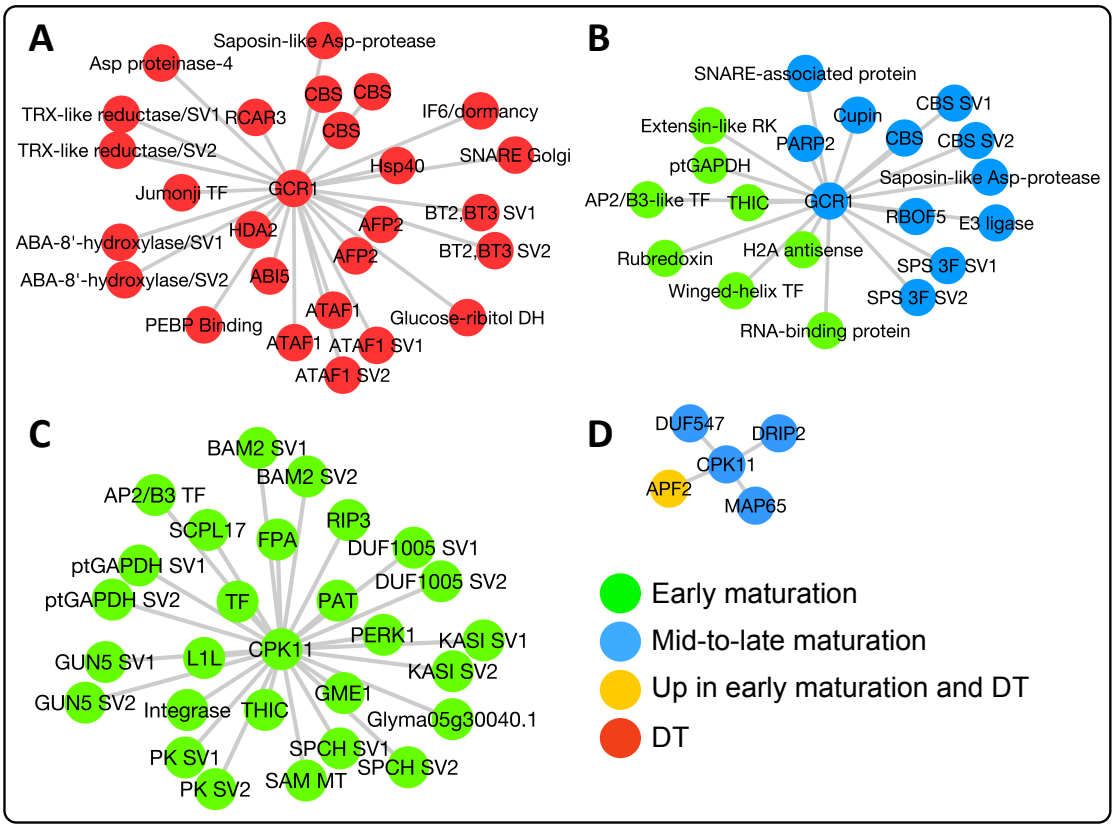

Figure 5.7 Sub-networks showing the nearest neighbors of the known and novel GCR1 and CPK11 SVs. (A) Known GCR1, (B) novel GCR1, (C) known CPK11, and (D) novel CPK11. The sub-networks were generated by extracting the first neighbors of these four transcripts from ArrayMining results shown in Figure 5.5. Nodes are color-coded based on their corresponding super-cluster.

The nearest neighbors belonging to the DT-based sub-network were primarily related to signaling or defense. Three coding and one noncoding ATAF1 transcripts were also among the nearest neighbors of GCR1, as were lncRNAs of heat shock protein (Hsp40) and BT2/ BT3. Only two lncRNAs of unknown function were identified as the nearest neighbors of the novel GCR1. A number of coding transcripts were also identified as strongly co-expressing with either the known or novel GCR1 SV. These transcripts encoded enzymes involved in seed filling or germination-related metabolism, including glucose/ribitol dehydrogenase, saposin-like Asp proteases, thiamine biosynthesis protein C (THIC), plastidic glyceraldehyde-3-phosphate dehydrogenase A (GAPDH), and sucrose-phosphate synthase. Homologs of regulatory proteins, including a histone deacetylase (HDAC), Hsp40, and several SVs encoding cystathionine beta-synthase (CBS)-domain-containing proteins that are associated with SnRK1 and energy sensing (Fang et al., 2011; Gissot et al., 2006) were also identified. Thioredoxin, rubredoxin, and

nicotinamide adenine dinucleotide phosphate (NADPH oxidase, RBOF5) represented the redox-related proteins. It is notable that many SVs present in the sub-networks with the centered GCR1 SVs were “singletons”, that is to say that only one member of a given SV group was present as a node in the sub-network. Interestingly, counterparts of some of these singletons associated with the known CPK11 SV and also belonged to the FUS3 regulon (GAPDH, THIC, saposin-like Asp proteases, and AP2/B3-like TF).

The full length CPK11 (Glyma06g16920.1) strongly co-expressed with several ABI3- and/or FUS3-regulated genes encoding alternatively spliced regulatory proteins (L1L, GUN5) (Figure 5.5B). Both BAM2 SVs and a Ser carboxypeptidase-like protease (SCPL17) were identified as the nearest neighbors of CPK11 that also belonged to the FUS3 regulon. The novel SVs of S-adenosyl-L-methionine methyltransferase (SAM MT), pyruvate kinase, and SPEECHLESS (SPCH) were also identified as the nearest CPK11 neighbors and reclassified by CodeWise as lncRNAs with high probabilities (84%, 99%, and nearly 100%). For metabolism-related processes, the corresponding SVs encoded enzymes involved in fatty acid/oil biosynthesis or storage (fatty acid synthase KAS1), amino acid metabolism (Asp/2-oxoglutarate aminotransferase PAT), and ascorbate biosynthesis (guanosine diphosphate (GDP)-mannose-3,5-epimerase 2 GME1). The CPK11 SV Glyma06g16920.N2 had only four nearest neighbors, two of which were connected to the B3 network (AFP2 and DRIP2) and the novel AFP2 SV was predicted to be lncRNA with a 95% probability.

5.6 Discussion

Current high-throughput transcriptomics data reveal a global occurrence of diverse types of transcriptional and post-transcriptional events, leading to the formation of transcripts of different coding potential and strand orientation. Knowing the coding potential and other characteristics of transcripts, including sequence similarity and presence of functional domains in resulting proteins, represents a first step towards discovering novel functionalities. We developed an integrated computational framework involving (i) a transcriptome-wide analysis of functional domains in proteins and an in-house SVM classifier, CodeWise, that categorizes transcripts as coding and noncoding and (ii) a relatively small-scale network analysis of 2,938 transcripts focusing on temporally driven co-expression and co-regulation.

5.6.1 Landscape of transcripts in developing soybean embryos

CodeWise was used to classify all transcripts detected in developing soybean embryos and, in combination with other tools, to analyze a set of 2,938 differentially expressed and alternatively spliced transcripts in terms of coding potential, expression timing, and changes in number and types of domains. The time period in seed development examined in this study extended from early maturation through the acquisition of dormancy and DT. Our analyses demonstrated the existence of a changing population of multiple types of transcripts over this part of soybean embryo development.

Interestingly, a relatively high proportion of coding and noncoding transcripts with no known domains were detected overall (27% of total), especially during DT. Many coding transcripts were predicted to encode small proteins lacking known domains (<120 amino acids; encoded by small ORFs). Considerable conservation of small ORFs across five leguminous species (including soybean) and Arabidopsis has been demonstrated (Guillen et al., 2013), suggesting that these are bona-fide proteins that probably act through conserved mechanisms. Known pathways, including sucrose signaling, in which small ORFs participate as "peptoswitches", were identified in plants (Jorgensen & Dorantes-Acosta, 2012). We have identified instances of transcripts that were classified as coding with no known domains that are associated with ABA signaling, but their potential function as peptoswitches remains to be investigated.

Long noncoding and antisense transcripts have also been implicated in regulating development and signaling in plants (Bardou et al., 2014; Di et al., 2014; James et al., 2012). With respect to coding/long noncoding or sense/antisense transcript pairs, a starting hypothesis is that their co-expression leads to either chromatin modification and/or degradation or stabilization of the sense mRNA (H. Wang et al., 2014; Jeremy E. Wilusz et al., 2009; L. Yang et al., 2014). The majority of SVs (80%) derived from the same gene (SV group) belonged to the same super-cluster, including long noncoding and antisense transcripts and their respective coding partners. The overall significance of these co-expression results is not yet clear, but it could be reflecting important conserved transcriptional and/or post-transcriptional regulatory mechanisms.

5.6.2 ABA- and FUS3-related transcripts were highly connected within the co-expression network of developing soybean embryos

ArrayMining (Glaab et al., 2009) was used to generate an isoform-specific co-expression regulatory network for the set of 2,938 transcripts (Figure 5.5A). The resulting co-expression network showed three different kinds of strong associations among the transcripts present in the different super-clusters (Figure 5.5B). First, transcripts from ABA- and FUS3-related regulons were identified within the overall network, revealing a specific sub-network. Transcripts within this sub-network were tightly clustered primarily around the three major super-clusters (early maturation, mid-to-late maturation, and DT). This clustering validated the original arrangement of the data into these temporally-based super-clusters. Second, GCR1 and CPK11 SVs expressed at different phases were found to have mostly distinct nearest neighbors, though some SVs were shared between the GCR1 and CPK11 sub-networks, providing a link between AS-related regulation of ABA- and FUS3-mediated signaling. In the case of the canonical GCR1 SV, the nature of its nearest neighbors may correspond to a specific coordinated regulatory mechanism involving AS, chromatin remodeling, redox-related processes, and signaling during DT. The presence of several lncRNAs among the nearest neighbors of GCR1 suggests that AS events involving the production of these lncRNAs are part of a distinct regulatory mechanism related to GCR1 action. Third, five hubs (including a lncRNA transcribed from a homolog of At1g60940, SnRK 2.10) with a large number of associated nodes were identified computationally in the case of DT, providing a connection between AS, redox regulation, and signaling pathways.

5.6.3 Evidence for post-transcriptional events leading to coordinated pre-mRNA splicing

Transcripts of some of the best-studied TFs and ABA biosynthetic genes that are known to regulate seed development (ABI3, FUS3, and 9-cis-epoxycarotenoid dioxygenases (NCED) 1, 4, and 5 (R. Finkelstein, 2013)) were present at relatively high and stable levels throughout soybean embryo maturation. Activities of ABI3 and FUS3 are also regulated through protein phosphorylation (Lynch, Erickson, & Finkelstein, 2012; Tsai & Gazzarrini, 2012) and proteosomal degradation (Q. S. Lu et al., 2010;

Stone, 2014). ABA-mediated signaling leads to the induction of specific SnRK kinases activating FUS3 and ABI3 (Bhaskara, Nguyen, & Verslues, 2012; Lynch et al., 2012; Schweighofer, Hirt, & Meskiene, 2004). Phosphorylation of FUS3 increases the stability of these short-lived proteins (Tsai & Gazzarrini, 2012). However, SnRK1.1 (represented by Glyma08g26180 and Glyma08g26191) transcript levels remained stable during soybean embryo development (E. Collakova et al., 2013), suggesting that any differential regulation mediated by this kinase would have to be at its translational and/or posttranslational levels.

While FUS3 transcript levels and, possibly, protein activity remained stable in developing soybean embryos, many SV pairs of ABA-related and FUS3-regulated genes were differentially expressed and did not co-express with FUS3. Differential expression of SVs originating from the same pre-RNAs suggests the occurrence of post-transcriptional events, which can globally influence transcript levels and stability. This is consistent with the observation that many ABA-related and FUS3-regulated transcripts that originated from different, but functionally related genes were co-expressed in the data set. It appears that specific splicing components can regulate differential splicing of groups of pre-mRNAs during specific stages of embryo development, leading to differential temporal expression of these SVs. It is tempting to hypothesize that this coordinated splicing ("co-splicing") may be a common regulatory mechanism employed in signaling processes within the embryo developmental programs. AS was proposed as a global regulatory mechanism in seed dormancy (Graeber, Nakabayashi, Miatton, Leubner-Metzger, & Soppe, 2012), and it also could be the case in developmental transitions within embryo maturation.

5.6.4 Potential roles for alternate pathways and antisense regulation in phytohormone interactions during late seed maturation and germination

The majority of ABA-related SVs corresponded to Arabidopsis genes already documented to participate in dormancy or, in some cases, germination. It is also interesting to note that a SV of a homolog of RCAR3/ PYRABACTIN RESISTANCE 1-LIKE (PYL 8) implicated in ABA signaling that promotes dormancy (Saavedra et al., 2010) was differentially expressed during DT, whereas an SV corresponding to PYL6

was expressed during the mid-to-late phase. Given the differences in the population of SVs that are ABA-related and were expressed during one or other of the two phases, it is possible that distinct ABA signaling pathways are in operation during the two developmental phases.

Transcripts of putative soybean homologs of Arabidopsis genes known to be associated either with ABA-related events (including, but not restricted to signaling), and/or to be targets of FUS3 were associated with specific AS events or antisense expression. AS resulted in altered numbers or types of domains and production of coding and noncoding transcripts, which has consequences for molecular interactions, epigenetic events, regulation of protein activity, and subsequently function. This information was incorporated into proposed signaling pathways using the Beacon editor (Figure 5.8), with extensive use of published information from genetic or biochemical studies regarding observed mutant phenotypes or biochemical characteristics of the proteins involved.

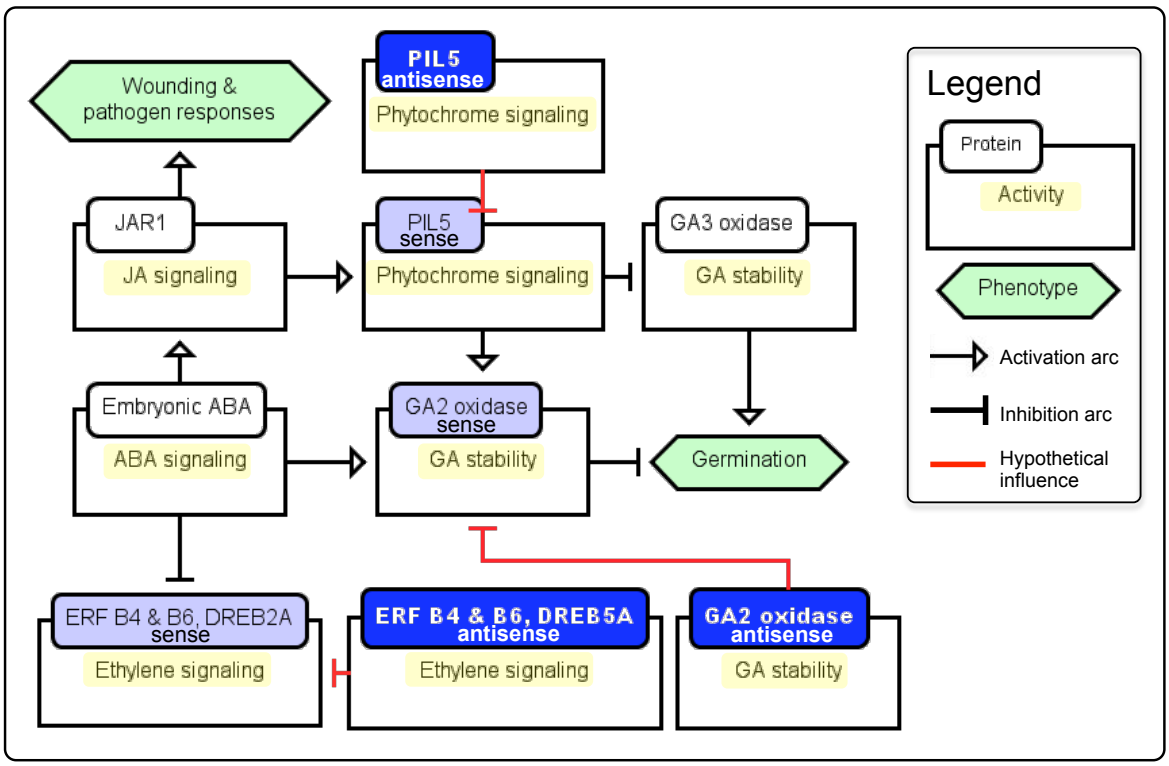

Figure 5.8 Proposed interactions of ABA with GA, ethylene, and phytochrome signaling pathways involving potential post-transcriptional regulation by antisense transcripts during desiccation phases and/or seed germination. These simplified signaling pathways were drawn in the Beacon editor. Signaling components involving sense-

antisense transcript pairs are shown in two different shades of blue. The influence arcs do not imply the direct interactions; they indicate a positive or negative influence between the corresponding glyphs.

The occurrence and expression changes of ERFs, PIL5, and GA2 oxidase antisense transcripts were validated by qPCR. The occurrence of these antisense transcripts during DT can reflect the well-documented antagonism of ABA with ethylene (Holdsworth et al., 2008). The presence of antisense transcripts corresponding to PIL5 and the GA degrading enzyme GA2 oxidase that is activated by PIL5 (Oh et al., 2006) is not readily explicable as an expected repressive effect on GA metabolism during dormancy (Figure 5.8). There are conflicting reports on whether PIL5 acts to trigger release from dormancy or inhibit germination (Graeber et al., 2012; Oh et al., 2006). It is possible that this is an instance of a positive regulatory effect of an antisense transcript on RNA stability (Jabnoune et al., 2013) or that these antisense transcripts are stored for germination to suppress PIL5 and GA2 oxidase to release dormancy.

5.6.5 Inferring transcript and protein functions in the context of regulation of seed filling

In Arabidopsis, seed maturation starts with the expression of LEC1, which induces transcription of L1L, LEC2, FUS3, and ABI3 (R. Finkelstein, 2013; Kwong et al., 2003). LEC1 and L1L represent HAP3-type subunits of heterotrimeric CCAAT-box binding factors (Calvenzani et al., 2012; Lee et al., 2003; Yazawa & Kamada, 2007), which activate transcription of genes involved in the synthesis and accumulation of seed storage compounds (Siobhan A. Braybrook & Harada, 2008; Kagaya et al., 2005; Mendes et al., 2013). ABI3 and FUS3 are positive regulators of the ABI5 family of TFs, including bZIP66, promoting accumulation of seed storage compounds (Bensmihen, Giraudat, & Parcy, 2005; R. Finkelstein, 2013). L1L, HAP2A, and bZIP66 are components of the B3 network that were alternatively spliced. The novel soybean L1L variant is a noncoding antisense transcript confirmed by qPCR to be expressed in the early maturation phase. Because the L1L transcripts showed negatively correlated expression profiles, it is possible that the antisense transcript negatively regulates levels of the sense transcript in soybean. LEC1-mediated transient activation of L1L (Mendes et al., 2013) could be inhibited post-transcriptionally through this antisense transcript to confine the L1L

presence to the early phases of seed filling. Although the L1L and HAP2A proteins physically interact (Mostafavi, Ray, Warde-Farley, Grouios, & Morris, 2008; Yazawa & Kamada, 2007; Zuberi et al., 2013), their transcripts were present at different times during soybean embryo development, which makes concurrent protein-protein interactions and transcriptional regulation unlikely. Genetic evidence in Arabidopsis supports the co-regulation of HAP2A and LEC1/ L1L transcription in early seed filling (Shu et al., 2013). These two HAP2A isoforms are probably not the bona-fide L1L interacting partners in soybean.

The B3 network contains transcriptional regulators. Activities of these TFs are also regulated at the post-transcriptional level. Several ABA-related and/or FUS3-regulated genes encoding protein kinases (BAM2, CPK11) and components of the 26S proteasome (DRIP2 and RAD23) were alternatively spliced and, in some cases, could be associated with specific TFs. Arabidopsis Ca^{2+}-dependent protein kinase CPK11 acts in parallel with SnRK2 kinases to phosphorylate and activate specific bZIP TFs (ABFs, ABI3, and ABI5) involved in promoting dormancy (Lynch et al., 2012; Uno et al., 2000; Yoshida et al., 2015; Zhu et al., 2007). The full-length CPK11 (Glyma06g16920.1) strongly co-expressed with L1L and several FUS3-regulated metabolic genes. Both BAM2 transcripts, SCPL17, and AP2/B3 TF were also identified as the nearest neighbors of CPK11 regulated by FUS3. While SCPL17 and AP2/B3 are functionally uncharacterized, BAM2 is involved in flower and fruit development (DeYoung et al., 2006; Durbak & Tax, 2011). Association of these transcripts with CPK11 and FUS3 implicates their potential involvement in early seed filling signaling.

Interestingly, the novel CPK11 SV was expressed during mid-to-late seed filling and lacked both EFh domains present in the canonical SV. The EFh domains occur as pairs and are responsible for changing protein conformation upon Ca^{2+} binding to modulate protein activity (Grabarek, 2006). The novel CPK11 variant could phosphorylate its targets during mid-to-late seed filling independently of Ca^{2+}-mediated signaling. Its direct co-expressers AFP2 and DRIP2 are associated with the B3 network through ABA/ABI3/ABI5 and FUS3 signaling, respectively. DRIP2 ubiquitinates the positive regulator of ABA-independent drought responses DREB2A (Q. Liu et al., 1998; Mizoi et al., 2013; Qin et al., 2008). Protein-protein interactions of the novel DRIP2 SV are likely

compromised due to the absence of the C3HC4-type RING-finger domain and functions of this SV in the B3 network and early seed filling phases remain to be elucidated.

RAD23 proteins are similar to UBQ and are involved in transporting ubiquitinated proteins to the 26S proteasome for degradation (Farmer et al., 2010; Fatimababy et al., 2010). RAD23 transcription was suppressed in response to ABA and in the protein phosphatase 2C *abi1* mutant (Hoth et al., 2002). In addition, RAD23 was identified as an interacting partner of a rice ABI3 homolog (Schultz & Quatrano, 1997), placing it downstream of FUS3, SnRKs, and CPK11 in the B3 network as a negative regulator of ABI3 activity. The known RAD23B SV belonged to the early super-cluster and could be involved in delivering ABI3 for degradation to the 26S proteasome in developing soybean embryos. It is not clear whether the novel SV was active as it lacked a UBQ superfamily domain (Watkins, Sung, Prakash, & Prakash, 1993).

5.7 Conclusions

This report demonstrates the usefulness of our integrated computational framework for the analysis of transcriptomics data, leading to prediction of experimentally testable and specific hypotheses concerning the functions of expressed transcripts. The behavior of many of the coding transcripts identified here has not been studied previously in seeds or at all in the case of long noncoding and antisense RNAs. Taken together, a common functional theme integrates the hubs related to DT in regulation, stress responses, and phytohormone signaling and suggests the existence of distinct ABA-related pathways, specific to different phases of soybean seed development. Several components of the B3 regulatory seed filling network were subjected to AS, potentially leading to differential expression and regulation as well as novel functionalities. Our computational approaches facilitated identification of other regulators possibly involved in seed filling and desiccation and dormancy induction phases of soybean embryo development.

6 Potential targets of VIVIPAROUS1/ABI3-LIKE1 (VAL1) repression in developing *Arabidopsis thaliana* embryos[1]

This chapter is produced from (Schneider et al., 2015) with permission from *The Plant Journal*.

Schneider, A., **Aghamirzaie, D.**, Elmarakeby, H., Poudel, A. N., Koo, A. J., Heath, L. S., Collakova, E. (2015). Potential targets of VIVIPAROUS1/ABI3‐LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos. *The Plant Journal*.

Abstract. Developing Arabidopsis seeds accumulate oils and seed storage proteins synthesized by the pathways of primary metabolism. Seed development and metabolism are positively regulated by transcription factors belonging to the LAFL regulatory network. The *VAL* gene family encodes repressors of the seed maturation program in germinating seeds, although they are also expressed during seed maturation. The possible regulatory role of VAL1 in seed development has not been studied to date. Reverse genetics revealed that *val1* mutant seeds accumulated elevated levels of proteins compared to the wild type, suggesting that VAL1 functions as a repressor of seed metabolism. However, metabolomes and the levels of ABA, auxin, and jasmonate derivatives did not change significantly in developing embryos in the absence of VAL1. Two *VAL1* splice variants were identified through RNA sequencing analysis: a full-length and a truncated form lacking the plant-homeodomain-like domain associated with epigenetic repression. None of the transcripts encoding the core LAFL network transcription factors were affected in *val1* embryos. Instead, activation of *VAL1* by FUSCA3 appears to result in repression of a subset of seed maturation genes downstream of core LAFL regulators as 39% of transcripts in the FUSCA3 regulon were de-repressed in the *val1* mutant. The LEC1 and LEC2 regulons also responded but to a lesser extent. Additional 832 transcripts that were not LAFL targets were de-repressed in *val1* mutant

[1] Andrew Schneider performed all the experiments and was involved in the biological data interpretation. Delasa Aghamirzaie and Haitham Elmarakeby were involved in the computational data analysis. Lenwood Heath advised in data analysis sections. Andrew Schneider, Ruth Grene and Eva Collakova performed all the biological data mining of the results.

embryos. These transcripts are candidate targets of VAL1, acting through epigenetic and/or transcriptional repression.

6.1 Introduction

The evolution of seeds allowed plants to conquer land by developing a dormant state that facilitates survival in seasonally unfavorable environments. Seed storage compounds, primarily consisting of oils, proteins, and carbohydrates, are synthesized in metabolically active developing seeds to provide energy and structural components for germinating seedlings (Baud, Dubreucq, Miquel, Rochat, & Lepiniec, 2008; Penfield, Pinfield-Wells, & Graham, 2006). Seed development can be divided into three major phases (Baud et al., 2008; David W Meinke, 1995). First, in embryogenesis, cells divide and differentiate to form shoot and root meristems and the embryo axis and cotyledons. Second, the maturation phase is characterized primarily by biosynthesis and accumulation of seed storage compounds. Third, seed development ends with the acquisition of desiccation tolerance and dormancy.

Seed storage compounds are synthesized through established metabolic pathways, while the regulation of these pathways is less well understood (Baud et al., 2008; H. Jia, Suzuki, & McCarty, 2014). In the model oilseed species *Arabidopsis thaliana*, regulation has been primarily studied at the level of transcription. Several transcription factors (TFs) are known to positively and globally regulate distinct aspects of seed development (H. Jia et al., 2014). These global regulators include the CAAT-box family protein LEAFY COTYLEDON 1 (LEC1) and the B3-family proteins ABCISIC ACID INSENSTIVE 3 (ABI3), FUSCA 3 (FUS3), and LEC2, together making up part of the core LEC1, ABI3, FUS3, LEC2 (LAFL) regulatory network (Sun et al., 2013). LEC1 and LEC2 function primarily during embryogenesis, while FUS3 and ABI3 activate maturation-specific processes. Together, the core LAFL TFs form a complex transcriptional network regulating seed development, with many overlapping targets, including mutual and auto regulation, and the participation of other TFs, such as the LEC1 homolog LEC1-LIKE and the basic leucine zipper TF bZIP67 (Mendes et al., 2013).

Most, but not all, of the known seed development regulators have been shown to function as activators of gene expression. However, TRANSPARENT TESTA GLABRA

1 (TTG1) was identified recently as a repressor of ABI3, LEC2, and several genes involved in seed storage compound synthesis in developing seeds (M. Chen et al., 2015). Other repressors have been defined to date exclusively in the context of repression of the seed maturation program in germinating seeds and seedlings (H. Jia et al., 2014). The VIVIPAROUS1/ABI3-LIKE (VAL) TFs are repressors of the core LAFL network in developing seedlings and are required for the transition from the embryonic to the vegetative state. Single knockouts of *VAL1* or *VAL2* genes do not have major effects on seedling and vegetative development, but the double *val1/val2* knockout results in seedlings that abort after 7 to 10 days following germination (Suzuki et al., 2007; Tsukagoshi, Morikami, & Nakamura, 2007). Seed storage proteins are substantially elevated in *val1*/*val2* seedlings, a phenotype related to the retention of the embryonic state. Microarray analysis showed that the core LAFL network genes were globally up-regulated in these seedlings (Suzuki et al., 2007). Although VAL1 and 2 are expressed in developing oilseeds, with the highest transcript levels during the transition from early to middle maturation, their potential roles in seed development and metabolism have not been investigated (Schmid et al., 2005).

VAL1 (At2g30470) contains four domains with potential roles in transcriptional regulation and/or chromatin/DNA binding (H. Jia et al., 2014). The B3 DNA-binding domain is utilized by many plant-specific transcriptional activators, such as ABI3, to bind directly to the Sph/RY motif (CATGCATG) in promoter regions of target genes (Monke et al., 2004). However, direct DNA binding by VAL1 to this domain has not been demonstrated in Arabidopsis (Guerriero et al., 2009; Suzuki et al., 2007; Swaminathan, Peterson, & Jack, 2008; Tsukagoshi, Saijo, Shibata, Morikami, & Nakamura, 2005). The B3 domain of GERMINATION-DEFECTIVE 1, which is a putative homolog of VAL1 in rice, was found to bind to the Sph/RY motif and potentially repress expression of reporter genes in Sph/RY-specific manner (Guo et al., 2013). Two of the domains associated with epigenetic regulation, a domain very similar to the canonical plant homeodomain (PHD) zinc finger domain, and a conserved zinc finger Cys- and Trp-containing (CW) domain, bind to histone 3 proteins that are trimethylated at the fourth Lys residue (H3K4me3) to activate transcription (Hoppmann et al., 2011; R. Sanchez & Zhou, 2011).

The PHD-like (PHD-L) domain, present in VAL1, contains three potential zinc fingers as opposed to the two Zn fingers that are present in the canonical PHD domains (Mouriz, Lopez-Gonzalez, Jarillo, & Pineiro, 2015; R. Sanchez & Zhou, 2011; Suzuki et al., 2007). The VAL1 PHD-L domain has been implicated in the repression of seed maturation genes in seedlings through the promotion of trimethylation of Lys residue 27 on histone 3 (H3K27me3) (Veerappan, Chen, Reichert, & Allen, 2014; Veerappan et al., 2012). In contrast to the H3K4me3-related activation of gene expression by the canonical PHD domain, trimethylation of K27 in histone 3 results in repression catalyzed by POLYCOMB REPRESSIVE COMPLEX 2 (PRC2) methyltransferases (Chanvivattana et al., 2004; Schubert et al., 2006). The VAL1 CW domain, on the other hand, has been demonstrated to bind directly to H3K4me3 (Hoppmann et al., 2011) and to interact with HISTONE DEACETYLASE 19 (HDA19) to repress seed maturation genes in seedlings (Zhou et al., 2013). VAL1 also contains an ethylene-responsive element binding factor-associated amphiphilic repression (EAR) motif, which was found to be necessary for repression of a reporter construct driven by a sugar responsive promoter (Tsukagoshi et al., 2005). EAR motifs are ubiquitous among repressors and likely function through recruitment of histone modifiers that target genes for chromatin modification (Kagale & Rozwadowski, 2011).

Here, VAL1 was investigated in the context of epigenetic and transcriptional regulation of seed development and metabolism. The *val1* mutant accumulated elevated protein levels in dry seeds compared to the wild type. Potential functions of *VAL1* in developing embryos were further investigated in transcriptomics and metabolomics time-course experiments covering the seed maturation and early desiccation phases of embryo development. Studies of Arabidopsis seed development have largely been limited to whole seeds or siliques due to the small seed size (M. Chen et al., 2015; B. H. Le et al., 2010; Sun et al., 2013). Only recently, laser capture microdissection technology enabled obtaining a global perspective on gene expression changes in various cell types throughout Arabidopsis seed development (Belmonte et al., 2013). This study represents a detailed investigation of the embryo-specific transcriptome and metabolome in Arabidopsis to provide correlation-driven bioinformatics predictions related to the action

and potential targets of a global epigenetic and transcriptional repressor functioning during seed development.

6.2 Results[1]

6.2.1 The *val1* mutant accumulates elevated levels of seed storage proteins

VAL1 has a known role in seedling establishment as a repressor of the embryonic program in Arabidopsis (Suzuki et al., 2007; Tsukagoshi et al., 2007). However, VAL1 is also expressed in developing seeds and siliques, where its negative effects on the embryonic program would appear contradictory. To investigate possible functions of VAL1 in developing embryos, various seed- and embryo-related phenotypes of the homozygous *val1* mutant (SALK_088606C) were analyzed, starting with assessing potential changes in seed storage compound levels. Wild type and *val1* mutant seeds accumulated 80.6 ± 8.6 and 93.3 ± 11.28 μg mg^{-1}, respectively, of proteins per dry weight (DW) of seed, representing an approximate 12% increase in protein content in *val1* seeds (22 biological replicates for wild type, 16 for *val1*, *p*-value < 0.00083). In contrast, the amount and composition of seed storage proteins did not appear to be affected in the *val1* mutant when analyzed through SDS-PAGE, but a marginal increase in protein, such as 12%, and with an 11% standard deviation, is difficult to capture by protein gels. Free and lipid-derived fatty acid levels of *val1* seeds remained unchanged.

6.2.2 The SALK_088606C mutant contains a single T-DNA insertion

The SALK_088606C mutant contains an insertion in the seventh exon of the *VAL1* gene (Figure 6.1A), which encodes part of the B3 domain. This insertion mutant has been characterized previously in studies using seedlings and mature plants (Sharma, Bender, Boyle, & Fobert, 2013; Sun et al., 2013; Suzuki et al., 2007; Tsukagoshi et al., 2007; Tsukagoshi et al., 2005; Veerappan et al., 2014; Veerappan et al., 2012), but not in seeds. In addition, the results of the T-DNA SEQ project (http://www.ncbi.nlm.nih.gov/nucgss/KO428765) showed that the parental line SALK_088606 contained a second T-DNA insertion in the 5'-untranslated region of an

[1] Andrew Schneider and Eva Collakova performed all the experiments in this chapter.

unrelated gene (At5g63350). PCR-based genotyping on plants grown from seed directly obtained from the Arabidopsis Biological Resource Center (ABRC, The Ohio State University) confirmed that this insertion is not present in the SALK_088606C mutant, and the expression pattern of At5g63350 is unaltered in *val1* mutant embryos, suggesting that this mutant line is free of any additional insertions other than that present in the 7th exon of the *VAL1* gene.

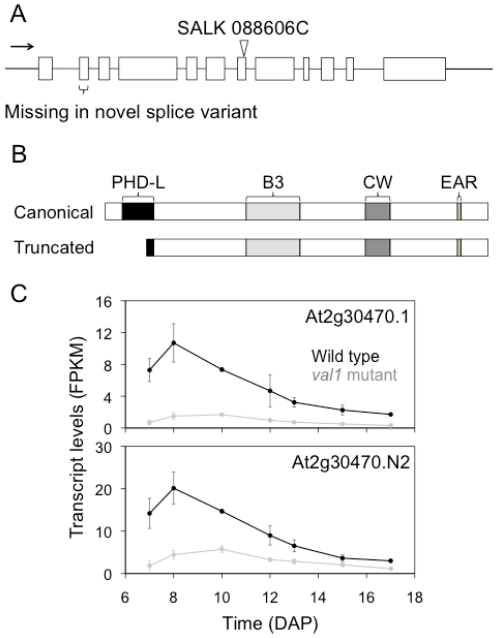

Figure 6.1 The *VAL1* gene and corresponding SVs. (a) The *VAL1* gene. The SALK_088606C mutant line carries a T-DNA insertion in the seventh exon, which encodes part of the B3 domain. The novel SV is missing the entire second exon. (b) Splicing of the *VAL1* premature RNA yields two SVs encoding two different protein variants. The full-length protein contains all four domains. The novel SV is missing 77% of the PHD-L domain. (c) Expression of *VAL1* SVs in developing Arabidopsis embryos. The full-length SV (At2g30470.1) is expressed nearly two fold less than the novel SV (At2g30470.N2) at each time point in the wild type. Transcript levels of both SVs appears reduced in developing *val1* embryos based on RNA-Seq transcriptomics.

6.2.3 Temporal aspects of Arabidopsis embryo development in wild type and *val1* mutant

The majority of seed biomass consists of storage compounds that are predominantly found in the embryo of dry seeds (Sébastien Baud et al., 2002; Baud et al., 2008; Higashi et al., 2006). The result showing increased protein levels in *val1* dry seeds suggests a role for *VAL1* in the regulation of metabolism in developing Arabidopsis embryos. In order to elucidate the specific role(s) of VAL1 in this context, wild type and mutant embryos at

the maturation and desiccation phases were investigated at the transcriptome and metabolome levels. The time course (Figure 6.2A) comprised early maturation (7 and 8 days after pollination (DAP)), middle maturation (10, 12 and 13 DAP), and late maturation/early desiccation (15 and 17 DAP) phases. The transcriptomes, metabolomes, and levels of selected phytohormones were analyzed in the wild type and *val1* mutant embryos, free from contaminating seed coats and endosperm. Results from these analyses are described in the following sections.

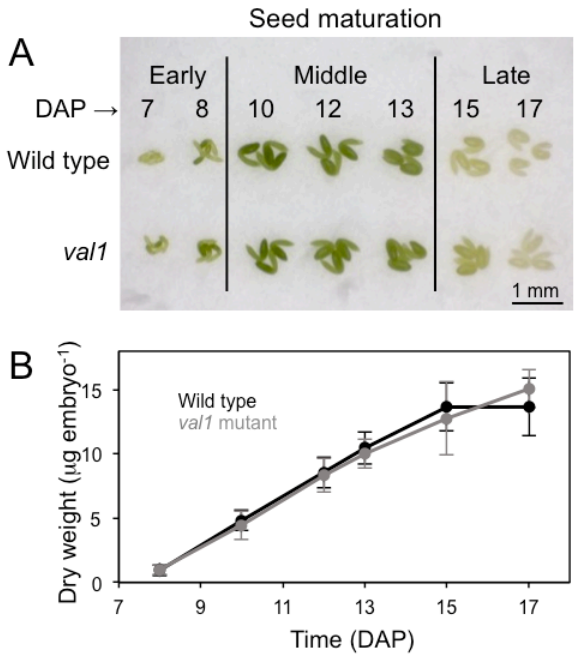

Figure 6.2 Developing wild type and *val1* mutant embryos show not visible phenotypes. (a) Embryos dissected from seeds of specified age from a representative wild type and *val1* mutant, respectively, plant are shown. Separation of embryo development into three major phases of seed maturation (early, middle, and late seed filling) is shown by vertical lines. (b) Biomass accumulation in developing wild type and *val1* mutant embryos. Averages ± SD of nine biological replicates for each line are shown. No differences in the appearance or DW of wild type and mutant embryos of comparable developmental stages were observed.

6.2.4 The *VAL1* gene encodes two splice variants and the *val1* mutant is likely a knock out

To investigate potential functions of *VAL1* in seed maturation, high-throughput RNA sequencing (RNA-Seq) was performed to analyze steady-state mRNA levels in developing wild type and *val1* embryos to identify transcripts that were differentially expressed over the time course between the wild type and *val1* mutant. Two *VAL1*-

derived splice variants (SVs) were detected. A "canonical" SV (At2g30470.1) encoding a protein possessing all four domains (PHD-L, B3, CW, and EAR) was detected in addition to an alternatively spliced SV (At2g30470.N2) that lacked the second exon. This truncated SV was predicted to be coding (coding probability was 1 and non-coding probability was 10^{-7}) by CodeWise. CodeWise is a support vector machine classifier that uses multiple sequence features in combination with the calculated value of the free energy of secondary RNA structure to classify transcripts as coding or non-coding with over 96% accuracy (Aghamirzaie et al., 2015). The resulting protein variant has an 85-amino acid truncation from the amino terminus that removes the majority (77%) of the PHD-L domain (Figure 6.1B). The two SVs were co-expressed throughout the time course in both wild type and mutant embryos (Figure 6.1C). *VAL1* transcript levels were 7.1 and 4.5 fold higher in wild type than in *val1* for the canonical and predicted SVs, respectively. Overall, there were 1.8- to 10.4-fold higher levels of the *VAL1* transcript levels, expressed as FPKM, in developing wild type compared with mutant *val1* embryos. The alternatively spliced SV was also more abundant (nearly two fold at each time point) than was the canonical SV in the wild type. Maximal expression for both *VAL1* SVs in the wild type and *val1* mutant was at 8 and 10 DAP, respectively.

The distribution of the reads across any given transcript is not apparent from FPKM values, the form of the data presented above. Examination of the actual read distribution is needed to determine whether insertion of the T-DNA resulted in truncation or alterations of transcripts in the *val1* mutant. Mapped reads across the *VAL1* transcripts for both wild type and *val1* are shown in Figure 6.3A. While the read density in the 5'-region upstream of the T-DNA in the *val1* mutant was similar to the wild type, reads were nearly absent from the region downstream of the T-DNA insertion site. Thus, it is likely that transcription of the *VAL1* gene in the mutant continued into the T-DNA, but rarely proceeded beyond the insertion site.

To unambiguously determine whether any wild type/full length *VAL1* transcripts were formed in the *val1* mutant embryos, the wild type region flanking the T-DNA insertion site was targeted for PCR amplification, using the wild type and mutant cDNA and the *VAL1*-specific primers originally used for genotyping. The expected 670-bp band representing the wild type portion of cDNA was robustly amplified in the wild type, even

when the wild type cDNA sample was diluted to equalize the possible *VAL1* template amounts between the wild type and mutant. In contrast, no bands were detected in the case of the *val1* mutant embryos (Figure 6.3B). Taken together, these results show that the T-DNA was not spliced out and, because no wild type *VAL1* transcripts were detected in the *val1* mutant, this mutant is a knockout as opposed to a knock down, provided that the resulting truncated protein (about 50% truncation) is not active.

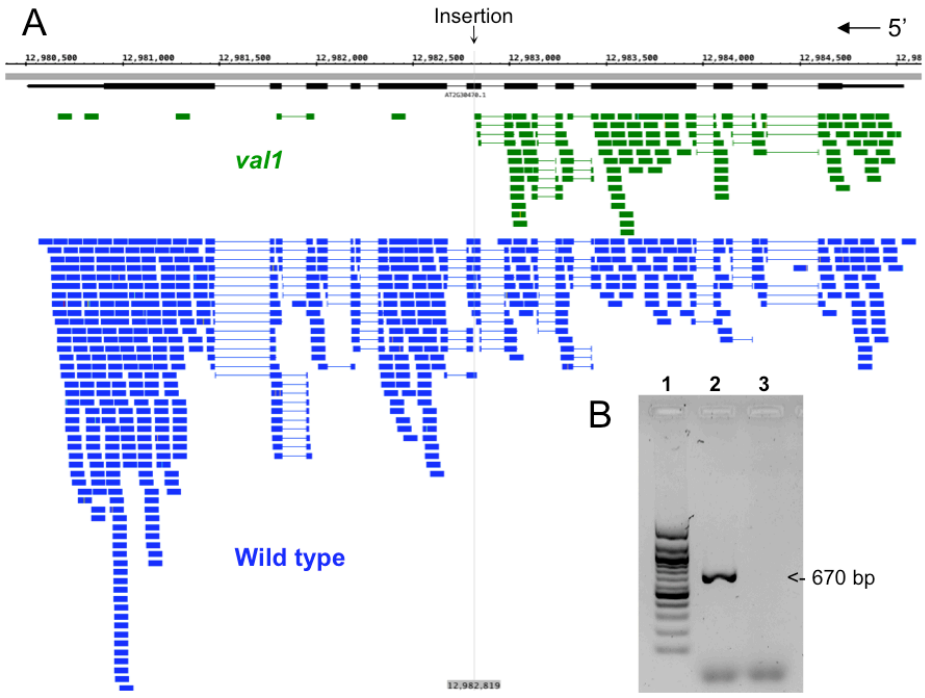

Figure 6.3 The *val1* mutant does not accumulate wild type *VAL1* transcripts in developing embryos. (a) Density of reads mapped to the wild type *VAL1* gene in wild type and *val1* embryos at a representative time point. The majority of reads in the *val1* mutant mapped to the region upstream of the T-DNA insertion site. Based on the density of the reads, transcription efficiency for the *val1* gene was not affected in the mutant. (b) PCR of the region flanking the T-DNA insertion site of the *val1* gene in 8-day-old embryos using the genotyping primers. Lanes: 1. 100-bp ladder; 2. wild type cDNA template diluted 12 fold; 3. *val1*cDNA undiluted template. An expected 670-bp band was amplified from the wild type cDNA, but not from the mutant.

Semi-quantitative reverse transcription PCR (qPCR) results for the two *VAL1* SVs in the mutant showed transcript levels that appeared to be comparable to the wild type. However, both *VAL1* SVs peaked at day 10 in the mutant rather than at day 8, which was observed for the wild type, which is seemingly inconsistent with the RNA-Seq results in Figure 6.1C. In order to distinguish these SVs, TaqMan primers and probes were designed to match unique sequences in the two SVs that were localized at the 5'-end of the gene, upstream of the T-DNA insertion site. Transcription in the 5'-end of the *VAL1*

gene was similar in the wild type and *val1* mutant based on read distribution (Figure 6.3A).

6.2.5 VAL1 regulates embryo development through FUS3 without a direct effect on the expression of the core LAFL TF genes

VAL1 represses expression of the core LAFL genes in seedlings in conjunction with VAL2 (Suzuki et al., 2007). Knocking out the *VAL1* gene resulted in increased protein levels in dry seeds, which is consistent with phenotypes observed in germinating *val1/val2* homozygous mutant seedlings. Therefore, it was possible that VAL1 has similar effects on developing Arabidopsis embryo transcriptomes, through a fine tuning of the seed developmental program by repression of the accumulation of seed storage compounds. A comparison of the wild type and *val1* transcriptomes revealed that 3,293 transcripts (encoded by 2,483 genes) were up-regulated, whereas 2,194 transcripts (encoded by 1,606 genes) were down-regulated at at least four time points in developing *val1* embryos compared to the wild type. The rationale for choosing this particular way of evaluating differential expression of genes and transcripts is provided in Materials and Methods. Unless otherwise stated, it will be used as the definition of differential expression between developing wild type and *val1* mutant embryos throughout this manuscript. Of the up-regulated transcripts, 244 (encoded by 190 genes) were significantly de-repressed (p-value < 0.05) at every time point in *val1* embryos relative to the wild type.

To determine whether VAL1 has similar effects on the transcriptome in developing embryos as in seedlings, the degree of overlap between the corresponding sets of differentially expressed transcripts in developing *val1* embryos and seedlings was determined. For Arabidopsis seedlings, the data set consisted of microarray data involving two biological replicates and four genotypes (wild type, single *val1* and *val2* mutants, and the double *val1/val2* mutant), in which 856 and 685 genes were up-regulated and down-regulated, respectively, at least four fold in *val1/val2* mutants compared to the wild type (Suzuki et al., 2007). We used our data set to identify those genes that were differentially expressed in both *val1/val2* seedlings and *val1* developing embryos. This cross-comparison of the data of Suzuki et al., (2007) with our data

revealed that 314 genes (466 transcripts) were up-regulated and 75 genes (105 transcripts) were down-regulated in both *val1*/*val2* seedlings and *val1* developing embryos. The same approach was applied to the subset of 188 genes (244 transcripts) that were significantly de-repressed at all time points as these transcripts are encoded by genes regulated by VAL1 consistently throughout our time course. This cross-comparison yielded 50 genes (68 transcripts) present in both data sets. Based on these results, a large proportion of the same genes appears to be repressed in both developing embryos and germinating seeds, suggesting that they could share similar LAFL-related regulatory pathways or at least some of their components. Down-regulation in the absence of VAL1 is likely representative of repression of repressors rather than direct activation. As such, there is an increased possibility for involvement of other (and diverse) TFs in these two different programs (vegetative growth versus embryonic program), which could explain the lower proportion of shared down-regulated than up-regulated genes.

In contrast to germinating seedlings, expression of the LAFL network genes was not affected by the *val1* mutation in developing embryos despite global changes detected in the transcriptomes of the *val1* mutant. The LAFL TFs are part of a complex transcriptional regulatory network promoting the seed developmental program. Each LAFL TF has its own regulon with both unique and overlapping targets (H. Jia et al., 2014). *VAL1* is transcriptionally activated by *FUS3* (F. Wang & Perry, 2013) and could regulate LAFL network activity through repression of transcription of individual members of these regulons rather than the individual LAFL TFs. To test this hypothesis, we determined whether any of the 3,293 transcripts that were de-repressed in *val1* mutant embryos relative to the wild type also belonged to any of the LAFL network regulons.

This strategy relies on the availability of the corresponding regulons in developing Arabidopsis embryos or seeds. Only the ABI3 regulon has been identified in developing seeds during the middle maturation phase (Monke et al., 2012). The available FUS3 regulon was obtained from embryonic tissue culture that closely resembles developing embryos during seed maturation (Harding, Tang, Nichols, Fernandez, & Perry, 2003; F. Wang & Perry, 2013). The available LEC1 and LEC2 regulon data sets were obtained from seedlings ectopically expressing the LECs (S. A. Braybrook et al., 2006; Junker et

al., 2012). These seedlings resembled developing embryos during the seed filling stage as they accumulated seed storage compounds and transcripts associated with the embryonic program (S. A. Braybrook et al., 2006; Junker et al., 2012). With the exception of LEC1, which also showed functions unrelated to embryo development that are connected to etiolation (Junker et al., 2012), these regulons likely reliably represent regulons in developing embryos or seeds. The LEC1 and LEC2 regulons (S. A. Braybrook et al., 2006; Junker et al., 2012) contained 58 and 149 transcripts (1 and 4% of the 3,293 de-repressed transcripts, respectively) that were up-regulated in *val1* embryos compared to the wild type. These transcripts were SVs that originated from 41 and 109 genes, respectively. Only 14 transcripts (13 genes, 0.4% of the 3,293 transcripts) belonging to the ABI3 regulon (Monke et al., 2012) were de-repressed in developing *val1* embryos.

In the case of the FUS3 regulon (F. Wang & Perry, 2013), the numbers of de-repressed transcripts in the mutant were higher than for the other LAFL regulons, with 507 up-regulated transcripts (327 genes, 15% of the 3,293 transcripts). Only 49 of these transcripts were significantly de-repressed at all time points. Several of these transcripts are known members of the regulatory seed development network, including the TF CUP-SHAPED COTYLEDONS 1 (CUC1), and proteins involved in phytohormone synthesis and signaling, including GIBBERLLIN METHYLTRANSFERASE 2 (GAMT2), GIBBERELLIN-REQUIRING 1 (GA1), PICKLE-RELATED 2 (PKR2), and AUXIN RESISTANT 3 (AXR3; a target of both FUS3 and LEC2).

6.2.6 Epigenetic and transcriptional repression of target genes by VAL1

VAL1 has the potential to repress gene expression either through epigenetic means or through direct effects on transcription as it contains domains involved in (i) interactions with histones or histone modifying proteins and (ii) domains implicated in interactions with *cis* element sequences. In the case of epigenetic repression, the VAL1 PHD-L domain promotes PRC2-catalyzed H3K27me3 histone modification in seedlings (Veerappan et al., 2014). Epigenetic modifications are temporary as they are cell specific and depend on developmental and environmental cues. Ideally, embryo-specific data sets, which are not available, would be used to identify VAL1-regulated genes associated with the H3K27me3 mark. In the absence of such specific information, for our analyses, the

genome-wide H3K27me3 data from whole seedlings generated by Zhang *et al*. (2007) was used. The majority of H3K27me3-related events were associated with genes involved in the regulation of seed and flower development and also root function (X. Zhang et al., 2007). It is reasonable to speculate that some of these seed-specific genes associating with histone 3 carrying the K27me3 mark that are repressed by VAL1 during seed germination would also be associated with this mark in developing embryos. Taken together, the results of the following analyses should therefore be considered as testable predictions.

To assess whether any of the 2,483 genes that were significantly up-regulated in developing *val1* mutant embryos could be repressed through this type of chromatin modification, they were evaluated for the presence of the H3K27me3 mark as defined above. Out of the 2,483 up-regulated genes in the mutant, 588 genes (22%) contained the H3K27me3 mark, while 53 genes out of 189 genes (27%) that were significantly up-regulated throughout development contained this epigenetic mark. The FUS3 regulon had the largest number (31%) of up-regulated genes containing the H3K27me3 epigenetic mark. Regardless of which regulon they belong to, these genes are candidates for direct regulation by the full-length VAL1 protein variant that has an intact PHD-L domain. In addition, 680 transcripts (out of 709 transcripts) originating from 549 genes with this epigenetic mark were identified. Among these 680 transcripts, 504 are predicted to be repressed directly through VAL1-mediated epigenetic regulation, while 176 transcripts could be repressed indirectly by the LAFL regulators or by VAL1-related epigenetic regulation (Figure 6.4). They are encoded by genes that could be epigenetically regulated by the full-length VAL1 variant, but independently of direct activation by LAFL TFs.

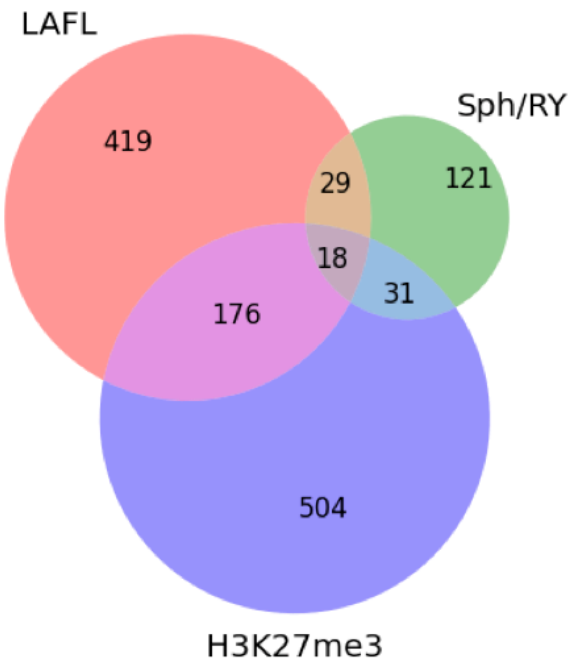

Figure 6.4 Venn diagram depicting the intersection of transcripts associated with the Sph/RY motif and H3K27me3 mark, and members of the LAFL regulatory network. Several TFs belonging to the LAFL regulatory network could be indirectly implicated in the up-regulation of transcripts in *val1* embryos through the Sph/RY motif. Exclusion of known targets of LAFL regulatory network (29 and 18) containing the Sph/RY motif resulted in 121 de-repressed transcripts representing potential direct targets of VAL1. VAL1 repressed the majority of its potential targets through epigenetic regulation.

The B3 domain in TFs such as ABI3 recognizes and binds to the Sph/RY motif in its target genes and activates expression of these genes (Monke et al., 2004). However, VAL1 is known to date only as a repressor, despite the presence of a B3 domain in its protein coding sequence. The B3 domain of a putative rice VAL1 homolog appeared to repress expression of reporter genes in an Sph/RY-specific manner (Guo et al., 2013). The possibility remains, however, that the VAL1 B3 domain could behave as an activator *in vivo*, although this has not yet been reported. A search for the presence of the Sph/RY motif was conducted in the promoter regions (1000 base pairs upstream of the start codon) of all genes that were up-regulated and down-regulated, respectively, in the absence of VAL1 using the MEME Suite (T.L. Bailey et al., 2009; Bailey et al., 2015). Nearly 50% of those genes (100 genes) that were up-regulated at all time points contained at least one Sph/RY motif in their promoters as well as the H3K27me3 mark. In the population of 3,293 transcripts that were up-regulated in the *val1* mutant at at least four time points, 199 originated from genes containing the Sph/RY motif.

6.2.7 Identification of candidate VAL1 B3-domain-specific regulons

Based on the large number of differentially expressed genes that were observed in developing *val1* embryos, VAL1 directly or indirectly affects expression of genes in developing Arabidopsis embryos on a global scale. Several members of the LAFL network and other TFs involved in regulating seed development contain the B3 domain, and there is a high regulatory complexity among these TFs and their targets (Sun et al., 2013; Swaminathan et al., 2008). Rigorously distinguishing VAL1 targets from target genes of these other TFs is not possible without experimentation. Considering the lack of success in this area (Guerriero et al., 2009; Suzuki et al., 2007; Swaminathan et al., 2008; Tsukagoshi et al., 2005), bioinformatics approaches were taken to predict sets of direct targets of VAL1.

Our approach utilized a search for the Sph/RY motif in the promoters of up-regulated genes independently of the core LAFL TFs. This approach is not ideal as it will also inevitably exclude genes that are regulated directly by both VAL1 and the B3-domain-containing core LAFL TFs. Based on this search, 152 transcripts were identified that: (i) were up-regulated in *val1* mutant embryos, (ii) contained at least one Sph/RY motif in their corresponding promoters (p-value <0.001), (iii) were not part of the core LAFL TF regulons, and (iv) did (31) or did not (121) have the H3K27me3 methylation mark (Figure 6.4). Many of these potential VAL1 targets encode regulatory proteins, including the auxin signaling TF AUXIN RESISTANT 3 (AXR3), the putative repressor PICKLE-RELATED 2 (PKR2) and the embryogenesis regulating TF WUSCHEL-RELATED HOMEOBOX 2 (WOX2) (Lie, Kelsom, & Wu, 2012). In summary, among 832 transcripts that were likely repressed directly by VAL1 (not members of any LAFL regulatory network in the case of Sph/RY motif), 81% (680) were predicted to be epigenetically regulated through H3K27me3, rather than transcriptionally regulated (18%). These results suggest that VAL1 predominantly functions as an epigenetic rather than as a transcriptional repressor.

6.2.8 Metabolomes were not affected in developing *val1* embryos

To investigate the effect of the *val1* mutation on the metabolomes of developing embryos, three independent metabonomics studies were conducted. A total of 51

compounds were detected, including sugars, sugar acids, sugar alcohols, sugar phosphates, phenolics, organic acids, organic amines, and amino acids, in addition to lipid-derived fatty acids and hydrophobic proteins. A combination of GC-MS, GC coupled to flame ionization detection (FID), ultra-performance liquid chromatography (UPLC), and the hydrophobic protein assay were employed. Embryos were taken between 8 – 17 DAP for these analyses, to match the timing in the transcriptomics experiment, though it was not possible to acquire sufficient quantities of 7 DAP embryos to achieve reliable dry weight and metabolite or phytohormone measurements. The wild type and *val1* embryos were similar with respect to both timing and accumulation of all detectable metabolites.

6.2.9 Changes in the *val1* transcriptomes are not caused by alterations in phytohormone levels

Seed development involves several phase transitions that are regulated by phytohormones, in particular the ratio of ABA and active forms of gibberellins (GAs) (R. Finkelstein, 2013; R. R. Finkelstein, Gampala, & Rock, 2002; Nambara & Marion-Poll, 2005). Our results show that *VAL1* transcript levels peaked in developing embryos during the transition from early to middle maturation, suggesting that VAL1 may be involved in the regulation of this transition. This timing also includes the period during which maternal ABA levels increase dramatically to promote seed maturation and to inhibit precocious germination (Kanno et al., 2010). To investigate whether differences in levels of ABA, jasmonate (JA), or auxin could be contributing to the transcriptional changes and metabolic phenotypes observed in the *val1* mutant, the levels of selected phytohormones were analyzed by UPLC-MS/MS. These selected phytohormones included ABA, auxin, 12-oxo-phytodienoic acid (OPDA), JA, and jasmonoyl-isoleucine (JA-Ile). No differences in the levels of these phytohormones were identified between developing wild type and *val1* embryos, suggesting that the phenotypes observed in *val1* are independent of these phytohormones. However, substantial up-regulation in the expression of GA catabolic, biosynthetic, and signaling genes was observed throughout the time course. Potential changes in GA levels could alter the ABA/GA ratios and affect the seed maturation program.

6.3 Discussion

6.3.1 VAL1 can recognize target genes through two distinct mechanisms

The full-length VAL1 protein contains four domains (PHD-L, B3, CW, and EAR) that are known in other proteins and, for some of the domains in VAL1, to bind to chromatin directly or through other proteins and/or to DNA (Hoppmann et al., 2011; Kagale & Rozwadowski, 2011; Mouriz et al., 2015; R. Sanchez & Zhou, 2011; Sun et al., 2013; Swaminathan et al., 2008). The PHD-L domain in VAL1 promotes PRC2-mediated H3K27me3 histone modification, which results in epigenetic repression of transcription (Chanvivattana et al., 2004; Schubert et al., 2006; Veerappan et al., 2014; Veerappan et al., 2012). Both the VAL1 CW domain and EAR motifs were shown to function as transcriptional repressors (Guo et al., 2013; Tsukagoshi et al., 2005). However, they also act as epigenetic regulators, as the CW domain in VAL1 was also shown to bind to H3K4me3 and interact with HDA19 in seedlings (Hoppmann et al., 2011; Zhou et al., 2013) and EAR motifs function through direct DNA binding as well as through recruitment of histone modifiers (Kagale & Rozwadowski, 2011). Unlike these domains, the B3 domain binds to Sph/RY motifs and was proposed to facilitate targeting such genes for epigenetic modification (Suzuki et al., 2007). Taken together, these domains potentially provide VAL1 with both epigenetic and direct transcriptional regulatory capabilities associated with repression of gene expression.

An additional layer of complexity related to the presence or absence of these domains is suggested by our finding that, due to AS, the VAL1 protein can take on two distinct forms. While the full-length VAL1 has four functional domains, the truncated VAL1 protein variant (encoded by a coding transcript) lacks the PHD-L domain, leaving a protein that is similar to VAL3 (H. Jia et al., 2014), containing intact B3 and CW domains and the EAR motif. This protein variant still could target genes through the B3 domain by binding directly to the Sph/RY motif and it still contains the epigenetic/transcriptional CW domain and EAR motif. Only the PHD-L domain in VAL1 facilitates the deposition of the repressive H3K27me3 mark in seedlings (Veerappan et al., 2014; Veerappan et al., 2012). If this holds true in developing embryos, these two

VAL1 protein variants likely have different targets because of the PHD-L domain as well as overlapping targets through the other three domains.

6.3.2 VAL1 is a global epigenetic and transcriptional regulator acting downstream of core LAFL transcriptional regulators in developing Arabidopsis embryos

The role of VALs as major repressors of the embryo developmental program during germination and in seedlings is well established (Guerriero et al., 2009; H. Jia et al., 2014; Sharma et al., 2013; Sun et al., 2013; Suzuki et al., 2007; Veerappan et al., 2014; Veerappan et al., 2012; C. Yang et al., 2013). Taking into account that VAL1 has the potential to regulate gene expression at both the epigenetic and transcriptional level, it was not unexpected to observe extensive changes in the transcriptomes between developing wild type and *val1* mutant embryos. VAL1 appears to have an important role in regulating embryo maturation, as 3,293 transcripts were up-regulated, and 2,194 transcripts were down-regulated in the VAL1 absence, relative to the wild type. Consistent with its repressive effects on gene expression, the majority of differentially expressed transcripts were de-repressed in *val1* mutant embryos. However, a large proportion of transcripts were repressed in developing Arabidopsis embryos in the absence of VAL1, suggesting a gene activation and/or repression of repressors of target genes by VAL1 in wild type embryos. Inhibition of repressors would activate the corresponding target genes indirectly by VAL1. In the absence of VAL1, such gene activation would be diminished, resulting in a decrease in steady-state transcript levels compared to the wild type. The PHD-L and CW domains and the EAR motif are involved in epigenetic and/or transcriptional repression that can result in repression of repressors, whereas B3 is the only domain that could be involved in transcriptional activation. However, there is a possibility that the B3 domain in VAL1 acts as a repressor in an Sph/RY-specific manner (Guo et al., 2013). Our data support this hypothesis as no genes that were down-regulated in *val1* embryos had the Sph/RY motif, so they would have to be down-regulated by VAL1-mediated epigenetic repression of repressors rather than activators and by means of domains other than B3.

VAL1 and 2 were shown to negatively regulate expression of the genes encoding the core LAFL TFs in germinating seedlings (Suzuki et al., 2007). VAL1 appears to function

differently during Arabidopsis embryo development, as expression of the core LAFL network genes were not affected in *val1* embryos. These results are in agreement with the conclusions of (Sun et al., 2013) for developing seeds in middle and late maturation phases and suggest that VAL1 activity is not a consequence of direct repression of any of the four core LAFL TFs during seed development. Based on our data, VAL1 likely globally regulates other genes, many representing the targets of specific core LAFL TFs, particularly FUS3. Both *VAL1* SVs co-express with the *FUS3* transcripts, while the *ABI3* transcript shows maximum transcript levels during late maturation, which strengthens the connection between VAL1 and FUS3 regulation. In addition, both ABI3 and VAL1 are activated by FUS3 (F. Wang & Perry, 2013), and, based on our results, VAL1 appears to repress FUS3 targets in particular, and to some extent those of LEC1 and 2, but not ABI3 targets.

For example, transcripts of the FUS3 target *WOX2* were found to be elevated in *val1* embryos relative to the wild type. The *WOX2* gene contains both the H3K27me3 mark and the Sph/RY-binding motif and represents a candidate target that could be directly repressed by VAL1. WOX2, in turn, would activate the *CUC1* gene, which does not contain the trimethylation mark or the Sph/RY motif. Both CUC1 and WOX2 are TFs involved in the formation of apical-basal axis and cotyledons in the embryo (Breuninger, Rikirsch, Hermann, Ueda, & Laux, 2008; Ikeda, Banno, Niu, Howell, & Chua, 2006; Kwon et al., 2006; Lie et al., 2012; M. Ueda, Zhang, & Laux, 2011) and their expression is needed during early embryogenesis. *WOX2* transcript levels decreased during early embryo maturation as *VAL1* transcript levels increased and remained completely repressed during middle maturation when VAL1 was present. It is reasonable to hypothesize that VAL1 would counteract FUS3-mediated activation of *WOX2* transcription to lower its transcript levels when WOX2 is no longer needed in fully developed embryos.

Embryo development depends heavily on the activity of the core LAFL TFs. Rather than interfering with expression of these TFs in developing embryos, VAL1 could repress a subset of FUS3 and perhaps LEC1 and 2 target genes at specific developmental stages at a time when the LAFL TFs responsible for originally activating these genes are still very much expressed and enabling active transcription. Through this mechanism, genes

whose action is no longer required could be effectively repressed without interfering with overall LAFL expression by repressing specific LAFL targets. Alternatively, VAL1 could provide fine tuning of gene expression for genes that are still required, but their expression needs to be modulated. Interactions similar to FUS3 and VAL1 were observed between FUS3 and the repressor TTG1. TTG1 suppresses the expression of seed storage compound genes by repressing the *ABI3* and *LEC2* genes during seed development (Mihaela Pertea et al., 2015). In contrast to VAL1, FUS3 represses, rather than activates, *TTG1* transcription to maintain the expression of the target genes. Taken together, we propose a simplified model describing the relationship between FUS3 and VAL1, including their involvement in regulating other seed-development-related genes (Figure 6.5).

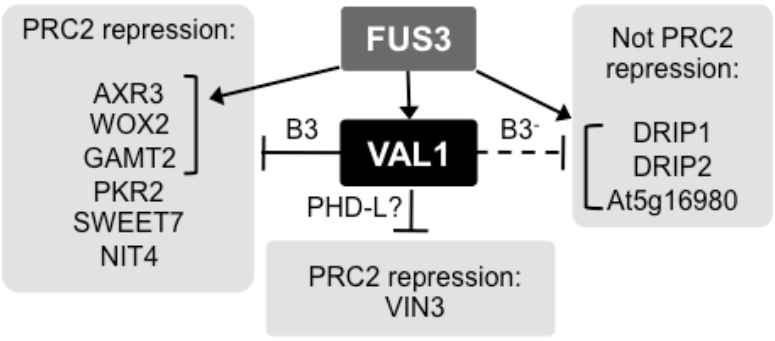

Figure 6.5 VAL1 represses genes in the FUS3 regulon through several potential mechanisms. The VAL1 B3 domain could target several genes for PRC2-mediated deposition of H3K27me3 marks, including the FUS3-regulated AXR3,WOX2, and GAMT2. In addition, the CHD3 repressor PKR2 is repressed completely in wild type embryos and upregulated in the *val1* mutant. The putative sucrose transporter SWEET7 and the nitrilase NIT4 are repressed by PRC2 and also have the RY/Sph motif in their promoters. Genes encoding the members of the FUS3 regulon, including VIN3 lack the RY/Sph motif, but are repressed through PRC2 activity. These two genes could be targeted by the VAL1 PHD-L domain. DRIP1 and 2, and At5g16980 are regulated by FUS3, but do not have the RY/Sph motif, nor they are repressed by PRC2. They could be indirectly repressed by VAL1 or targeted through a different, B3-independent (B3$^-$) mechanism. In the *val1* mutant, these genes are de-repressed to different degrees that are consistent with whether they are regulated by VAL1 alone or both VAL1 and FUS3. FUS3-activated genes are indicated by brackets.

6.3.3 VAL1 is not essential for embryo development and metabolism

With the exception of elevated protein levels in dry seeds, the physical appearance, growth, and metabolomes of developing wild type and *val1* mutant embryos were remarkably similar. These observations were unexpected considering the extensive global changes detected in the corresponding transcriptomes. There were no major changes in

transcript or in protein levels of any of the seed storage proteins in developing *val1* mutant embryos that could explain elevated dry seed protein levels. The modest increase in protein levels that was observed in dry seeds could be attributed to the endosperm and seed coat, which contribute to nearly one third of the seed dry mass in Arabidopsis and are known to accumulate seed storage compounds (Y. Li, Beisson, Pollard, & Ohlrogge, 2006). Involvement of VAL1 in regulating seed storage protein levels in endosperm and seed coat cannot be ruled out. Overall, the involvement of VAL1 in suppressing the accumulation of seed storage proteins is not entirely unexpected, as seed storage protein levels and the corresponding transcripts were found to be elevated in *val1*/*val2* mutant seedlings (Suzuki et al., 2007; Tsukagoshi et al., 2007).

6.4 Experimental procedures[1]

6.4.1 Chemicals

All reagents and metabolite standards were of analytical or higher purity and purchased from Sigma-Aldrich (St. Louis, MO) or Thermo Fisher Scientific (Waltham, MA). Solvents used for embryo work, extracting metabolites, and LC-MS/MS were MS grade, while solvents used for UPLC, GC, and other applications were LC or GC grade. Multiply labeled internal standards were obtained from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA).

6.4.2 Plant growth, silique and seed harvesting, and embryo dissections

Seeds were cold stratified (4 ^{o}C) for 3 – 5 days in 0.1% agarose and sown onto moist Farfard superfine germinating mix (Sun Gro Horticulture, Vancouver, British Columbia, Canada) in 2-inch pots. Plants were grown in a controlled environment at ~115 μmol s^{-1} m^{-2}, 20/18 ^{o}C day/night, 65% humidity, and a 16-hour photoperiod. Watering was done with deionized water every 3 – 4 days until the plants were 3 – 4 weeks old. After that, the plants received Miracle Gro All-Purpose Plant Food (1 tsp per 2 gallons) with every watering. For storage compound and metabolite analyses in dry seeds, a newly opened flower on the main stem was marked (0 DAP), and, after 21 days, the four siliques above

[1] Andrew Schneider and Eva Collakova performed the experiments in this chapter.

and below this marked silique were collected to retrieve sufficient amount of material for dry seed analyses.

All three time course experiments (transcriptomics, metabolomics, and phytohormone level analysis) were set up in the same manner. To ensure similar nutrient allocation and light exposure to the developing seeds, the time course was set up in stages: stage 1 (7 and 8 DAP), stage 2 (10, 12, and 13 DAP), and stage 3 (15 and 17 DAP), such that embryos belonging to the same stage were harvested on the same day and individual stages within days apart. Flowers were marked, starting with the latest time point within each stage. Each set of stages represented a biological replicate. Seeds and embryos from the marked siliques were dissected at low temperatures to slow down any metabolic processes. This was achieved by performing these dissections on wet glass-fiber filters placed on Petri dishes completely filled with frozen water and ensuring that temperatures were maintained between -5 and 5 ^{o}C by frequent measurements with an infrared thermometer. We found that these temperatures are the lowest that can be used without freezing the seeds, which would cause cell damage that would be problematic for the subsequent embryo washing steps.

For different applications, slightly different approaches in embryo harvesting were taken. For example, transcriptomics did not require DW determination, and embryos could be harvested 1–2 minutes faster than for metabolomics. Depending on the developmental stage of the embryos and application, embryo dissection from a single silique (40–60 seeds) took between 10 and 25 minutes. We acknowledge that these are long times for metabolomics purposes, as metabolite levels can change within minutes (Collakova et al., 2008; Fiehn, 2002; Ma, Jazmin, Young, & Allen, 2014). However, care was taken to perform dissections at low temperatures, and wild type and *val1* mutant embryos were dissected in parallel by two persons. In addition, the metabolomics experiment was performed three times and all conclusions are based on comparing wild type and *val1* mutant embryos. Therefore, it is reasonable to assume that any changes to the metabolomes due to sample preparations would be to the same degree in both wild type and *val1* mutant embryos. It is noteworthy that 6,949 wild type and 7,023 *val1* embryos were collected for the entire study. The following text provides descriptions of

two different approaches undertaken to dissect sufficient amounts of Arabidopsis embryos for transcriptomics and metabolomics/ phytohormone analyses.

For transcriptomics, embryos were rinsed off from the glass-fiber filter into ice-cold MS-grade water, briefly centrifuged at 4 ^{o}C prior to an additional wash to remove any remaining endosperm, snap-frozen in liquid nitrogen, and stored at -80 ^{o}C until processed. For metabolomics and phytohormone analyses, the procedure was similar, except that dissected embryos were transferred with a metal probe into ice-cold MS-grade water to prevent transfer of glass fibers into the sample, which would interfere with the subsequent DW determination on freeze-dried embryos. The lyophilization was performed on a FreeZone 2.5 freeze dryer (Labconco, Kansas City, MO) for two days. The lyophilized embryos were stored at room temperature in dark in a desiccator connected to continuous vacuum source at -0.08 MPa until all samples within each experiment were collected, such that they could be extracted and analyzed at the same time.

6.4.3 Analyses of seed storage compounds in dry seeds

For dry seed studies, 1.000 mg (± 5%) of dry seed was measured on an XP-26 analytical microbalance (Mettler Toledo, Columbus, OH) and lipids and proteins were extracted and analyzed by GC-FID and a fluorescent hydrophobic protein assay as described in (E. Collakova et al., 2013; Eva Collakova et al., 2013). Briefly, lipid-derived fatty acids were analyzed as fatty acid methyl esters using heptadecanoic acid as an internal standard on an Agilent 7890A series GC-FID (Agilent Technologies, Santa Clara, CA) equipped with a 30-m DB-23 column (0.25 mm x 0.25 µm, Agilent Technologies). Proteins were solubilized according to (Hou et al., 2005) and a plate-reader-based Marker Gene Hydrophobic Protein Analysis Kit (Marker Gene Technologies, Inc., Eugene, OR) using bovine serum albumin as a standard was used according to the manufacturer's recommendations. The fluorescent compound 6-(*p*-toluidino)-2-naphthalenesulfonic acid used in this kit binds preferentially to hydrophobic regions of proteins and is not influenced by high concentrations of strong detergents needed to solubilize highly hydrophobic seed storage proteins. Protein composition was determined on 15% gels by using SDS-PAGE according to (Hou et al., 2005).

6.4.4 Metabonomics of Arabidopsis embryo development

Metabonomics (a time-course metabolomics), including data processing and analysis, was performed on Arabidopsis embryos and dry seeds as described for developing soybean embryos in (E. Collakova et al., 2013) with minor modifications. Briefly, due to their small size, it was not possible to obtain a milligram of dry material for analysis of all stages of embryo development, and, depending on the stage, between 0.3 and 1.4 mg of DW was collected for each sample. Biphasic extractions were used to separate polar and non-polar metabolites and insoluble proteins. Internal standards included (i) heptadecanoic acid for free and lipid-derived fatty acids, (ii) [U-$^{13}C_6$]-glucose for sugars, sugar alcohols, sugar acids, and sugar-phosphates, (iii) [2,2,4,4-D_4]-citrate for carboxylic and organic acids, and (iv) norvaline for amino acids and organic amines. Organic phase was analyzed for fatty acids by GC-FID as described for dry seeds. A portion of aqueous phase (5%) was used for amino acid and organic amine analysis using Waters AccQ-Tag^{TM} Ultra Kit and an H-class Acquity UPLC-FLD equipped with a 10-cm Waters AccQ-Tag^{TM} Ultra C18 (1.7 μm x 2.1 mm) column (Waters Corporation, Milford, MA) using Waters 10.2-min method for analyzing free amino acids as described in (E. Collakova et al., 2013). The remaining aqueous phase was used to analyze other polar metabolites, which were separated as trimethyl silyl derivatives on an Agilent 7890A series GC equipped with a DB-5MS-DG column (30 m length × 0.25 mm × 0.25 μm with a 10-m pre-column, Agilent Technologies) and analyzed on an Agilent 5975C series single quadrupole MS as described (E. Collakova et al., 2013). The remaining insoluble interphase contained hydrophobic proteins and cell-wall material and was used to analyze relative seed storage protein levels. We confirmed that seed storage proteins partitioned in the interphase and not in the aqueous phase.

6.4.5 Phytohormone analysis using UPLC-MS/MS[1]

Absolute levels of ABA, IAA, OPDA, JA, and JA-Ile were quantified in developing embryos (three biological replicates, six time points as for the metabonomics experiment) by UPLC coupled to electrospray ionization (ESI) tandem MS. We were unable to

[1] Arati N. Poudel and Abraham J. Koo performed experiments for this section.

analyze GA levels due to their low levels in developing embryos and rather large amounts of material (e.g., 200 mg) needed for their analyses (Varbanova et al., 2007) that could not be obtained. Phytohormone extractions and quantification were based on previously established methods (Koo et al., 2014; Muller & Munne-Bosch, 2011) with minor modifications. Briefly, phytohormones were extracted from 100 - 300 lyophilized embryos in a 300-µl volume with a mixture of methanol:isopropanol:glacial acetic acid at the 40/59/1 (v/v/v) ratios at 4 °C for 30 minutes. Multiply labeled authentic compounds were used for each phytohormone as internal standards to obtain absolute levels. Five µl of extract were injected on a UPLC BEH C18 column (1.7 µm, 2.1 x 50 mm; Waters) maintained at 40 °C and attached to an H-class Acquity UPLC system (Waters, Milford, MA). A 3-min gradient program was applied using methanol and 0.15% aqueous formic acid as mobile phases with a 0.4 ml min^{-1} flow rate. Multiple reaction monitoring (MRM) was employed to detect characteristic precursor to product ion transitions for phytohormones and their internal standards: ABA (m/z 263 > 153), d_6-ABA (269 > 159), OPDA (291 > 165), d_5-OPDA (296 > 170), JA (209 > 59), dihydro-JA (211 > 59), JA-Ile (322 > 130), and $^{13}C_6$-JA-Ile (328 > 136) using Waters Xevo TQ-S MS/MS operated at ESI negative ion mode. A second channel was set to monitor IAA (m/z 176 > 130) and d_5-IAA (181 > 135) simultaneously in a positive ion mode. Data were acquired under the control of MassLynx 4.1 software and chromatographic peaks were integrated using TargetLynx Application Manager (Waters). Absolute quantification of analytes was based on standard curves comparing analyte responses to the corresponding authentic internal standards.

6.4.6 RNA-Seq analysis

RNA was extracted from frozen embryos according to (Onate-Sanchez & Vicente-Carbajosa, 2008) and purified by using the RNeasy kit (Qiagen, Limburg, Netherlands). RNA quality and integrity analysis was performed by the Virginia Bioinformatics Institute (Virginia Tech, Blacksburg, VA). Reverse transcription, library preparation, and paired-end RNA-Seq generating 75-bp reads (HiSeq 2500 Ultra-High-Throughput sequencer, Illumina, San Diego, CA), including removing low quality reads and adapter sequence trimming, was performed by Beckman Coulter Genomics (Danvers, MA) using

established protocols. Any additional data processing was done in house. Each library was sequenced four times and samples were randomly multiplexed among eight lanes. Therefore, the abundances of a transcript in these four different sequencing runs are expected to be very close, unless there is a problem with a lane. If abundance of a transcript was significantly different in a particular sequencing run (out of four runs) from the average abundance of all the runs, it was captured and removed using a t-test (p-value < 0.05). The abundance of a transcript in a sample was calculated using the average abundance of the runs, which passed the t-test. This approach enabled retaining high quality data without sacrificing the number of biological replicates.

All subsequent steps, starting with mapping reads to the TAIR10 version of the Arabidopsis genome using Tophat2, were performed using the Tuxedo Suite Trapnell, 2009 #228;Trapnell, 2010 #403;Roberts, 2011 #404} as described previously (Aghamirzaie et al., 2013), except that StringTie (Mihaela Pertea et al., 2015) instead of Cufflinks was used for transcript assembly. On average, 95% of the reads mapped to the genome and 1.5% of the reads were multi-mapped. The only defective sample was the 17 DAP replicate 3 for *val1*, in which only 5.9% reads mapped to the genome, and it was accordingly removed from further analyses. Assembled transcripts from all samples were merged using Cuffmerge to generate a reference transcriptome GTF file (transcriptome.gtf) using Arabidopsis reference GTF as a guide, which facilitates identification of different types of transcripts using Cuffcompare class codes. StringTie-B, together with transcriptome.gtf, was used to prepare read coverage tables from aligned reads. The RNA-Seq data was deposited to the GEO database (GSE7469).

Limma was chosen for performing differential expression analysis (Ritchie et al., 2015). Limma accepts raw counts as input, but StringTie only provides raw counts for exons and introns of a transcript. Therefore, an in-house script was written to calculate raw count abundance of each transcript. These raw counts were normalized using the Voom package (Law et al., 2014), as required for Limma prior to performing differential gene expression analysis. A moderated t-statistic was used to assess differential transcript expression between the wild type and *val1* mutant embryos at each time point. Empirical Bayes method was used to shrink the probe-wise sample variances towards a common value (G. K. Smyth, 2004). An F-statistic was used to test if a transcript is differentially

expressed at any time point. The *p*-values were adjusted for multiple testing using the Benjamini and Hochberg method to control the false discovery rate (Benjamini & Hochberg, 1995). Transcripts with adjusted *p*-value < 0.05 were declared to be differentially expressed. Raw count abundance values were normalized to the library size and transcript sizes to obtain Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values.

Based on the initial Limma results, 5,320 and 3,625 transcripts were up-regulated and down-regulated, respectively, in developing *val1* embryos relative to the wild type at at least one time point. However, data mining revealed that many of these transcripts should not be considered differentially expressed due to very small changes usually associated with a single time point. To eliminate such transcripts from the results, Limma results were subsequently filtered using two criteria to identify transcripts that were differentially expressed between the mutant and the wild type embryos over the time-course with a high confidence. First, a 2-fold change cutoff was applied to identify transcripts that were significantly differentially expressed between the wild type and the *val1* mutant at at least one time point. Second, a transcript was considered as differentially expressed between wild type and *val1* mutant if it showed higher or lower transcript levels in the *val1* mutant than in the wild type at at least four time points. The latter facilitated categorizing transcripts into two mutually exclusive classes of up-regulated or down-regulated transcripts in developing *val1* embryos compared to the wild type, and it was used as the default definition of differential expression. These two filtering criteria led to the removal of many false positives and resulted in the detection of transcripts that were de-repressed significantly in the absence of VAL1. Database analyses and all further data mining (e.g., how many transcripts in the data set were part of each regulon) were performed using a combination of Excel (Microsoft) and in-house scripting.

The TAIR website was used to identify the upstream 1000 bp of the 2,483 differentially expressed genes. Two tools from the MEME suite (T.L. Bailey et al., 2009; Bailey et al., 2015) were used at their default settings to find the Sph/RY motif (CATGCATG) in these input sequences. First, Analysis of Motif Enrichment (AME) (McLeay & Bailey, 2010) was used to identify the *p*-value of Sph/RY motif enrichment

in the input sequences. Second, Find Individual Motif Occurrences (FIMO) (C. E. Grant, Bailey, & Noble, 2011) was used to identify the locations of this motif in each transcript (*p*-value < 0,001). Expression profiles of each transcript for the wild type and *val1* mutant detected in this study can be viewed at http://dumuzi.ppws.vt.edu:5000/ using transcript TCONS IDs provided in all Supplemental Tables (Schneider et al., 2015).

6.4.7 Semi-quantitative qPCR

The RNA-seq data was validated using TaqMan qPCR. Briefly, RNA was converted to cDNA using the TaqMan Reverse Transcription kit (Life Technologies, Carlsbad, CA) following the manufacturer's instructions. TaqMan reactions were performed with commercially available primers and probe kits acquired from Life Technologies. Expression profiles of the two *VAL1* SVs that were detected were validated with custom-designed primers and probes for each time point in both wild type and *val1* mutant embryos. Other selected transcripts were validated for selected time points. The expression of many commonly used reference genes (e.g., tubulin, elongation factor alpha, etc.) changed dramatically in developing embryos, including those that are seed specific (Dekkers et al., 2012). To identify suitable reference transcripts, the average expression at each time point was taken and the fold change between one time point and the next was compared across the time course. Positive and negative cutoffs of 1.15 and 0.85, respectively, were used to identify transcripts that showed no changes in their levels throughout the time course. The At2g3053 transcript had the most stable expression across the time course and was selected for transcript expression normalization purposes. Normalization of each sample was performed with the $2^{-\Delta CT}$ method (Schmittgen & Livak, 2008).

7 Toward understanding of splicing regulation through construction of co-splicing networks from transcriptomics data[1]

Aghamirzaie, D., Collakova, E., Li, S., Grene, R. Toward understanding of splicing regulation through construction of co-splicing networks from transcriptomics data. *in preparstion*.

Abstract

Almost every gene in a eukaryotic organism undergoes alternative splicing. Alternative splicing occurs as the consequence of the action of a complex machinery called the splicesome, which contains several splicing factors and splicing related proteins. RNA-Seq provides an opportunity to capture known and novel splice variants. However, studies have often taken the form of intensive investigations of specific isoforms, or splice variants (SVs), encoded by selected genes, rather than focusing on understanding the mechanisms through which SVs arise. Here, we present a computational framework to predict isoform-specific regulation by specific splicing factors (SFs) that occur during Arabidopsis embryo development. The isoform population present in developing Arabidopsis embryos are shown here to contain a multiplicity of transcripts, which vary with respect to predicted protein sequence, domain composition, and coding potential. Isoform-specific expression was used as a quantitative measure for the regulation of splicing over the time-course of seed development. *de novo* RNA-motif discovery for each SF - containing a RNA-binding domain - among its splicing regulatory regions of target pre-mRNAs was carried out on a population of co-expressing product transcripts to construct a preliminary co-splicing network. Through this multi-stage analysis of co-expression and co-splicing, a group of SFs was identified as asset of candidate players in splicing events that occur during seed development in Arabidopsis. A detailed analysis of the inferred targets of a selected group of splicing factors was carried out, revealing an unexpected potential role for the unfolded protein response in seed development. This framework can be used as a starting point to understand how splicing regulation occurs in the cell.

[1] Delasa Aghamirzaie performed all the computational data analysis and developed all the pipelines for co-splicing network constructions. Song Li advised D.A. in developing co-splicing networks and motif analyses. Ruth Grene and Eva Collakova performed biological data mining of the co-splicing networks.

7.1 Background

Many biological processes in eukaryotic organisms are regulated post-transcriptionally by alternative splicing (AS), leading to the production of more than one coding or noncoding transcript from a single locus (Carvalho, Feijão, & Duque, 2013; Hubé & Francastel, 2015; Santosh, Varshney, & Yadava, 2015). Several types of AS events can occur during pre-mRNA splicing, including exon skipping, intron retention, and the use of alternative acceptor and donor splice sites, resulting, in some cases, in transcripts with premature stop codons and altered coding potential (Reddy, Marquez, Kalyna, & Barta, 2013). As a result, AS can yield proteins with distinct numbers and types of functional domains or truncated proteins and peptides with altered biological functions, which provides a basis for splicing-related protein diversity (Aghamirzaie et al., 2015; Dubrovina et al., 2012). Because many of these peptide and protein variants are not abundant and are often too short to be captured by proteomics approaches, alternative approaches are needed to assess the influence of AS on protein diversity.

To examine the extent of AS on the production of different protein isoforms on a global scale for a model system, transcriptomes of maturing *Arabidopsis thaliana* embryos (Schneider et al (2015)) were analyzed to: (i) classify transcripts into coding and non-coding using a plant-specific support vector machine classifier CodeWise (Aghamirzaie et al., 2015) and (ii) evaluate similarities and differences in protein isoforms encoded by these transcripts. Information on protein global alignment scores, protein sequence differences, and the presence or absence of functional protein domains provided insights into potential protein diversity associated with AS in developing Arabidopsis embryos.

Over 150 regulatory components, including small ribonucleoproteins and specific SFs, collectively referred to as splicing-related proteins (SRPs), are involved in the action of the complex spliceosome machinery during precursor mRNA (pre-mRNA) splicing (Valadkhan & Jaladat, 2010). SRPs are involved in protein-RNA and/or protein-protein interactions within a spliceosome. The final splicing outcome is the result of the action of several SRPs. The specificity of splicing is brought about through the action of SFs, which bind to their target pre-mRNAs in a position-dependent and RNA-motif-specific manner, acting as exonic or intronic splicing enhancers or silencers (Mo Chen & Manley,

2009). Positions of splicing regulatory elements determine the action of the cognate splicing factor, because they affect the representation or misrepresentation of the splice site to the splicing factor, which ultimately results in inclusion or exclusion of the corresponding exon (Mo Chen & Manley, 2009). This being the case, it is plausible to assume that the production of spliced transcripts is dependent, in part, on the presence and activity of each SF required for the splicing of its corresponding pre-mRNAs and that some coordination exists between expression of an SF and the transcripts produced by that SF. Coordinated splicing (co-splicing) is defined here as a metric for the identification of the action of the spliceosome on a group of pre-mRNAs to produce a population of coordinately expressed and spliced transcripts.

To globally investigate the splicing regulation through occurrence of co-splicing events, preliminary co-splicing networks and sub-networks were constructed, using a multi-stage bioinformatics approach. Differentially expressed transcripts were obtained from developing Arabidopsis embryos (Schneider et al., 2015), and they were further mined for correlations with the expression of known splicing related proteins to produce preliminary co-expression networks. The members of these networks, in turn, were analyzed for the presence of enriched RNA-binding motifs. The motifs that were identified in this way potentially play a role in the splicing of transcripts during embryo development. Three networks were examined in detail, each of which, unexpectedly, contained transcripts encoded by genes known to participate in the unfolded protein response (UPR), revealing the possible involvement of ER-based post-transcriptional events in seed development

The methods reported here can be applied to generate predictive co-splicing networks for any temporal or other multi-point paired-end RNA-Seq data obtained from any eukaryotic organism.

7.2 Methods

7.2.1 RNA-Seq analysis pipeline and the identification of 7,960 differentially expressed transcripts

The RNA-Seq data set (GEO accession number GSE74692) used in this report comprises seven time points, with three biological and four technical replicates per time

point, representing different phases of Arabidopsis embryo development, from the onset of seed filling (7 days after pollination (DAP)) to the onset of seed desiccation (17 DAP). Read mapping, transcriptome assembly, and differential expression analyses were carried out using Tophat2 (D. Kim et al., 2013), StringTie (Mihaela Pertea et al., 2015), and Limma (Ritchie et al., 2015) as described in (Schneider et al., 2015). The Arabidopsis reference genome (TAIR10 version) was used to guide the transcriptome assembly process, yielding 41,933 known and 12,054 previously unreported expressed transcripts. Transcripts were defined as differentially expressed if their expression: (i) changed by at least 2 fold in a comparison of at least two time-points and (ii) was significantly different (p-value < 0.05 and false discovery rate (FDR) < 0.05) at at least two consecutive time points. This analysis led to the detection of 7,960 differentially expressed transcripts. This population was used for the study of SVs and their potential relationships with SRPs in co-expression and co-splicing networks. CodeWise (Aghamirzaie et al., 2015) was used to assess the coding potential of novel assembled transcripts. The set of 7,960 transcripts was normalized using z-score and Cumulative distribution function (CDF) transformation, which makes them between 0 and 1. Normalized expression data were subsequently clustered into 50 groups to identify major expression trends using a k-means algorithm available in scikit-learn package (Pedregosa et al., 2011), with the following parameters: init: 'k-means++', n_init=1000, max_iter=1000. This setting stabilizes the k-means results, as each single run takes 1000 iterations, with 1000 initial different centroid seeds. Clusters containing transcripts that were expressed during the same phase of embryo development were subsequently merged into six super-clusters.

7.2.2 Effects of AS on protein diversity

To assess the effects of AS on protein diversity, protein isoforms were compared with respect to differences in their sequences as well as their conserved domains. The canonical transcript of a gene is defined as the transcript containing all the exons in the reference genome. Three metrics were defined for this purpose: (i) peptide length ratio with respect to the canonical peptide, (ii) global alignment score, and (iii) conserved

domain category. Let us assume that gene X produces two SVs, SV1 and SV2, encoding protein isoforms X1 and X2. First, the peptide length ratio is defined as:

$$Peptide\ length\ ratio = {}^{length(X2)}\Big/_{length(X1)}$$

Second, to identify sequence differences at the amino acid level, protein isoform X2 was aligned to the canonical protein isoform (X1) using pairwise global alignment with the following parameters: gap penalty = 0 and match score = 1. This setting makes the interpretation of the AS events straightforward in the context of general protein sequence differences between two proteins. The global alignment score was divided into two groups: (i) if alignment score is equal to the length of the non-canonical transcripts, then no mismatch exists between two protein isoforms (only gaps exist), which implies that SVs differed with respect to UTRs and/or exon skipping has occurred and (ii) if alignment score was less than non-canonical protein isoform, then some mismatches exist between two protein isoforms, which implies that intron retention has occurred.

Third, to predict how AS affects the potential functionality of protein variants, the corresponding protein isoforms were compared with respect to their conserved functional domains. Batch conserved domain search (CD-Search) (Marchler-Bauer et al., 2011) with the default setting was used to identify conserved domains in protein isoforms. Non-specific hits were subsequently filtered from the CD-Search outputs, leaving only superfamily and specific hits. Protein isoforms were categorized into four groups based on differences in their domains: (i) disparate domains - presence/absence of at least one domain in one of the protein isoforms, (ii) identical domains - conserved domains are exactly the same in protein isoforms, (iii) similar domains - at least one domain was truncated in one of the isoforms, and (iv) no domains – protein isoforms did not have any domains (found predominantly in proteins of unknown function).

7.2.3 Co-splicing network construction

Identification of 146 differentially expressed SRPs and 13 SFs containing RNA-binding domains. A list of genes encoding SRPs in Arabidopsis was obtained by combining entries from the Arabidopsis Splicing-Related Genes (ASRG) database (395 SRPs) (B.-B. Wang & Brendel, 2004) and the results of a proteomics analysis performed

on isolated Arabidopsis spliceosomes (additional 89 SRPs) (Koncz, deJong, Villacorta, Szakonyi, & Koncz, 2012). An additional 13 SRPs were identified from the literature, yielding a total of 497 SRPs. The list of 497 SRPs was compared to the list of 7,960 differentially expressed transcripts to extract 146 differentially expressed SRPs, which were searched for the presence of at least one RNA-binding domain to identify 13 SFs.

Co-expression network construction. Spearman correlation analysis (p-value < 0.001, Spearman correlation coefficient > 0.95) was performed to identify transcripts whose expression patterns highly correlated with at least one of 146 differentially expressed SRPs. SRPs and their correlated transcripts were further visualized as a co-expression network in Cytoscape (Shannon et al., 2003) using an organic layout.

***De novo* RNA-binding motif discovery and co-splicing network construction.** Several steps were performed for each selected SF to construct a potential co-splicing network (Figure 7.1). First, transcripts whose expression was highly correlated with the expression of at least one of 13 SFs were identified. Second, 30-nucleotide regions (R) were extracted from upstream (minus signs) and downstream (plus signs) of exon/intron junctions (5'- and 3'-splicing sites "ss") as follows: R_1 (-31:-1 5'ss), R_2 (0:30 5'ss), R_3 (-30:0 3'ss), and R4 (1:31 3'ss) sequences for each exon in a transcript, yielding four sets of sequences for each region (SF_Ri.fasta, $1 \leq I \leq 4$) as suggested in RNAmotif tool (Cereda et al., 2014). We refer to an R region as R_i ($1 \leq I \leq 4$). Third, each SF had a separate SF_Ri.fasta containing these 30-nucleotide exonic and intronic sequences from co-expressing transcripts. *de novo* motif discovery was performed on 13 SF_Ri.fasta files using MEME (Timothy L Bailey et al., 2009) in zoops mode (Zero or One Occurrence Per Sequence) to identify consensus k-mer motif ($4 \leq k \leq 7$) in each R region separately (motif E-value < 0.01). Fourth, transcripts with the significant motif located in at least one of their R regions with p-value < 0.05 (compared with background noise) were identified. We hypothesized that if an exon or intron contains a significant motif in an Ri region then the co-expressing SF will be involved in splicing that particular transcript through binding to that RNA motif. A transcript has generally more than one exon and, therefore, a conserved motif can be present in multiple Ri regions. An edge (corresponding to an Ri) was formed between an SF and a predicted product transcript in the co-splicing network if at least one of the exons or introns in the co-expressed

transcript contained a significant conserved motif in the Ri region. Motifs were compared with published RNA binding motifs using the TOMTOM tool, which is also available in the MEME suite (Gupta, Stamatoyannopoulos, Bailey, & Noble, 2007).

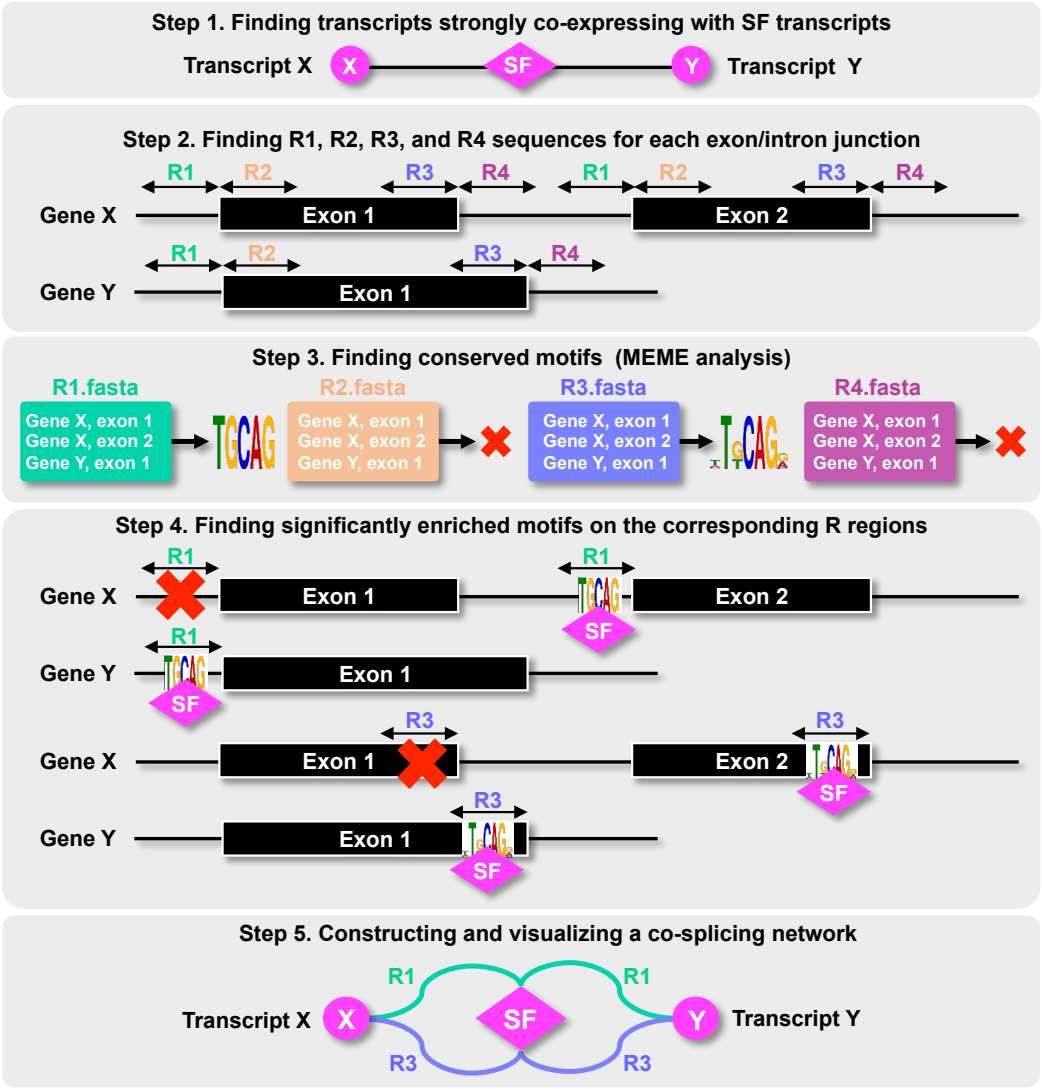

Figure 7.1 Co-splicing network construction pipeline. The following steps were performed to construct a co-splicing network: First, the sets of transcripts highly co-expressing with at least one SRP (Spearman correlation coefficient > 0.95) were used to generate a co-expression network related to 146 SRPs. Second, 30-nucleotide regions surrounding each exon/intron splice junctions (R1 through R4) were extracted for each transcript co-expressing with any of 13 selected SFs. In this example, transcripts X and Y are encoded by genes X and Y, respectively. Third, the resulting sequences were subjected to MEME analysis separately for each of 13 SF groups to find conserved motifs. Fourth, transcripts containing significantly enriched motifs at each region (p-value < 0.05) were retrieved from the MEME results. Fifth, for each transcript that had a significantly enriched motif in at least one of the R regions (e.g., R1 and R3 for co-expressing genes X and Y), the corresponding edges representing the individual R regions were constructed between each SF and its potential product using an organic layout in Cytoscape.

7.3 Results

7.3.1 Transcriptomics data

The RNA-Seq time course comprised data obtained from embryos at the following stages: (i) early maturation (7 and 8 days after pollination (DAP)), (ii) middle maturation (10, 12 and 13 DAP), and (iii) early desiccation (15 and 17 DAP) phases (Schneider et al., 2015). At the early maturation stages, Arabidopsis embryos are already fully differentiated, and, as they transition from torpedo to early bent cotyledon stage, they have already accumulated seed storage compounds (oil and protein). Embryos at the middle maturation stage show steady-state accumulation of seed storage compounds (Schneider et al., 2015) and as they start losing water during early desiccation stages, they start acquiring desiccation tolerance (S. Baud et al., 2002; Baud et al., 2008; David W Meinke, 1995). The accumulation of seed storage compounds and acquisition of desiccation tolerance prepares them for dormancy and germination. These phases of embryo development are characterized by specific metabolic, developmental, and signaling processes, only some of which are well-defined, as yet.

A data set of 7,960 transcripts was identified that showed statistically significant changes in developing Arabidopsis embryos. Transcripts were defined as differentially expressed if their expression changed by at least 2 fold between at least two time points with a significant difference (false discovery rate (FDR) < 0.05) at at least two consecutive time points. To identify trends in transcript expression, k-means clustering was performed to obtain 50 clusters. Several clusters contained transcripts that showed similar expression profiles in developing Arabidopsis embryos, and were grouped based on expression at specific phases of embryo development. This grouping yielded six large clusters, referred to here as super-clusters, containing six combinations of the defined three embryo maturation phases (Figure 7.2). These super-clusters comprised transcripts expressed at: (i) early maturation, (ii) early and middle maturation, (iii) middle maturation, (iv) middle maturation and early desiccation, (v) early desiccation, and (vi) both early maturation and desiccation tolerance phases. Color-coded grouping of transcripts into super-clusters facilitated visualization of co-expression and co-splicing networks from the temporal perspective of embryo development.

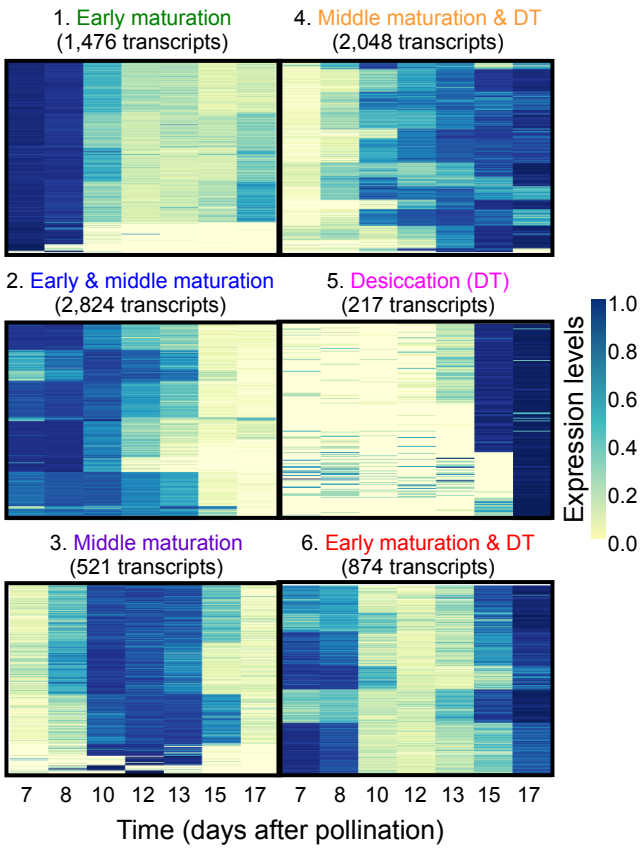

Figure 7.2 Classification of transcripts to super-clusters. The set of 7,960 differentially expressed transcripts were grouped into 50 clusters using k-means clustering. Clusters were further merged into 6 super-clusters based on the extent of transcript expression rather than the actual expression profiles during the three major developmental phases in Arabidopsis seed maturation (early and middle maturation, and early desiccation). The colors assigned to each super-cluster were used to visualize the nodes (transcripts) in all networks within this manuscript to obtain temporal information on transcript expression in developing Arabidopsis embryos.

7.3.2 Characterization of transcripts in developing Arabidopsis embryos

The population of 7,960 differentially expressed transcripts was further categorized according to their types (genic or intergenic, sense or antisense, coding or non-coding). Categorization of the novel transcripts was performed using Cuffcompare classes (Aghamirzaie et al., 2015; Aghamirzaie et al., 2013; C. Trapnell, Hendrickson, et al., 2012). CodeWise (Aghamirzaie et al., 2015) was used to assess coding potential of the transcripts. 429 ncRNAs were significantly differentially expressed during Arabidopsis embryo development. Although some transcripts were expressed only at one developmental phase, the expression of the majority of the transcripts (~73%) spanned two or more developmental phases (Table 7.1).

Table 7.1 differentially expressed transcripts (7,960) in developing Arabidopsis embryos belong to different classes (known, novel splice junction, exon skipping, antisense, and intergenic) and they can be coding or noncoding

Transcripts (Cuffcompare class)	Coding	Noncoding
Known (=)	5990	144
Novel splice junction (j)	1474	70
Exon skipping (o)	62	25
Antisense (x and s)	3	124
Intergenic	2	66
Total	7531	429

7.3.3 Alternative splicing and protein diversity in Arabidopsis embryo development

We identified genes that were alternatively spliced as the set of the genes that produced at least two splice variants so that at least one of the isoforms was significantly differentially expressed (belonged to the set of 7,960 transcripts). This analyses let to the detection of 3,008 alternatively spliced genes. Among this population, 203 genes produced both a coding and a noncoding SV. A canonical transcript was defined as an isoform that contained all the exons in the reference genome. To assess how AS affects protein diversity, each protein isoform encoded by an alternatively spliced transcript was compared with the canonical protein isoform. The peptide ratio represents how the encoded protein isoforms differ with respect to the length of the canonical isoform, as deposited in TAIR10. In cases in which an SV was predicted to be noncoding, the corresponding peptide length ratio was found to be 0.25 on average (Figure 7.3A), suggesting that a premature stop codon caused truncation of the peptide.

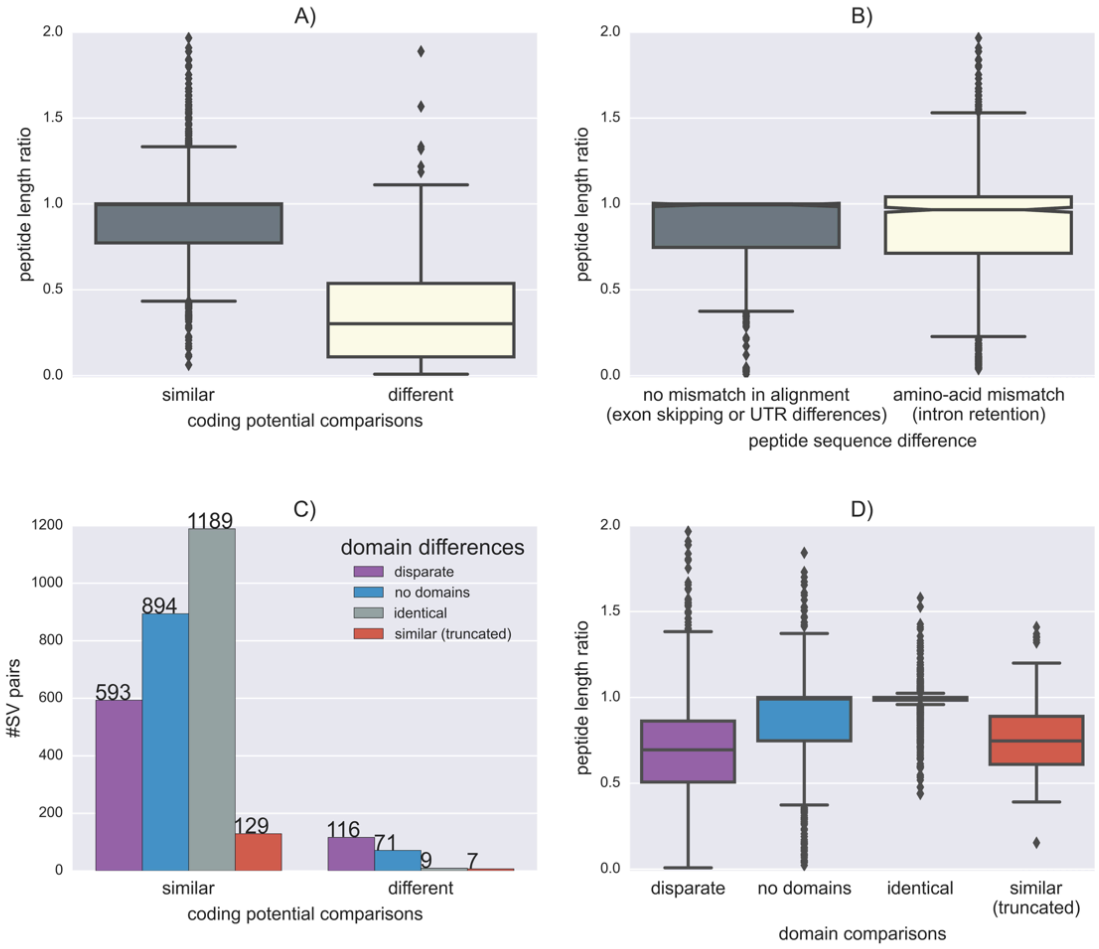

Figure 7.3 Protein diversity categorization. We identified 3,008 genes that were alternatively spliced in the set of 7,960 transcripts. These transcripts were classified either as coding or non-coding using CodeWise. The coding transcripts were translated in silico and then they were evaluated in terms of similarities or differences of amino acid sequences and functional domains. Peptide length ratio reflects how much proteins or peptides encoded by the same gene differ from the canonical protein. Peptide ratio of 1 reflects the complete match between two peptides. A) Relationship between the peptide length ratio and coding potential. B) Amino acid sequence comparisons between peptides encoded by the same genes as reflected by pairwise global alignment score. When there is only gaps and matches in the alignment (no mismatch), the peptide length ratio was close to 1. However, there were cases with new encoded amino acids compared with the canonical SV, mostly due to intron retention. C) Influence of differences in conserved domains among SV pairs on the coding potential. Alternative splicing might result in peptide isoforms with identical domains, disparate domains, or similar domains (truncated). The majority of peptide isoforms with similar coding potentials had identical and similar domains (1,189 and 129). Only 593 protein isoforms gained or lost a domain completely due to an AS event. D) Effect of peptide length differences on the presence or absence of conserved domains.

In no cases were two protein isoforms found to be exactly the same length and some sequence differences between two protein isoforms existed. Intron retention can result in a protein extension or production of a short peptide. However, in most cases, the lengths of different protein isoforms detected in the transcript population were very similar (Figure 7.3B). Forty percent of inferred protein isoforms contained identical domains

(1,198 out of 3,008 pairs), so they differed in their UTR regions, or AS did not affect the domain composition. 27% of protein isoforms either lost a domain completely or had truncated domains (Figure 7.3C). Comparing conserved domain differences with the population of protein isoforms' lengths shows that short protein isoforms had either truncated domains or had lost a domain completely, producing disparate domains (Figure 7.3D).

7.3.4 Co-expression network analysis

The set of 7,960 significantly differentially expressed transcripts contained 146 SRPs. These SRPs were distributed among all super-clusters, but the majority of SRP transcripts was expressed during early maturation and "early maturation and desiccation" phases of embryo development. This disproportion could reflect a preponderance of AS-related regulation during specific phases of embryo development. Alternatively the list of 146 SRPs is incomplete, since functions of many Arabidopsis genes remain to be determined. To identify associations between SRPs and their potential products, Spearman correlation analysis was performed on the set of 146 SRPs and the set of 7,814 remaining differentially expressed transcripts. This analysis led to the identification of 6,341 transcripts whose expression was highly correlated with at least one SRP. The resulting co-expression network is shown in Figure 7.4.

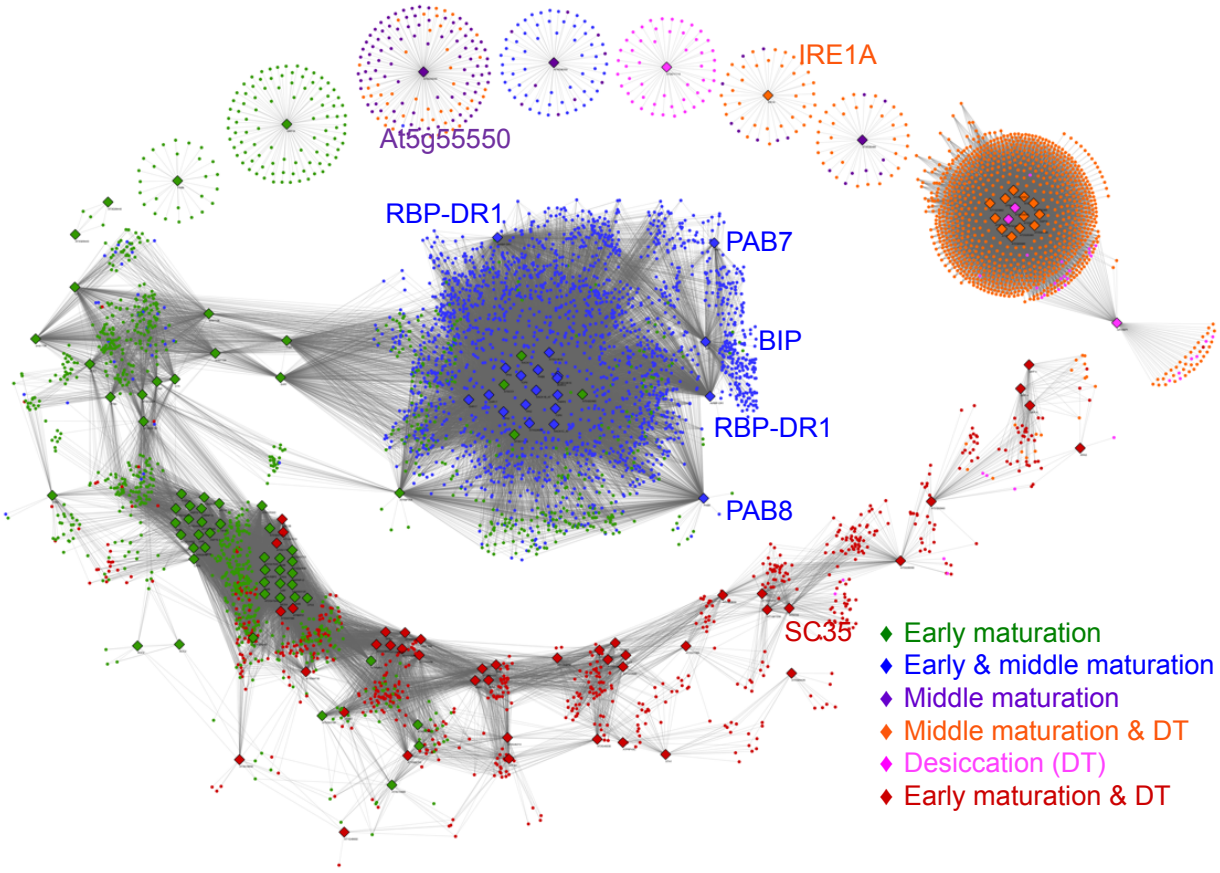

Figure 7.4 Association of differentially expressed transcripts with SRPs in a co-expression network. Spearman correlation analysis was performed between 146 SRPs and 7,960 differentially expressed transcripts. Transcripts showing temporal trends that highly correlated with each SRP (gray edges: Spearman correlation coefficient > 0.95) were extracted (6,341 transcripts) and visualized as a co-expression network in Cytoscape. SRPs are shown as diamonds and transcripts as circles. Nodes are color-coded based on super-clusters. SFs/SRPs of interest are also shown.

As is shown in this network, some transcripts co-expressed only with a single SRP, forming individual sub-networks (e.g., IRE1A or At5g55550). The majority of transcripts, however, co-expressed with more than one SRP, resulting in a highly interconnected large sub-network with visible potential associations among individual SRPs (e.g., the RBP-DR1, PAB7, and BIP sub-networks). Although most edges in this co-expression network likely reflect transcriptional guilt-by-associations among transcripts, some have the potential to reflect co-splicing relationships, or a combination of co-expression and co-splicing. To further explore possible specific splicing-related associations between SRPs and their product transcripts, co-splicing networks and sub-networks were constructed and investigated.

7.3.5 Co-splicing network inference

Because not every SRP is involved in pre-mRNA-protein interactions and the goal was to identify SRPs responsible for splicing specificity, the population of 146 SRPs was mined to identify genes encoding 13 splicing factor proteins (SFs) that contain potential RNA-binding capabilities based on experimental evidence and/or the presence of at least a single RNA-binding domain. Transcripts belonging to the sub-networks of these 13 SFs were used for subsequent motif discovery and co-splicing network construction (See Figure 7.1). The nearest neighbors (2,692 transcripts) were selected for each of these SFs from the co-expression network (Figure 7.4). *De novo* motif discovery was performed using MEME (Timothy L Bailey et al., 2009; Bailey et al., 2015) to identify consensus sequences in 30-nucleotide (R1 – 4) R-regions, near splice sites for each exon/intron junction of co-expressed transcripts. R1 and R4 are within intronic regions, while R2 and R3 are within exonic regions (Fig. 2). The notion of using R regions was adapted from the RNAmotif tool (Cereda et al., 2014), which showed that it could find the experimentally validated RNA motifs in cis-regulatory regions. The differences between the RNAmotif method and our method are as following: first, RNAmotif combines R2 and R3 regions, whereas we separated the exonic regions to be able to detect distinct RNA motifs at the 5' and 3' ends of exons. Second, we hypothesized that a SF and its potential target may be co-expressed. Third, we set the requirement that a motif should be present in at least one of the splicing regions with p-value < 0.05 to be incorporated in the final co-splicing network. Fourth, we construct co-splicing networks only for those SFs possessing a RNA-binding domain.

To construct a preliminary co-splicing network related to the population of 13 SFs, an edge was formed between transcripts co-expressing with any of the 13 SFs when at least one conserved RNA motif was present in any one of the four R regions, defined above. This preliminary network contained 2,632 transcripts connected through at least one R-related edge to a specific SF. To eliminate motifs that were not statistically significantly occurring within the R regions of exons and introns under investigation, a statistical filter (p-value < 0.05) was applied. We defined Ri ($1 \leq i \leq 4$) ratios as the proportion of all exon/intron regions containing a significantly occurring motif to the total number of exons of co-spliced transcripts. If all exon/intron Ri regions contained a significant

conserved motif, then the Ri ratio would be 1. As expected, most Ri regions did not contain conserved motifs in regions associated with all exons or introns. Overall, the R1, R2, R3, and R4 ratios were about 0.3, 0.2, 0,2, and 0.25, respectively (Figure 7.5). Therefore, less than half of the exons and introns contained an enriched motif in their Ri regions. In summary, a transcript was defined as a potential product of an SF if the expression of that transcript was highly correlated with that SF in developing embryos and a statistically significantly occurring (enriched) conserved motif existed in at least one of the R-regions (p-value < 0.05).

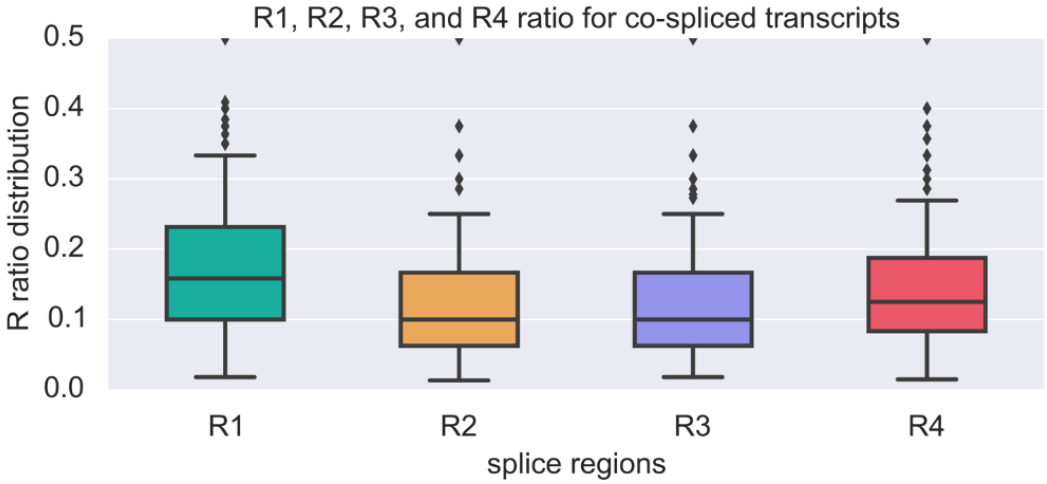

Figure 7.5 R ratio distribution in the co-splicing network. R ratio distributions reflect how frequently present conserved motifs are in exons of a given transcript (e.g., if R ratio = 1 then the motif will exist in all exons). Overall, the R ratio distributions for potential targets of 13 SFs for R1 through R4 were less that 0.15, indicating that less than 15% of the exon/intron junction regions possessed a conserved motif, which was expected after several layers of filtering based motif E-value (E-value < 0.01) and motif enrichment p-value in each region (p-value < 0.05)

We compared the sequence motifs detected in the R1 to R4 regions for each cluster, and identified several motifs that were specifically enriched in each region. For example, we found the motif AGGT enriched in more than half of the R1 regions, and motif TGCAG in most of the R4 regions. C/T rich motifs were found in the majority of R2 and R3 regions with a consensus sequence of CTTCTT. This motif is similar to an intronic RNA binding motif (M2 motif) in SR45 associated transcripts (Xing, Wang, Hamilton, Ben-Hur, & Reddy, 2015). However, the CTTCTT motif in our analysis is found in the exon regions (R2 and R3) for different splicing factors such as RSZ22a and AtRBP-DR1 suggesting a potential exonic function for this motif. We also identified motif

GAAGAAG in three R2 regions. This motif is similar to the M1 motif identified in SR45 associated transcripts. The M1 motif is enriched in exon sequences and is in agreement with our observations.

Several motifs were specifically enriched in certain R regions. One example is the motif AGGTAAG, found in the R1 region of the co-splicing cluster of IRE1A (Figure 7A). This motif is similar to the motif of DAZAP1 (consensus UAGGUAG) (Ray et al., 2013) found in human. We also identified the R2 motif AGCTGG for ATSC35, which is similar to the motif of SAMD4A in human (consensus GCTGG) (Ray et al., 2013).

7.3.6 Identification of differentially spliced transcripts and construction of a splicing-specific sub-network

The preliminary networks presented above are based on transcriptional co-expression and/or co-splicing associations between SRPs and transcripts that showed expression trends nearly identical to the trends of these SRPs. As mentioned above, it is because expression of genes is regulated at the transcriptional level. As such, it is not easy to distinguish whether the correlation of expression profiles among transcripts is due to the action of transcription factors (TFs) or SFs or a combination of both types of these regulatory mechanisms. The expression profile of a pre-mRNA is dependent on the action of specific TFs, but pre-mRNAs are transient and usually not captured by RNA-Seq data as most splicing co-occurs with transcription in the nucleus. The resulting transcripts can either show trends that are similar to those of their corresponding pre-mRNA (transcriptional), trends that correspond to the action of an SF (splicing), or a combination of the two. The only way to computationally distinguish co-splicing from transcription-related co-expression is to identify SVs that show different expression profiles. SVs are encoded by the same gene and any differences in their expression profiles can be explained by a post-transcriptional event, e.g., differential splicing when distinct SFs would bind to their motifs on pre-mRNA to facilitate splicing, in our case, at different developmental stages.

Differentially spliced transcripts were defined here as SVs of a gene encoding a pre-mRNA predicted to be spliced by distinct SFs (based on the constructed co-splicing network) belonging to different super-clusters. Fifty two genes were identified that encode differentially spliced transcripts (SVs) that had at least one SV connected to one of the 13 SFs and had at least one other either differentially expressed SV or an SV that showed stable expression during embryo development. The resulting splicing-specific network contained 11 SFs and 68 transcripts (Figure 7.6). For example, the TOE2 gene in this network has one SV (TOE2.2) associated with the SF At5g42820.2 and two other SVs (TOE2.N3 and TOE2.N4) were connected to the SFs At2g34590 and At4g14342. In contrast, Fes1A.1 and 3 co-express with the SF SC35 (expressed during early maturation and desiccation phases), while the third SVs (N4) is only expressed during the seed desiccation phase and is not associated with any of the 11 SFs within this sub-network.

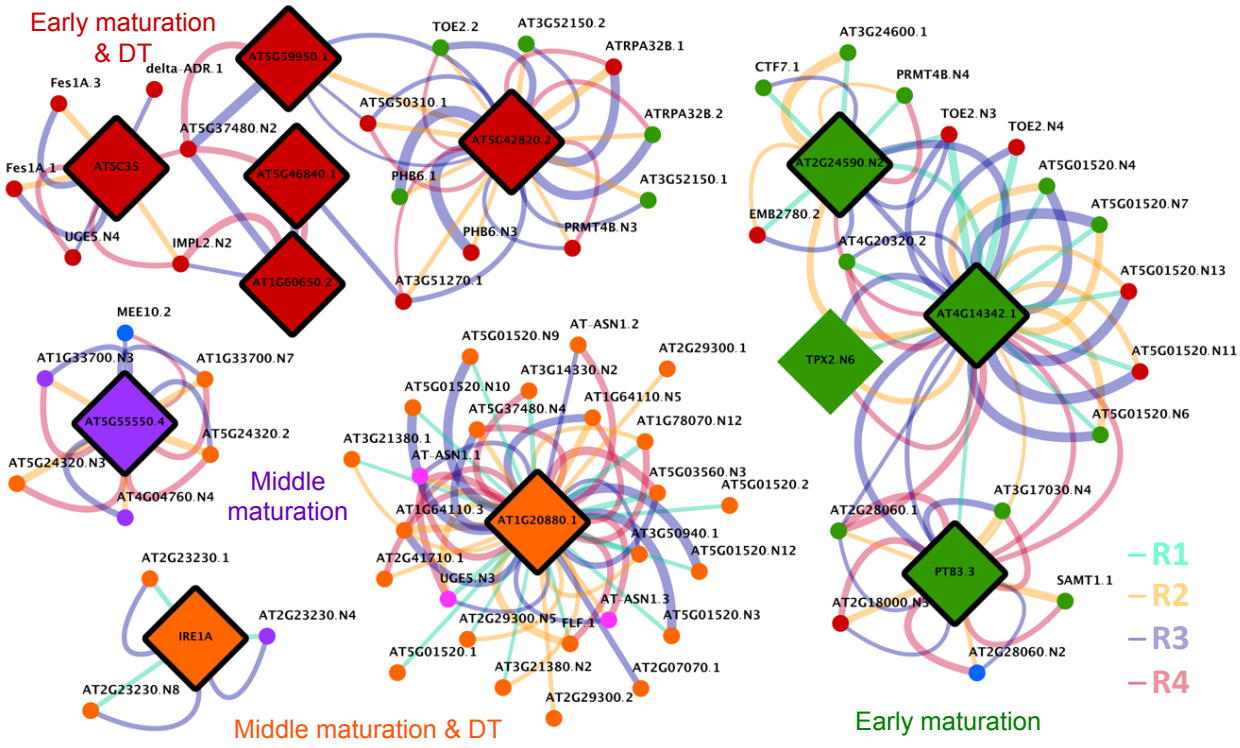

Figure 7.6 Splicing-specific sub-network of differentially expressed and spliced transcripts. This weighed sub-network shows only predicted splicing-related associations of potential product transcripts with their respective SFs based on co-splicing and the presence of at least one enriched and conserved RNA motif in at least one R region where the SF would presumably bind.

7.3.7 Inferred splicing events associated with the UPR

Endoplasmic reticulum (ER) stress occurs in plants under certain conditions (J. X. Liu & Howell, 2016), one of which may be the intense production of secretory proteins during specific developmental phases. This is manifest as the accumulation of unfolded or misfolded proteins in the ER, called the unfolded protein response (UPR). The role of the UPR is to sense ER protein-folding activities and to signal the genome to modulate, accordingly, the expression of genes affecting the protein folding machinery. In the preliminary co-splicing network presented here, the expression of transcripts encoded by unfolded protein response (UPR)-related genes was found to be correlated with IREA1A, ATRBP-DR1, or AT1G20880, each encoding a cytoplasmic RNA-binding protein. The expression of a previously unknown SV encoded by At2g17520, IRE1A, an ER-resident transmembrane protein that carries out unconventional, extra-nuclear, "minor" splicing in the cytosol (Caceres & Misteli, 2007; Parra-Rojas, Moreno, Mitina, & Orellana, 2015) was found to be significantly correlated with the expression of 44 other SVs, as shown in Figure 7.7.

Presence of a motif in several co-splicing networks determines that a motif is not specific for a particular SF. In order to examine whether motifs in UPR-related co-splicing networks (IRE1A, AtRBP-DR1, and AT1G20880) are specific or non-specific, the presence of the reported motifs were checked in the rest of the co-splicing networks using FIMO search and then were subjected to Chi-squared test. The specificity test resulted that in all the cases the motif was significantly specific (*p*-value<10e-6). Therefore, the motifs in all ER-related co-splicing networks are specific for each region.

All transcripts whose expression was correlated with that of IRE1A fell into either the middle maturation or middle maturation and DT super-clusters. The expression of two SVs of bZIP17, encoded by At2g40950, correlated with the expression of IRE1A, suggesting that bZIP17 may be a product of the splicing activity of IRE1A. bZIP17 is a membrane-tethered, ER-localized, TF which is activated by release from the ER membrane (Howell, 2013). The expression of an SV of an alpha-mannosidase, encoded by At1g30000, an enzyme that is part of a group that forms part of the glycosylation mechanism in the ER (Iwata & Koizumi, 2012), also correlated with the expression of IRE1A.

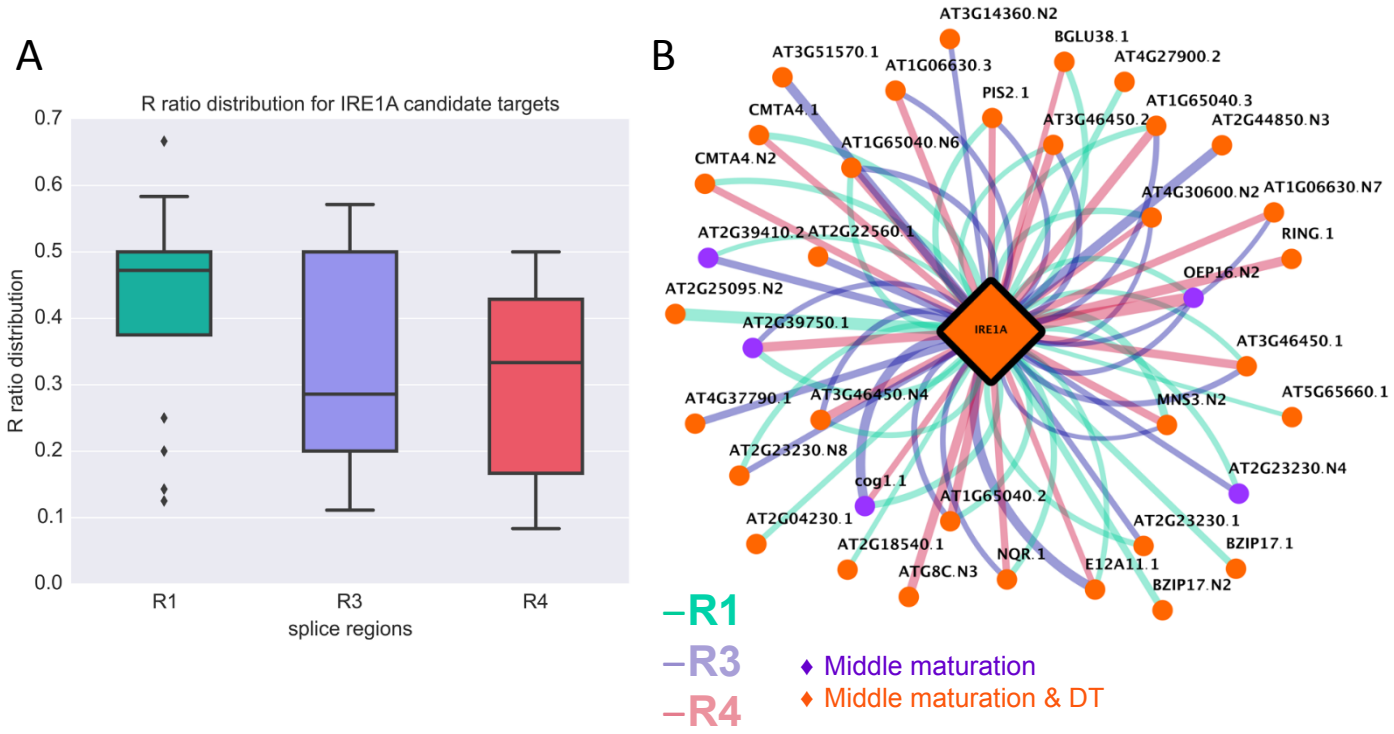

Figure 7.7 IRE1A co-splicing sub-network. A) R ratio distributions for IRE1A potential targets. R1, R3, and R4 motifs are present in more than 45%, 30%, and 32% of the exons in the targets. B A co-splicing network for IRE1A.

Although the expression of transcripts encoded by other genes that are known to be part of the UPR in plants (Howell, 2013; Song et al., 2015) were not correlated with the expression of IRE1A, many UPR-related genes were significantly expressed in developing Arabidopsis embryos, including 17 transcripts that were expressed in the same super-clusters as the group that co-expressed with IRE1A. In addition, 49 other transcripts encoded by UPR-related genes were present in the early and middle maturation super-cluster. Moreover, the expression of 16 transcripts encoded by genes associated with the UPR that were present in the early and middle maturation super-cluster was highly correlated with the expression of another SF, RBP-DR1, a cytoplasmic protein, encoded by At4g03110, which, to date has been associated only with the salicylic acid signaling pathway and responses to pathogens (Qi et al., 2010), and the regulation of flowering (H. S. Kim, Abbasi, & Choi, 2013). The ER-related transcripts in the inferred co-splicing network are shown in Figure 7.8. Among the inferred targets of RBP-DR1 that are known to be UPR-related were a BIP, (At5g42020, an HSP 70 cognate), sec61 (At5g50460), a pre-protein translocase, (At3g60540), two SVs of AtPDI5, (At1g21750), PDI6, (At1g77510), and calnexin 1 (At5g61790). BIP is known to be a global regulator

of the UPR response (Srivastava, Deng, Shah, Rao, & Howell, 2013). The expression of a third cytoplasmic, relatively unstudied, RNA-binding protein (At1g20880) was highly correlated with the expression of the UPR-related genes (At1g32560, At3g16990, and Ag1g01580), a late embryogenesis related protein, LEA, a haem-oxygenase-like protein, and ferric reduction oxidase.

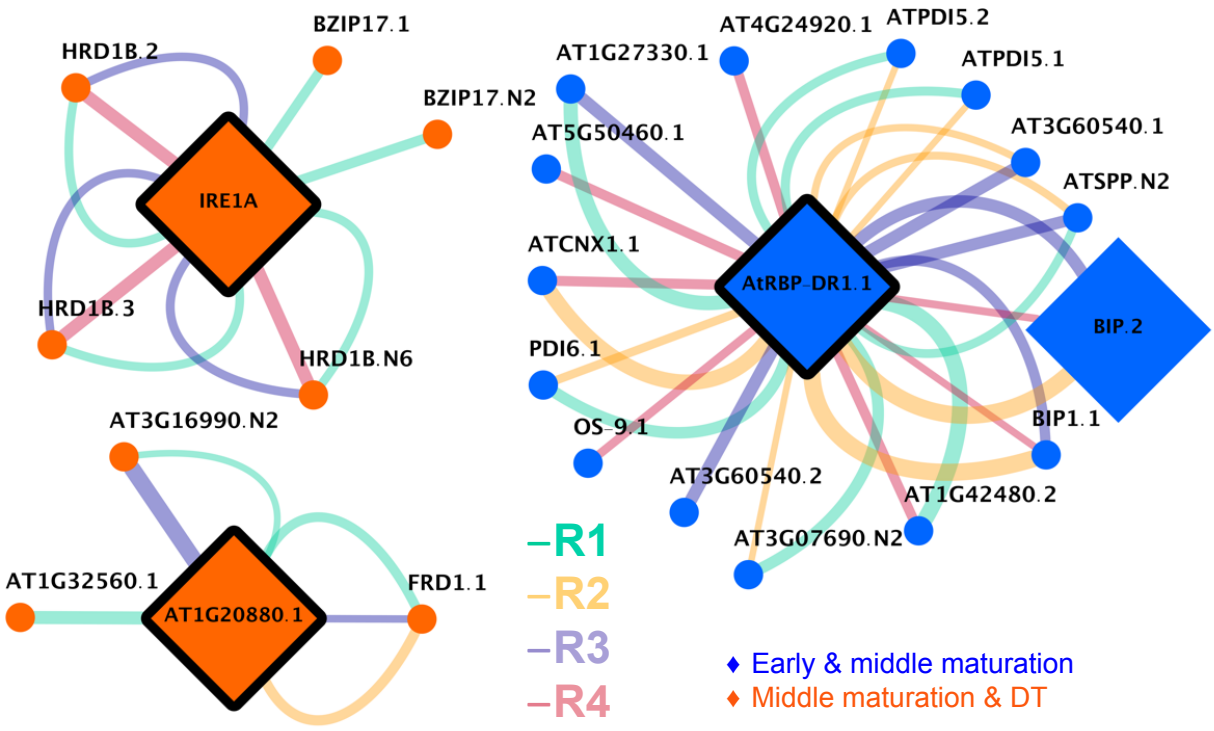

Figure 7.8 A co-splicing network for ER-related transcripts. IRE1A, AtRBP-DR1, and At1g20880 are SFs involved in ER-associated splicing and are expressed during early and middle maturation or middle maturation and desiccation phases of embryo development. Associations between these SFs and their predicted products are shown through weighed edges (thickness of the edges) representing the presence of significantly enriched conserved motifs in the corresponding R regions at the exon/intron junctions in the pre-mRNAs of these products.

However, the expression of bZIP60, encoded by At1g42990, the most well established target of IRE1A for well-studied unfolded protein responses, was not significantly correlated with the expression of IRE1A. Only the un-spliced form of bZIP60 was detected in the data set, and it was significantly expressed during embryo development at the early maturation phase. A search for common putative SF-binding motifs among the 45 SVs whose expression correlated with that of IRE1A revealed commonalities among R1, R3, and R4 groups (Figure 7.9). All detected R motifs were unique to the IRE1A group. This result suggests that novel direct targets of IRE1A may be present among this population of co-expressed and co-spliced transcripts.

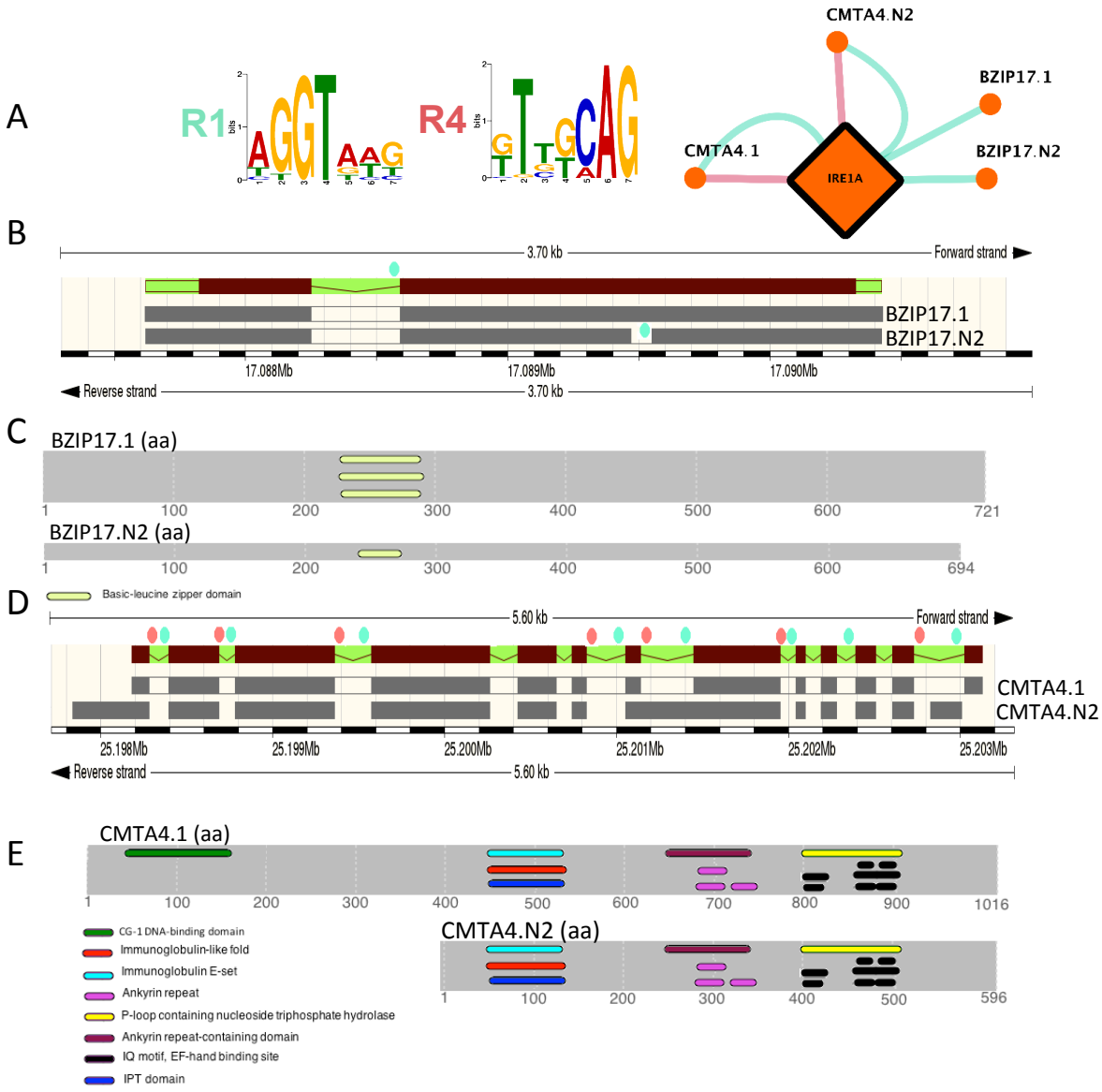

Figure 7.9 Gene structure and functional protein domains of bZIP17 and CMTA4 variants. bZIP17 and CMTA4 pre-mRNAs were identified as potential targets of IRE1A through co-expression and co-splicing network analyses. a Predicted enriched conserved motifs for IRE1A and the IRE1A co-splicing network. b bZIP17 gene/pre-mRNA structure and the location of the predicted conserved motif in the R1 intron/exon splice junctions (teal circles) on the pre-mRNA. c Protein isoforms encoded by the bZIP17.1 and bZIP17.N2 SVs with the location of basic-leucine zipper domain. The interruption of the second exon with a short cryptic intron in bZIP17.N2 results in the truncation of this domain. Interestingly, the conserved (A/N)GGT(N)3 motif is present in what would be considered the R1 region of this cryptic intron, but this sequence was not used for MEME analysis because it was not part of the conventional exon/intron gene structure reported in TAIR. d CMTA4 gene/pre-mRNA structure and the location of the predicted conserved motifs in the R1 and R4 regions flanking the exon/intron splice junctions (teal and salmon circles, respectively) on the pre-mRNA. The CMTA4.N2 SV retained two introns and is also different 5'- and 3'-UTR regions compared to CMTA4.1. e Protein isoforms encoded by the CMTA4 SVs with the location of functional domains. Protein truncation at the N terminus of CMTA4.N2 protein variant resulted in the loss of the CG-1 DNA-binding domain.

7.4 Discussion

A high-throughput RNA-Seq transcriptomics data set obtained from developing Arabidopsis embryos (Schneider et al., 2015) was used to investigate splicing regulation via prediction of associations between specific SFs and their products. The goal was to identify splicing regulatory elements as multivalent RNA binding motifs through which co-expressed splicing regulatory proteins and splicing factors are recruited to include/exclude an exon or cause intron retention. It has been shown that tissue-specific alternative splicing is controlled by splicing regulatory proteins that are differentially expressed (Mo Chen & Manley, 2009). Therefore, first the transcripts that were significantly differentially expressed were identified (7,960 transcripts in total) including 146 splicing related proteins. The population of differentially expressed transcripts included antisense, intergenic, and novel splice variants. We identified 3,008 genes (among the 7,960 transcripts) that were alternatively spliced, in which at least one splice variant was significantly differentially expressed.

Alternative splicing (AS) is known to generate protein diversity in eukaryotes. It has been previously shown, from proteomics data that AS-induced diversity at the transcriptome level can result in protein diversity (Edouard I Severing, van Dijk, & van Ham, 2011). We investigated how AS affected protein diversity in the set of 3,008 AS genes expressed during embryo development. First, the SVs were translated *in silico* to find the longest possible ORF for each transcript and then they were mined to identify differences in the protein isoforms' sequences as well as in their conserved domains. We determined that the majority of the protein isoforms detected had some differences in their respective sequences (based on pairwise alignment results), and that no two protein isoforms had exactly the same sequence. Despite differences in the protein sequences, about 40% of the protein isoform pairs (1,198 out of 3,008) contained identical domains. 23% of the protein isoforms had lost at least one domain (disparate domain category), and only 4% had truncated domains. The remaining isoforms did not have any known domains. The loss or truncation of at least one domain in 27% of the protein isoforms might cause severe effects on the function of a protein, and in some cases might make the protein nonfunctional.

7.4.1 Splicing regulation through co-splicing network inference

We further focused on an examination of the splicing regulation mechanism that occurred during the maturation and early desiccation phases of embryo development in Arabidopsis. This analysis revealed the possible existence of a novel phenomenon, which we call coordinated RNA splicing (co-splicing). Through co-splicing analysis we were able to identify groups of transcripts that were potential splicing targets of one or more splicing factors. These possible targets possessed consensus cis-regulatory elements in at least one of their splice junctions, leading to their inferred recognition by a splicing factor. The idea of co-splicing originated by analogy with the well-studied phenomenon of coordinated gene expression (co-expression). In the case of co-expression, expression of a set of genes that show similar expression profiles across multiple treatments, cell types, genotypes, or developmental stages are inferred to be regulated by the same TF(s). Such coordinated regulation can be used to assign possible function to uncharacterized genes based on their guilt-by-association relationships with functionally characterized genes and their respective roles in specific biological processes and pathways.

Plant splicesomes have not been isolated so far and mechanisms that regulate alternative splicing are poorly understood (Reddy et al., 2013). There have been several approaches to identify splicing regulation mechanism in plants, mostly in relation to the serine/arginine-rich (SR) protein family and nuclear ribonucleoproteins. SR proteins are one of the major regulators of plant responses to stress. Most of the studies to date concern the response of AS to stress (Mancini et al., 2016; Ryu, Kim, & Park, 2015; Schlaen et al., 2015). It has been shown that the final spliced product is the result of the action of several splicing regulatory proteins (Reddy et al., 2013). However, it is also known that pre-mRNA secondary and tertiary structure can also affect splicing (Deng et al., 2011; Reddy et al., 2013).

The co-splicing network inference reported here was performed on the SRPs and their candidate targets that were significantly differentially expressed. 13 splicing factors were identified and a corresponding co-splicing regulatory network was constructed. We are well aware that focusing on only the differentially expressed splicing factors might result in the elimination of SRPs that were ubiquitously expressed. However, we were

interested in the identification of developmentally related splicing elements and SRPs. In many cases, the Ri ratio which shows how frequent a significant motif was found in a transcript (with p-value < 0.05 compared with background noise), was less than 0.4, meaning that the motif is present only in 40% of the splice junctions. Therefore, some other splicing factors are most likely involved to produce the final splicing outcome. Splicing regulation is a collaborative process involving multiple SRPs. In order to get a comprehensive view of splicing regulation, the integration of chromatin landscape, RNA structure, and RBP-RNA interactomes together with transcriptomics data is necessary (Reddy et al., 2013). Therefore, the reported co-splicing networks in this analysis are all preliminary, but they can serve as a good starting point for understanding this complex phenomenon.

7.4.2 SF-specific co-splicing networks

Among the 13 preliminary co-splicing networks, AtRBP-DR1 and IRE1A networks were investigated in detail. AtRBP-DR1 contains an RNA-binding domain and is located in the cytoplasm (Qi et al., 2010). Its action has not previously been associated with the UPR. The data reported here show a highly significant correlation between the expression of this RNA-binding protein and the expression of several UPR-related genes. Furthermore, specific RNA-binding motifs in the inferred target genes have been identified. IRE1A, on the other hand is a well-studied transmembrane protein located in the membrane of the endoplasmic reticulum which is documented to engage in unconventional (i.e. non-nuclear) splicing of precursor mRNAs present in the cytosol through the action of its C-terminal RNase domain, which faces the cytosol (Iwata & Koizumi, 2012; Walley et al., 2015). Specifically, IRE1A is known to splice bZIP60 mRNA, encoding a transcription factor that mediates the activation of some of the genes involved in the unfolded protein response (UPR) in plants (Deng, Srivastava, & Howell, 2013). Two branches of the UPR are known in plants, one involving the splicing of bZIP60 and the other the release of two membrane-bound transcription factors in the ER, bZIP17 and bZIP28 (Walley et al., 2015). Two spliced forms of bZIP17 were correlated closely with the expression of IRE1A. Only the unspliced form of bZIP60 was significantly differentially expressed over the time course of embryo development

reported here, suggesting that IRE1A did not act to splice the corresponding pre-mRNA. There is evidence that the unspliced form of bZIP60 is transcriptionally active as well as the spliced form (Henriquez‐Valencia et al., 2015).

The results reported here suggest that the SV population whose expression was highly correlated with that of IRE1A, and for which the corresponding genes possess common RNA-binding motifs, contains direct targets of the novel SV of IRE1A, which was significantly expressed during the mid maturation and DT stages of embryo development. The same is true of the large population of transcripts whose expression was highly correlated with the expression of ATRBP-DR1.

IRE1A may have as yet unknown direct targets that modulate IRE1A signaling under specific conditions (Ruberti, Kim, Stefano, & Brandizzi, 2015). To date, the UPR has been studied almost exclusively under different stress conditions, and, indeed, the phenomenon was named in this way. However, (Iwata & Koizumi, 2012) characterize UPR-like events that occur within the ER during developmental periods during which high levels of secretory proteins are being synthesized as ERQC, rather than a response to a classical stress condition. This may be the circumstance that occurs during the maturation phases of Arabidopsis seed development. It should be noted also that, under normal conditions, loss of AtIRE1 causes changes in root growth (Y. Chen & Brandizzi, 2012), suggesting that this SF is also an essential part of development, in addition to its role in ER-related stress responses. Deng et al., (2013) present evidence that the role of IRE1A includes effects on vegetative growth and reproductive development. The transcriptomic data reported here suggest strongly that some form of this process is a part of embryo development since sixty –six transcripts encoded by genes associated with ERQC were differentially expressed over the time course of embryo development, (Suppl File z), several of which were highly correlated with the expression of one or other of IRE1A or ATRBP-DR1 (see above). The splicing actions of IRE1A, ATRBP-DR1 and AT1G20880 are likely to be a part of this quality control process. Candidate novel targets of IRE1A and AtRBP-DR1 are reported here. The UPR system has not previously been associated with the regulation of seed development, nor have specific SFs, nor have specific RNA binding motifs been identified as part of that process. It is interesting to

note that the three SFs that showed high correlation with the expression of UPR-related transcripts are all cytoplasmic proteins. This suggests that the processing of this category of transcripts may have a cytoplasmic component. “Minor” splicing may be an essential part of the process of protein synthesis during seed development.

8 Expresso: a database and Web server for exploring the interaction of transcription factors and their target genes in Arabidopsis thaliana using ChIP-Seq data[1]

Aghamirzaie, D., Velmurugan, K., Wu, S., Altarawy, D., Heath, L. S., Grene, R. Expresso: a database and Web server for exploring the interaction of transcription factors and their target genes in Arabidopsis thaliana using ChIP-Seq data.

Abstract

Motivation: The increasing availability of Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data affords the possibility of discovering function of transcription factors (TFs) in gene expression regulation. Even though *in vivo* transcriptional regulation often involves the concerted action of several TFs, results from each individual ChIP-Seq data set usually represents the action of a single TF. Therefore, a database in which available ChIP-Seq data sets are curated is necessary.

Results: The Expresso database and Web server is a tool for the processing and integration of curated Arabidopsis ChIP-Seq data, which, in turn, can be linked to a user's gene expression data. Candidate target genes of TFs were identified by motif analysis on publicly available GEO ChIP-Seq data sets. Expresso currently provides three services: (1) Identification of target genes of a given TF; (2) Identification of TFs that regulate a gene of interest; and (3) Computation of correlation between the expression of genes encoding TFs and their target genes.

Availability: Expresso is freely available at http://bioinformatics.cs.vt.edu/expresso/

8.1 Introduction

Chromatin immunoprecipitation (ChIP) is a method to investigate DNA-binding sites of DNA-binding proteins, such as transcription factors (TFs) (Valouev et al., 2008). ChIP can provide genome-wide information of *in vivo* protein-DNA interactions (Kaufmann et

[1] Delasa Aghamirzaie was involved in the data collection, data processing, and web development, Karthik Velmurugan contributed in the data analysis. Shuchi Wu contributed in the data collection and biological data mining. Doaa Altarawy contributed to the web server development.

al., 2010). Therefore, it has become an important tool to assay TF-associated gene regulations (Kaufmann et al., 2010; Park, 2009; Valouev et al., 2008). In a typical ChIP experiment, first the DNA-binding protein of interest is cross-linked to its binding sites. Then the chromatin is sheared, randomly, into short fragments and the protein-DNA complexes are purified by immunoprecipitation using a specific antibody against the DNA-binding protein of interest. Finally, genome-wide profiling of protein binding sites is produced by either genome-tiling arrays (ChIP-ChIP) or next-generation sequencing technologies (ChIP-Seq) (Kaufmann et al., 2010; Valouev et al., 2008). Compared to ChIP-ChIP, ChIP-Seq provides high-resolution data with a better signal-to-noise ratio. ChIP-seq also requires less initial material and is more cost-effective than ChIP-ChIP (Ho et al., 2011; Kaufmann et al., 2010; Valouev et al., 2008). As a consequence, ChIP-Seq has displaced ChIP-ChIP rapidly and is currently the most widely used technology for studying the protein-DNA interactions (Park, 2009; Valouev et al., 2008).

In contrast to the biomedical field, the use of ChIP-Seq in plant biology is limited (Kaufmann et al., 2010). For example, the GEO database (www.ncbi.nlm.nih.gov/gds) currently contains 8,486 ChIP-Seq human data sets, but has only 200 Arabidopsis data sets. The delay in the use of ChIP-Seq technology in plant research may be due to the specific properties of plant tissue such as the presence of the cell wall and abundant secondary metabolites that affect the quality of protein-DNA complex extraction (Kaufmann et al., 2010). However, with the improvement of ChIP-Seq protocols and cost-effective next generation sequencing, an increasing number of plant scientists are choosing ChIP-Seq. The Arabidopsis ChIP-Seq data sets currently deposited in GEO DataSets are increasing with the speed of one data set per month (Kaufmann et al., 2010; Valouev et al., 2008). Recently, Immick et al (2012) used ChIP-Seq technology to identify the SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 (SOC1) that binds to floral homeotic genes and to its own locus, suggesting that SOC1 has multiple functions in controlling flower development.

Besides predicting TF's target genes, ChIP-Seq technology can also reveal conserved binding motifs of any given TF (Valouev et al., 2008; Yong Zhang et al., 2008). Many software and online tools provide this capability such as rGADEM (http://bioconductor.org/packages/devel/bioc/html/rGADEM.html), cisfinder

(http://lgsun.grc.nia.nih.gov/CisFinder/) and MEME (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) (Timothy L Bailey et al., 2009; Mercier et al., 2011; Sharov & Ko, 2009). Multiple EM for Motif Elicitation (MEME) is a widely used tool for the identification of conserved motifs (Timothy L Bailey et al., 2009). First, DNA sequence peaks with false discovery rate (FDR) less than 0.01 are identified, the mid nucleotide position is located and then 200-500 bp DNA fragments are extracted (Timothy L Bailey et al., 2009; Immink et al., 2012; Valouev et al., 2008; Yu Zhang et al., 2013). Further, these DNA fragments are processed by MEME to obtain consensus motifs. MEME uses statistical modeling techniques to calculate the probability of a nucleotide for every base position in the predicted conserved motif. The output is usually in the format of position-dependent letter-probability matrices (Timothy L Bailey et al., 2009; Xia, 2012). The putative conserved motifs require further validation by literature review or biochemical methods such as an electrophoretic mobility shift assay (EMSA), which is a common electrophoresis technique used to study protein–DNA interaction. For example, in the study of PHYTOCHROME INTERACTING bHLH FACTORS 3 (PIF3), Zhang et al (2013) used both literature review and EMSA to validate conserved binding motifs for PIF3 predicted by MEME. First, they compared their result with previously reported conserved binding motifs for PIF3. They then selected CACGTG and CACATG from the previously reported MEME output motifs. Next, they used EMSA to test the binding activity between PIF3 and two promoter *cis*-elements. Each of the promoter *cis*-elements contained one of the two putative conserved binding motifs. Finally, they confirmed that both CACGTG and CACATG are conserved binding motifs for PIF3.

Overall, ChIP-Seq is a very effective method that provides high throughput and cost-effective predictions of target genes regulated by specific TFs and their corresponding binding motifs (Kaufmann et al., 2010; Valouev et al., 2008). However, the ChIP data sets currently available for Arabidopsis are isolated, fragmentary, and they lack a uniform format. Thus a major gap exists between the capabilities of *in vitro* methods such as ChIP-Seq and the goal of understanding the complexities of transcriptional regulation. The previously mentioned study of PIF3 conserved binding motifs, is a good example of this gap (Yu Zhang et al., 2013). Therefore, a relational database in which all available ChIP-Seq data sets are curated is essential and necessary.

We report on the construction of the Expresso Database to collect and integrate Arabidopsis ChIP-Seq data, which in turn can be linked to a user provided Arabidopsis gene expression data. Expresso compiles 20 groups of selected Arabidopsis ChIP-Seq data sets from NCBI GEO DataSets. All collected ChIP-Seq data sets were re-analyzed by Expresso processing pipeline to create a coherent and unified results for bridging the gap among multiple ChIP-Seq studies and provide a consensus access to TFs, target genes, and DNA-binding motifs. In summary, instead of going though separate ChIP-Seq data sets, Expresso provides a more rapid and integrated approach for the systematic study of the action of plant TFs.

8.2 Methods

The computational analysis pipeline for Expresso is shown in Figure 8.1. It comprises a major pre-processing step before the formatted and analyzed data can be uploaded into the database. Preprocessing includes formatting peak information, finding conserved motifs using MEME and shortlisting target genes for TFs. One of the major contributions of Expresso is that a user can upload their own Arabidopsis gene expression data and get a report of all known TF present in the Expresso Database and the corresponding target genes that are present in the user uploaded data. Additionally, the report will also contain correlation analysis results between the gene expression values of the reported TF-target gene pairs.

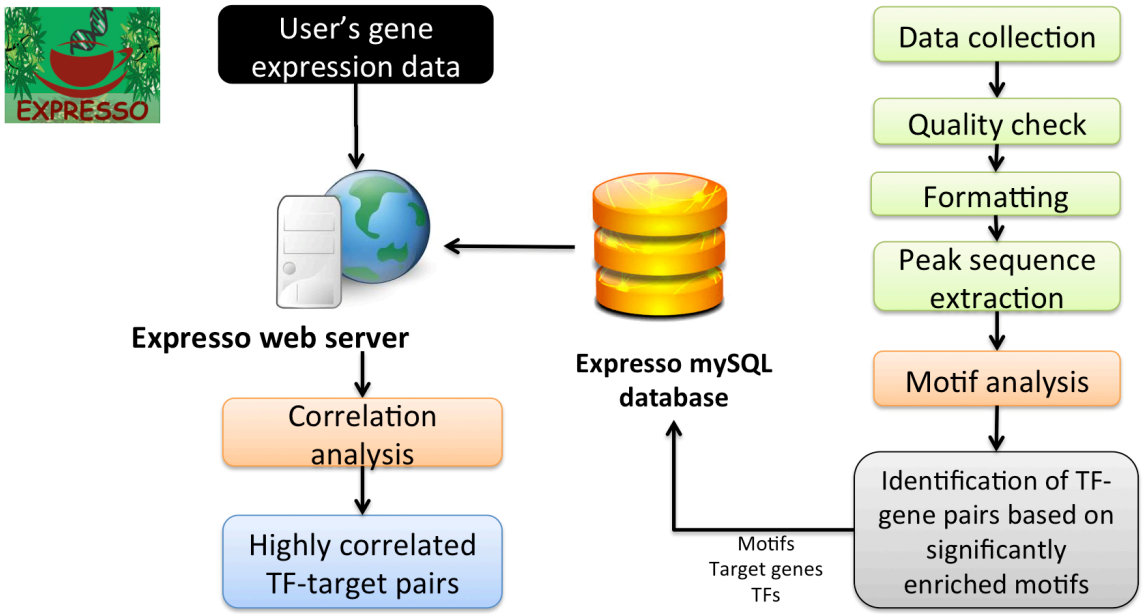

Figure 8.1 Expresso analysis pipeline

8.2.1 Data collection and formatting

We selected 20 Arabidopsis ChIP-Seq data sets from NCBI GEO Datasets (200 in total). We inspected the material and methods provided for each data set and selected 50 data sets with well-described experiment protocols that also contained specific information on plant genotypes, experimental controls, and antibodies used for immunoprecipitation. All original descriptions of these 20 ChIP-Seq experiments have been uploaded into the Expresso Database and are available on the website. We update our database on a monthly basis by adding new ChIP-Seq data sets.

Data-formatting primarily involves the extraction of a minimal set of information that is required for obtaining the peak sequence from the *Arabidopsis thaliana* genome. Of the 20 data sets, almost all were found to be in distinct formats. Therefore, custom-written python scripts were used to restructure the downloaded data into a unique format by extracting a specific set of information including: peak ID, chromosome number, peak start and end position and genes that correspond to every peak. While the peak ID, when not provided, can be manually created from the experiment, other information in the minimal set needs to be extracted from the downloaded GEO Datasets.

8.2.2 Motif finding

Given the chromosome number and peak start and end positions, the corresponding genomic sequence can be extracted. A python script was manually written for this purpose. The extracted peak sequences were organized in FASTA format with the minimal information set as the sequence header. The sequence was then trimmed based on the length of the peak summit given in the published article of each data set. While the distribution of the length of the untrimmed peak sequences of each data set varied widely, the reported peak summit lengths were usually 200 to 500 bases long upstream and downstream from the middle of the summit. For a few data sets, the summit length was not provided in the article, so the largest summit length found, 500 bases, was used.

The MEME motif discovery tool was used to extract motifs. A motif width of 5-30 bases was used for the minimum width and the maximum width parameters. The size of the motif width for every data set was decided based on the information provided in their corresponding published articles. The minimum width ranged from 6-8 bases and the maximum was found to be as high as 16 bases. Twenty was used for the number of motif parameter and was found to be sufficient after a trial run. The trial run gave eight statistically significant motifs based on the E-value provided by MEME.

The output provided by MEME contains a motif logo chart and a regular expression for each of the 20 motifs along with a comprehensive list of other information. The logo chart for regular expression [GTC][CTA]CACGTG[GA][CAG] is shown in Figure 8.2. The letters in the square brackets represent the possibilities for one nucleotide position and the letters without brackets represent the nucleotide frequency of the positions that have one dominant nucleotide letter frequency. MEME furnishes the output motif information in three different formats: text, XML and HTML. Biopython's MEME output parser was used to parse the output to get only the peak ID and *p*-value for all the peaks. Before uploading the final set of processed data (TF, target genes, *p*-value and conserved binding motifs from MEME) to the Expresso Database, the conserved binding motifs predicted by MEME were validated by literature review. If the authors of each data set have reported conserved binding motifs, we require the MEME outputs to

contain at least one of those reported motifs (Table 8.1). If none of the reported motifs were found in the MEME output, the MEME motif search was re-initiated with adjusted parameters. The adjusted parameters were usually the length of the input peak sequences and the number of motifs to be predicted.

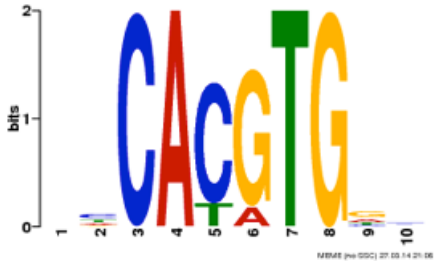

Figure 8.2 A motif logo chart for PIF3. The nucleotide position of each base pair of the motif is on the X- axis and the background letter (A, C, G, T) frequency on the Y – axis.

Table 8.1 TFs' conserved binding motifs validation

TF	Reported motifs	Expresso motifs
AGAMOU	CC[AT]6GG	Motif 5
PIF3	CAC[GA]TG	Motif 1
AP3	CC[AT]6GG and CACGTG	Motif 3 and Motif 4
PI	CC[AT]6GG	Motif 3
FLM	CC[AT]6GG	Motif 4
SOC1	CC[AT]6GG	Motif 1

8.2.3 Target gene finding

One of the essential tasks is to identify reliable candidate target genes for each TF. A set of target genes for each peak is usually available for each data set. These data sets, which may or may not have false discovery rates (FDR) assigned to them, may have a huge number of false positives. Therefore, distinguishing a false target gene/peak (false positive) from a correct target gene (true positive) is necessary.

Expresso uses two features to get high quality target genes from ChIP-seq data, as is shown in Figure 8.3. The first set is the set of target genes that was provided in the data set along with the other peak information. If the FDR value is provided in the published data, the set of peaks and target genes with FDR > 0.05 were filtered.

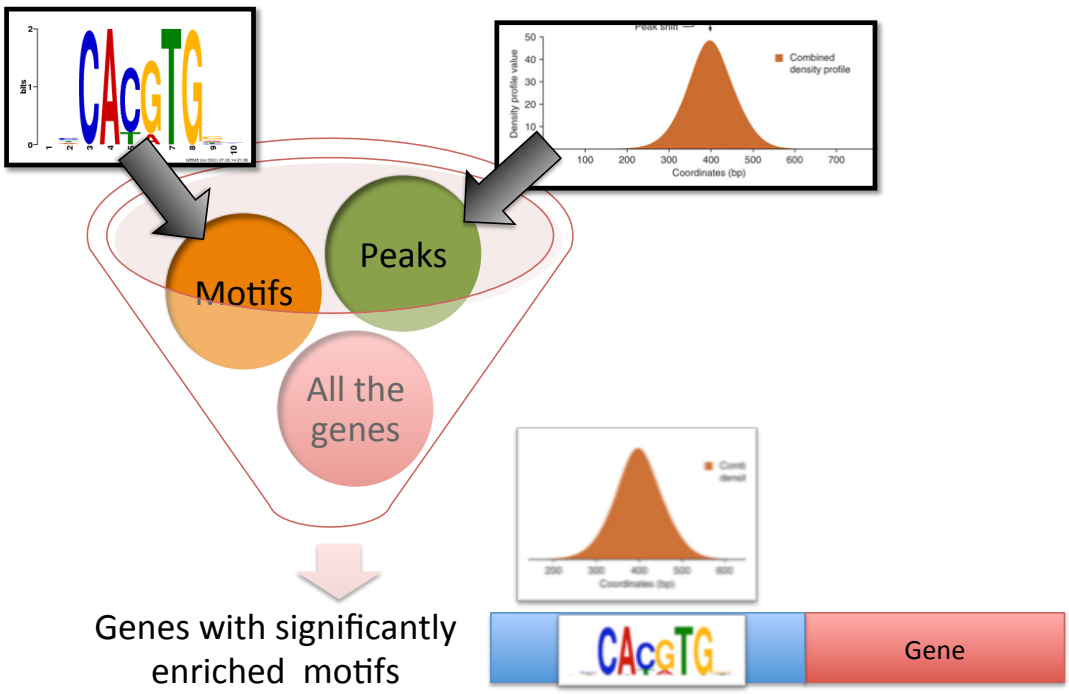

Figure 8.3 Expresso compiles two feature (motifs and peaks) and three filters to get high quality target genes for each TF.

8.2.4 Expresso database structure

Expresso is a relational mySQL database currently consisting of four main tables. The first table is the Genes table. This table consists of 33602 Arabidopsis genes in Arabidopsis Genome Initiative (AGI) format, the gene names, and a short annotation for each gene as provided by TAIR10. The AGI identifiers are used as a foreign key to other tables in the database.

The second table is the Experiments table, which keeps track of the information concerning current ChIP-seq experiments available in Expresso, such as NCBI GEO identifier, TF AGI identifier, TF name and a brief explanation about the experiment design.

The third table is the Motif table. It contains MEME output motifs for each data set. This table has several records for each motif including: number of sites, E-value, conserved sequence, experiment GEO identifier and the TF name.

The fourth table, the Motif-to-Gene table, relates genes to the conserved motifs. Each conserved motif appears in the upstream region of many genes and each gene may

have more than one motif. Therefore, this table has motif information for each gene (see AT3G19390 in Table 8.2). The Motif table and the Motif-to-Gene table are the core tables of the database, storing the relationship between TFs and target genes through motifs.

Table 8.2: An example of Motif-to-Gene table contents

id	Gene id	Motif	Experiment id	Transcription factor
1	AT3G19390	[CG]ACGTG[TG]	GSE39215	PIF3
2	AT3G19390	[GTC][CTA]CACGTG[GA][CAG]	GSE39215	PIF3

8.3 Results

The Expresso webserver has been designed in a user-friendly way to facilitate the exploration of available ChIP-seq data sets for Arabidopsis thaliana. A brief introduction to the Expresso project and its workflow is available on the Expresso Homepage (http://bioinformatics.cs.vt.edu/expresso/). All the ChIP-seq experiments in Expresso are available under the "Experiments" tab. Expresso currently provides three services for identifying: 1) the target genes of a given transcription factor, 2) the transcription factors that regulate a gene of interest and 3) the correlation of gene expression between transcription factors and their target genes.

8.3.1 Service 1: Identifying transcription factors' target genes

Users can select a transcription factor from the transcription factor drop-down menu and choose from the available list of transcription factors to view all reported target genes for a selected transcription factor. Since target genes for each transcription factor have been compiled from the peaks and motifs data, users can change their cut-off for the motif E-value. The default E-value is set to 0.05.

The information related to all target genes of a selected transcription factor includes the target gene AGI along with a link to TAIR, name of the target gene, a short

description from TAIR10 and the GEO identifier for the GEO Dataset based on which that gene was classified as a target gene for the selected transcription factor. It also shows the number of genes that have been retrieved for the selected transcription factor.

In the cases where the number of target genes is relatively high (more than 500), it may take a while for the data to be retrieved from the database and to be presented on the webpage. Therefore, the first 100 genes are shown and, in case the user is interested in viewing all target genes, the user can use the "Load more" button.

8.3.2 Service 2: Identifying transcription factors for a gene of interest

The "Genes" tab in the Expresso webserver enables the user to search for a particular gene from the database. Expresso finds all transcription factors that have the selected gene as their target gene and the resulting list of transcription factors is filtered using the motif E-value provided by the user (or the default). The transcription factor-binding site conserved motif found by MEME is reported along with corresponding transcription factors. The experiment that has resulted in the establishment of this relationship between selected target gene and transcription factor is also reported for further reference.

8.3.3 Service 3: Exploring gene expression data

The third service that Expresso provides is that the users can upload their own Arabidopsis gene expression data to look for possible Expresso known TF-Target gene pairs in the uploaded data. This analysis is performed in three consecutive steps:

1. Expresso searches for overlaps between transcriptions factors in its database and the user's gene expression data.
2. If an overlap exists, Expresso looks for further overlaps among target genes of the overlapping transcription factors (which are found in the first step) and genes present in the gene expression data.
3. Expresso computes the Pearson correlation coefficient for the expression of the transcription factor and its target genes, which have been identified in step 2. This

analysis is performed as an additional step to examine whether the expression of a transcription factor affects the expression of its target gene in the user uploaded gene expression data.

8.3.4 Correlation analysis

The correlation of the expression values between a transcription factor and a gene, over a series of samples is calculated using the Pearson Correlation Coefficient. A python script based on the interpretation of Pearson Correlation Coefficient was manually written to calculate the correlation between two given sets of expression values.

To demonstrate the correlation process, FPKM (Fragments Per Kilo Base per Million fragments mapped) values from a published RNA-Seq data (Segaran, 2007) were downloaded. About 100 genes (including some transcription factors) were selected randomly from this data sets, which has expression values for genes from different Arabidopsis tissues: leafs, seeds, roots and flowers. Fifty-four genes were found to be target genes of transcription factors in Expresso. 33% of the uploaded genes were found to be targets genes of multiple transcription factors.

Figure 8.4 shows an example to demonstrate that the correlation of gene expression between a transcription factor and its target genes can be used for inferring their relationship. Three out of four target genes of PIF3 show high correlation with the expression of PIF3, although one gene was found to have a negative correlation (R=-0.92). The fact that their expression patterns are correlated with PIF3 suggests that PIF3 plays a dominant role in regulating these three target genes. However, AT3G21700 was found to have a low correlation with PIF3, which suggests that there might be interference from other transcription factors in the regulation of AT3G21700 expression.

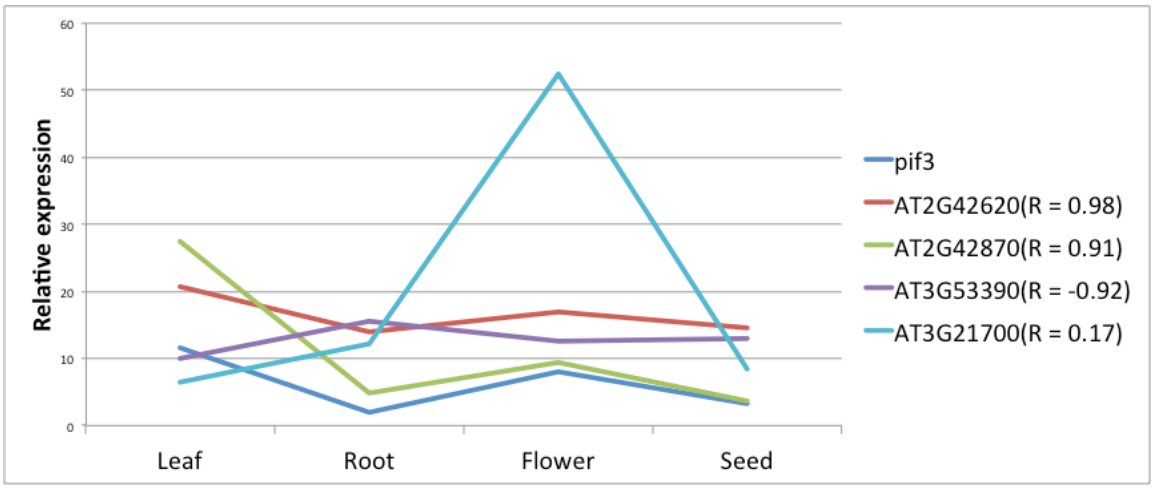

Figure 8.4 Gene expression pattern of PIF3 and its target genes in a series of Arabidopsis thaliana Col-0 tissue samples. The Pearson correlation coefficient (R) between the expression of the PIF3 gene and its target genes is shown in parenthesis.

8.4 Discussion

ChIP-Seq is a powerful technology that aides in the study of the action of transcription factors, predicting a given transcription factor's target genes and corresponding conserved binding motifs (Ho et al., 2011; Kaufmann et al., 2010; Park, 2009; Valouev et al., 2008) that has lead to the rapid increase in the number of plant TF ChIP-Seq experiments (Kaufmann et al., 2010). Expresso is designed to collect and integrate plant ChIP-Seq data. The Expresso webserver links the transcription factor and target gene pairs in a given gene expression data. Our test run shows that Expresso can precisely predict the target genes of a given transcription factor and their corresponding conserved binding motifs.

We used PIF3 as an example to test Expresso's functionality. First, we compared our MEME outputs of PIF3's conserved binding motifs with the published study by Zhang et al. The first-ranked motif we identified (CAC[GA]TG) is identical to the reported one (Yu Zhang et al., 2013). Then, we inspected PIF3's target genes identified by Expresso. Five target genes of PIF3 were randomly selected from the Expresso outputs (AT2G46970, AT2G18790, AT2G01570, AT3G23030 and AT2G43060) and were processed using the DAVID Functional Annotation tool (http://david.abcc.ncifcrf.gov/). These five genes are highly related with far red light response, which is identical to the reported function of PIF3 (Ni, Tepperman, & Quail,

1998; Yu Zhang et al., 2013). Expresso also provides precise information on the TF-binding site conserved motifs. We tested one co-target gene (AT4G27310) of PIF3 and another PIF gene family member, PIF5, to evaluate Expresso's function of identifying corresponding transcription factors. Expresso output includes both PIF3 and PIF5, which are identical to the published report. More interestingly, Expresso suggested that TIMING OF CAB EXPRESSION 1 (TOC1) also targets AT4G27310's promoter region and competes with PIF3 at the same binding site. Overall, these examples demonstrated that Expresso is a powerful tool for investigating the action of plant transcription factors.

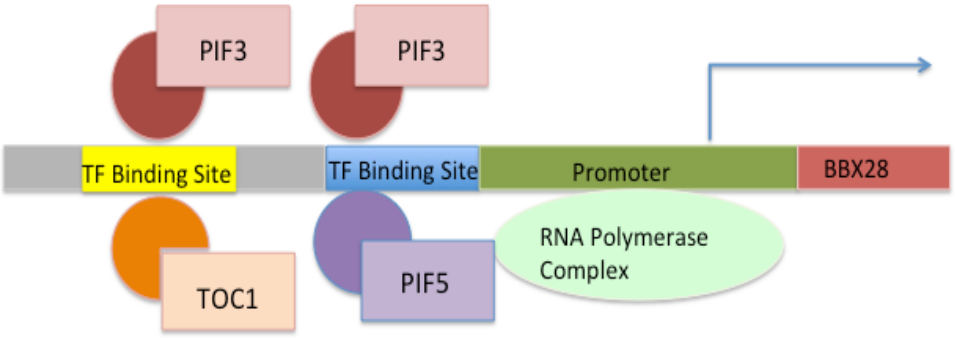

Figure 8.5 A model of the predicted co-target gene of PIF3, PIF5 and TOC1: AT4G27310.

8.5 Conclusion

The Expresso has been online for about 1.5 years and currently contains 20 Arabidopsis ChIP-Seq data sets. The database is being updated monthly. Expresso comprises a multi-stage preprocessing pipeline including data formatting, conserved motif analysis, comparison with motifs reported in the reference paper, and, finally, identifying target genes for corresponding transcription factors. One of the major services of Expresso is that users can upload their own gene expression data to the webserver, and Expresso reports all possible Expresso-known transcription factor and target gene pairs along with a Pearson correlation coefficient. In summary, Expresso provides a consolidated service to systematically study the action of transcription factors in plants.

9 Summary and outlook

9.1 Significance and contribution

RNA-Seq provides an unprecedented opportunity to capture the state of any given transcriptome. The aim of this dissertation was to mine for biologically meaningful information from large RNA-Seq data sets. In particular, my main focus was to apply in-house and developed bioinformatics tools to investigate transcriptome populations obtained from RNA-Seq analysis of developing soybean and Arabidopsis embryos. I developed several in-house scripts to identify alternative splicing events that occur during the major stages of seed development. As a consequence of my multi-stage analyses, I was able to identify several genes that were alternatively spliced and were differentially expressed during the time-course of seed development. Alternative splicing can produce noncoding splice variants or protein isoforms with incomplete domain compositions. I developed an in-house script to assess how alternative splicing affected the function of a splice variant, in particular with respect to its domain composition. I developed CodeWise for the prediction of the coding status of any given transcript. CodeWise was applied to Arabidopsis and soybean transcript populations and was able to classify known transcripts into their correct class with more than 96% accuracy. I concluded that RNA secondary structure, conserved domains, and sequence features (sequence length, open reading frame length, and UTR length) are the most informative features for accurate detection of noncoding RNAs from coding ones. Using CodeWise, I was able to detect 5,202 and 13,884 novel noncoding RNAs in Arabidopsis and soybean, respectively.

I developed a transcriptome-wide analysis framework in soybean RNA-Seq data to characterize all the detected transcripts with respect to sequence similarity, conserved domains, and coding potential. I constructed an isoform-specific co-expression network from this framework, and I showed that co-expression network analysis can be used for inferring new biological information/hypothesis generation. I identified highly connected nodes as hubs and I showed that the networks linked to the hubs are sources of discovery of possible regulatory mechanisms that occur at specific stages of seed development.

Alternative splicing leads to the production of more than one transcript from a single gene. Almost every pre-mRNA in an eukaryotic organism undergoes alternative splicing.

Misregulation of alternative splicing is associated with many diseases and cancer in the human genome. Alternative splicing occurs as the consequence of the action of a complex machinery called the splicesome, which contains several splicing factors and splicing related proteins. One of the aims of this dissertation was to identify splicing factors that create this huge diversity in Arabidopsis for a specific tissue and developmental process. I developed a multi-stage analysis pipeline, which includes the integration of co-expression and *de novo* motif discovery at splicing regulatory regions.

9.2 Future prospects

The methods that I developed in different chapters of this dissertation are applicable to any eukaryotic RNA-Seq data set. I developed several methods to identify the effect of alternative splicing on the possible function of that splice variant at the protein level. Co-expression network analysis and finding hubs can be used to find biologically meaningful information in any gene expression analysis. The co-splicing network construction using *de novo* motif discovery can be a starting point for future understanding of the complex machinery of splicing and alternative splicing. The motif analysis can be used to identify why a pre-mRNA is spliced differently. I will make all the in-house scripts for construction of co-splicing network and CodeWise available on github.

10 Appendix A: List of publications

Aghamirzaie, D., Collakova, E., Li, S., Grene, R. (2016).Toward understanding of splicing regulation through construction of co-splicing networks from transcriptomics data. *in preparation*.

Aghamirzaie, D., Velmurugan, K., Wu, S., Altarawy, D., Heath, L. S., Grene, R. (2016) Expresso: a database and Web server for exploring the interaction of transcription factors and their target genes in Arabidopsis thaliana using ChIP-Seq data. Submitted.

Aghamirzaie, D., Batra, D., Heath, L. S., Schneider, A., Grene, R., & Collakova, E. (2015). Transcriptome-wide functional characterization reveals novel relationships among differentially expressed transcripts in developing soybean embryos. *BMC Genomics, 16*(1), 928. doi:10.1186/s12864-015-2108-x

Schneider, A., **Aghamirzaie, D.**, Elmarakeby, H., Poudel, A. N., Koo, A. J., Heath, L. S., . . . Collakova, E. (2015). Potential targets of VIVIPAROUS1/ABI3‐LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos. *The Plant Journal*.

Aghamirzaie, D., Nabiyouni, M., Fang, Y., Klumas, C., Heath, L. S., Grene, R., & Collakova, E. (2013). Changes in RNA splicing in developing soybean (*Glycine max*) embryos. *Biology, 2*, 1311-1337. doi:10.3390/biology2041311

Collakova, E., **Aghamirzaie, D.**, Fang, Y., Klumas, C., Tabataba, F., Kakumanu, A., . . . Grene, R. (2013). Metabolic and transcriptional reprogramming in developing soybean (*Glycine max*) embryos. *Metabolites, 3*, 347-372. doi:10.3390/metabo3020347

11 References

Aghamirzaie, D., Batra, D., Heath, L. S., Schneider, A., Grene, R., & Collakova, E. (2015). Transcriptome-wide functional characterization reveals novel relationships among differentially expressed transcripts in developing soybean embryos. *Bmc Genomics, 16*(1), 1.

Aghamirzaie, D., Nabiyouni, M., Fang, Y., Klumas, C., Heath, L. S., Grene, R., & Collakova, E. (2013). Changes in RNA splicing in developing soybean (Glycine max) embryos. *Biology, 2*(4), 1311-1337.

Akhunov, E. D., Sehgal, S., Liang, H. Q., Wang, S. C., Akhunova, A. R., Kaur, G., . . . Gill, B. S. (2013). Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol., 161*(1), 252-265. doi:10.1104/pp.112.205161

Allen, D. K., Ohlrogge, J. B., & Shachar-Hill, Y. (2009). The role of light in soybean seed filling metabolism. *Plant J, 58*, 220-234.

Allen, D. K., Shachar-Hill, Y., & Ohlrogge, J. B. (2007). Compartment-specific labeling information in C-13 metabolic flux analysis of plants. *Phytochemistry, 68*(16-18), 2197-2210. doi:10.1016/j.phytochem.2007.04.010

Alonso, A. P., Piasecki, R. J., Wang, Y., LaClair, R. W., & Shachar-Hill, Y. (2010). Quantifying the Labeling and the Levels of Plant Cell Wall Precursors Using Ion Chromatography Tandem Mass Spectrometry. *Plant Physiol, 153*, 915-924. doi:10.1104/pp.110.155713

Amor, B. B., Wirth, S., Merchan, F., Laporte, P., d'Aubenton-Carafa, Y., Hirsch, J., . . . Deragon, J. M. (2009). Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. *Genome research, 19*(1), 57-69.

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome research, 22*(10), 2008-2017.

Andrews, S. J., & Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet., 15*(3), 193-204.

Angelovici, R., Galili, G., Fernie, A. R., & Fait, A. (2010). Seed desiccation: a bridge between maturation and germination. *Trends Plant Sci, 15*, 211-218. doi:S1360-1385(10)00006-3 [pii]

10.1016/j.tplants.2010.01.003

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., . . . Croning, M. D. R. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res., 29*(1), 37-40.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet., 25*(1), 25-29.

Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research, 38*(14), 4570-4578.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., . . . Nobel, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res., 37*, 202-208.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., . . . Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research, 37*(suppl 2), W202-W208.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res., 43*(W1), W39-49. doi:10.1093/nar/gkv416

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. E., . . . Xie, L. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res., 22*(9), 1646-1657.

Barbazuk, W. B., Fu, Y., & McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Res., 18*(9), 1381-1392. doi:10.1101/gr.053678.106

Bardou, F., Ariel, F., Simpson, C. G., Romero-Barrios, N., Laporte, P., Balzergue, S., . . . Crespi, M. (2014). Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. *Dev. Cell, 30*(2), 166-176. doi:10.1016/j.devcel.2014.06.017

Bates, P. D., Durrett, T. P., Ohlrogge, J. B., & Pollard, M. (2009). Analysis of Acyl Fluxes through Multiple Pathways of Triacylglycerol Synthesis in Developing Soybean Embryos. *Plant Physiol, 150*, 55-72. doi:10.1104/pp.109.137737

Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L., & Rochat, C. (2002). An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiology and Biochemistry, 40*(2), 151-160.

Baud, S., Boutin, J. P., Miquel, M., Lepiniec, L., & Rochat, C. (2002). An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. *Plant Physiol Biochem, 40*(2), 151-160.

Baud, S., Dubreucq, B., Miquel, M., Rochat, C., & Lepiniec, L. (2008). Storage reserve accumulation in Arabidopsis: metabolic and developmental control of seed filling. *The Arabidopsis Book*, e0113.

Bauer, S., Grossmann, S., Vingron, M., & Robinson, P. N. (2008). Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics, 24*(14), 1650-1651. doi:Doi 10.1093/Bioinformatics/Btn250

Belmonte, M. F., Kirkbride, R. C., Stone, S. L., Pelletier, J. M., Bui, A. Q., Yeung, E. C., . . . Harada, J. J. (2013). Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc. Natl. Acad. Sci. USA, 110*(5), E435-444. doi:10.1073/pnas.1222061110

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B, 57*(1), 289-300.

Bensmihen, S., Giraudat, J., & Parcy, F. (2005). Characterization of three homologous basic leucine zipper transcription factors (bZIP) of the ABI5 family during *Arabidopsis thaliana* embryo maturation. *J. Exp. Bot., 56*(412), 597-603. doi:10.1093/jxb/eri050

Bhasin, M., & Raghava, G. P. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res, 32*(Web Server issue), W414-419. doi:10.1093/nar/gkh350

Bhaskara, G. B., Nguyen, T. T., & Verslues, P. E. (2012). Unique drought resistance functions of the highly ABA-induced clade A protein phosphatase 2Cs. *Plant Physiol., 160*(1), 379-395. doi:10.1104/pp.112.202408

Blochl, A., Grenier-de March, G., Sourdioux, M., Peterbauer, T., & Richter, A. (2005). Induction of raffinose oligosaccharide biosynthesis by abscisic acid in somatic embryos of alfalfa (*Medicago sativa* L.). *Plant Sci, 168*, 1075-1082. doi:10.1016/j.plantsci.2004.12.004

Boerner, S., & McGinnis, K. M. (2012). Computational identification and functional predictions of long noncoding RNA in Zea mays. *PLoS One, 7*(8), e43047.

Bogamuwa, S., & Jang, J. C. (2013). The Arabidopsis tandem CCCH zinc finger proteins AtTZF4, 5 and 6 are involved in light-, abscisic acid- and gibberellic acid-mediated regulation of seed germination. *Plant Cell Environ., 36*(8), 1507-1519.

Borisjuk, L., Nguyen, T. H., Neuberger, T., Rutten, T., Tschiersch, H., Claus, B., . . . Rolletschek, H. (2005). Gradients of lipid storage, photosynthesis and plastid differentiation in developing soybean seeds. *New Phytol, 167*, 761-776. doi:10.1111/j.1469-8137.2005.01474.x

Braybrook, S. A., & Harada, J. J. (2008). LECs go crazy in embryo development. *Trends Plant Sci, 13*, 624-630. doi:10.1016/j.tplants.2008.09.008

Braybrook, S. A., Stone, S. L., Park, S., Bui, A. Q., Le, B. H., Fischer, R. L., . . . Harada, J. J. (2006). Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. *Proc. Natl. Acad. Sci. U.S.A., 103*(9), 3468-3473. doi:10.1073/pnas.0511331103

Breuninger, H., Rikirsch, E., Hermann, M., Ueda, M., & Laux, T. (2008). Differential expression of WOX genes mediates apical-basal axis formation in the Arabidopsis embryo. *Dev. Cell, 14*(6), 867-876. doi:10.1016/j.devcel.2008.03.008

Brummell, D. A., Chen, R. K. Y., Harris, J. C., Zhang, H. B., Hamiaux, C., Kralicek, A. V., & McKenzie, M. J. (2011). Induction of vacuolar invertase inhibitor mRNA in potato tubers contributes to cold-induced sweetening resistance and includes spliced hybrid mRNA variants. *J. Exp. Bot., 62*(10), 3519-3534. doi:10.1093/jxb/err043

Buitink, J., Leger, J. J., Guisle, I., Vu, B. L., Wuilleme, S., Lamirault, G., . . . Leprince, O. (2006). Transcriptome profiling uncovers metabolic and regulatory processes

occurring during the transition from desiccation-sensitive to desiccation-tolerant stages in *Medicago truncatula* seeds. *Plant J., 47*(5), 735-750.

Burland, T. G. (2000). DNASTAR's Lasergene sequence analysis software. *Methods Mol. Biol., 132*, 71-91.

Caceres, J. F., & Misteli, T. (2007). Division of labor: minor splicing in the cytoplasm. *Cell, 131*(4), 645-647.

Calvenzani, V., Testoni, B., Gusmaroli, G., Lorenzo, M., Gnesutta, N., Petroni, K., . . . Tonelli, C. (2012). Interactions and CCAAT-binding of *Arabidopsis thaliana* NF-Y subunits. *PLoS One, 7*(8), e42902. doi:10.1371/journal.pone.0042902

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

Carvalho, R. F., Feijão, C. V., & Duque, P. (2013). On the physiological significance of alternative splicing events in higher plants. *Protoplasma, 250*(3), 639-650.

Castillo, E. M., Delumen, B. O., Reyes, P. S., & Delumen, H. Z. (1990). Raffinose synthase and galactinol synthase in developing seeds and leaves of legumes *Journal of Agricultural and Food Chemistry, 38*, 351-355. doi:10.1021/jf00092a003

Cereda, M., Pozzoli, U., Rot, G., Juvan, P., Schweitzer, A., Clark, T., & Ule, J. (2014). RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol, 15*(1), R20.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol., 2*(3), 27.

Chanvivattana, Y., Bishopp, A., Schubert, D., Stock, C., Moon, Y. H., Sung, Z. R., & Goodrich, J. (2004). Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. *Development, 131*(21), 5263-5276. doi:10.1242/dev.01400

Chasin, L. A. (2008). Searching for splicing motifs. *Advances in experimental medicine and biology, 623*, 85.

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews Molecular cell biology, 10*(11), 741-754.

Chen, M., Zhang, B., Li, C., Kulaveerasingam, H., Chew, F. T., & Yu, H. (2015). TRANSPARENT TESTA GLABRA 1 regulates the accumulation of seed storage reserves in Arabidopsis. *Plant Physiol., 169*, 391-402. doi:10.1104/pp.15.00943

Chen, Y., & Brandizzi, F. (2012). AtIRE1A/AtIRE1B and AGB1 independently control two essential unfolded protein response pathways in Arabidopsis. *The Plant Journal, 69*(2), 266-277.

Chew, G.-L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., & Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development, 140*(13), 2828-2834.

Chia, T. Y., Pike, M. J., & Rawsthorne, S. (2005). Storage oil breakdown during embryo development of *Brassica napus* (L.). *J Exp Bot, 56*, 1285-1296. doi:eri129 [pii]

10.1093/jxb/eri129

Chipman, D. M., & Shaanan, B. (2001). The ACT domain family. *Curr. Opin. Struct. Biol., 11*(6), 694-700.

Chung, T., Wang, D. F., Kim, C. S., Yadegari, R., & Larkins, B. A. (2009). Plant SMU-1 and SMU-2 homologues regulate pre-mRNA splicing and multiple aspects of development. *Plant Physiol., 151*(3), 1498-1512. doi:10.1104/pp.109.141705

Clancy, S. (2008). RNA splicing: introns, exons and spliceosome. *Nature Education, 1*(1), 31.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., . . . Morillon, A. (2011). The reality of pervasive transcription. *PLoS Biol, 9*(7), e1000625.

Clemente, T. E., & Cahoon, E. B. (2009). Soybean Oil: Genetic Approaches for Modification of Functionality and Total Content. *Plant Physiol, 151*, 1030-1040. doi:10.1104/pp.109.146282

Collakova, E., Aghamirzaie, D., Fang, Y., Klumas, C., Tabataba, F., Kakumanu, A., . . . Grene, R. (2013). Metabolic and transcriptional reprogramming in developing soybean (*Glycine max*) embryos. *Metabolites, 3*, 347-372.

Collakova, E., Aghamirzaie, D., Fang, Y., Klumas, C., Tabataba, F., Kakumanu, A., . . . Grene, R. (2013). Metabolic and transcriptional reprogramming in developing soybean (Glycine max) embryos. *Metabolites, 3*(2), 347-372.

Collakova, E., Goyer, A., Naponelli, V., Krassovskaya, I., Gregory, J. F., Hanson, A. D., & Shachar-Hill, Y. (2008). *Arabidopsis* 10-formyl tetrahydrofolate deformylases are essential for photorespiration. *Plant Cell, 20*, 1818-1832.

Corpas, F. J., Barroso, J. B., Sandalio, L. M., Distefano, S., Palma, J. M., Lupianez, J. A., & Del Rio, L. A. (1998). A dehydrogenase-mediated recycling system of NADPH in plant peroxisomes. *Biochem J., 330*, 777-784.

Crawford, B. C., & Yanofsky, M. F. (2011). HALF FILLED promotes reproductive tract development and fertilization efficiency in *Arabidopsis thaliana*. *Development, 138*(14), 2999-3009. doi:10.1242/dev.067793

Cregg, P. J., Murphy, K., & Mardinoglu, A. (2012). Inclusion of interactions in mathematical modelling of implant assisted magnetic drug targeting. *Applied Mathematical Modelling, 36*(1), 1-34. doi:Doi 10.1016/J.Apm.2011.05.036

Dai, S., & Chen, S. (2012). Single-cell-type Proteomics: Toward a Holistic Understanding of Plant Function. *Mol Cell Proteomics, 11*, 1622-1630. doi:10.1074/mcp.R112.021550

Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C_4 plant metabolism. *Plant Physiol., 154*, 1871-1885. doi:pp.110.166488 [pii]

10.1104/pp.110.166488

Davies, E., Stankovic, B., Vian, A., & Wood, A. J. (2012). Where has all the message gone? *Plant Sci., 185*, 23-32.

Day, I. S., Golovkin, M., Palusa, S. G., Link, A., Ali, G. S., Thomas, J., . . . Reddy, A. S. N. (2012). Interactions of SR45, an SR-like protein, with spliceosomal proteins and an intronic sequence: insights into regulated splicing. *Plant J., 71*(6), 936-947. doi:10.1111/j.1365-313X.2012.05042.x

de Hoon, M. J. L., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics, 20*(9), 1453-1454.

Dekkers, B. J., Willems, L., Bassel, G. W., van Bolderen-Veldkamp, R. P., Ligterink, W., Hilhorst, H. W., & Bentsink, L. (2012). Identification of reference genes for RT-qPCR expression analysis in Arabidopsis and tomato seeds. *Plant Cell Physiol., 53*(1), 28-37. doi:10.1093/pcp/pcr113

Delmas, F., Sankaranarayanan, S., Deb, S., Widdup, E., Bournonville, C., Bollier, N., . . . Samuel, M. A. (2013). ABI3 controls embryo degreening through Mendel's I locus. *Proc. Natl. Acad. Sci. U S A*.

Deng, Y., Humbert, S., Liu, J.-X., Srivastava, R., Rothstein, S. J., & Howell, S. H. (2011). Heat induces the splicing by IRE1 of a mRNA encoding a transcription factor involved in the unfolded protein response in Arabidopsis. *Proceedings of the National Academy of Sciences, 108*(17), 7247-7252.

Deng, Y., Srivastava, R., & Howell, S. H. (2013). Protein kinase and ribonuclease domains of IRE1 confer stress tolerance, vegetative growth, and reproductive development in Arabidopsis. *Proceedings of the National Academy of Sciences, 110*(48), 19633-19638.

DeRidder, B. P., Shybut, M. E., Dyle, M. C., Kremling, K. A. G., & Shapiro, M. B. (2012). Changes at the 3 '-untranslated region stabilize Rubisco activase transcript levels during heat stress in Arabidopsis. *Planta, 236*(2), 463-476. doi:10.1007/s00425-012-1623-0

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., . . . Knowles, D. G. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res., 22*(9), 1775-1789.

DeYoung, B. J., Bickle, K. L., Schrage, K. J., Muskett, P., Patel, K., & Clark, S. E. (2006). The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in Arabidopsis. *Plant J., 45*(1), 1-16. doi:10.1111/j.1365-313X.2005.02592.x

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., . . . Lu, Z. J. (2014). Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *Plant J., 80*(5), 848-861. doi:10.1111/tpj.12679

Diaz-Meco, M. T., & Moscat, J. (2001). MEK5, a new target of the atypical protein kinase C isoforms in mitogenic signaling. *Mol. Cell. Biol., 21*(4), 1218-1227. doi:10.1128/MCB.21.4.1218-1227.2001

Dixon, D. P., Hawkins, T., Hussey, P. J., & Edwards, R. (2009). Enzyme activities and subcellular localization of members of the Arabidopsis glutathione transferase superfamily. *J. Exp. Bot., 60*(4), 1207-1218. doi:10.1093/jxb/ern365

Dubrovina, A., Kiselev, K., & Zhuravlev, Y. N. (2012). The role of canonical and noncanonical pre-mRNA splicing in plant stress responses. *BioMed research international, 2013*.

Duque, P. (2011). A role for SR proteins in plant stress responses. *Plant Signal. Behav., 6*(1), 49-54.

Duran, A. L., Yang, J., Wang, L. J., & Sumner, L. W. (2003). Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics, 19*(17), 2283-2293.

Durbak, A. R., & Tax, F. E. (2011). CLAVATA signaling pathway receptors of Arabidopsis regulate cell proliferation in fruit organ formation as well as in meristems. *Genetics, 189*(1), 177-194. doi:10.1534/genetics.111.130930

Durrett, T. P., Benning, C., & Ohlrogge, J. (2008). Plant triacylglycerols as feedstocks for the production of biofuels. *Plant J., 54*(4), 593-607. doi:10.1111/j.1365-313X.2008.03442.x

Eastmond, P. J., & Graham, I. A. (2001). Re-examining the role of the glyoxylate cycle in oilseeds. *Trends Plant Sci, 6*, 72-78. doi:S1360-1385(00)01835-5 [pii]

Fang, L., Hou, X., Lee, L. Y., Liu, L., Yan, X., & Yu, H. (2011). AtPV42a and AtPV42b redundantly regulate reproductive development in *Arabidopsis thaliana*. *PLoS One, 6*(4), e19033. doi:10.1371/journal.pone.0019033

Farmer, L. M., Book, A. J., Lee, K. H., Lin, Y. L., Fu, H., & Vierstra, R. D. (2010). The RAD23 family provides an essential connection between the 26S proteasome and ubiquitylated proteins in Arabidopsis. *Plant Cell, 22*(1), 124-142. doi:10.1105/tpc.109.072660

Fatimababy, A. S., Lin, Y. L., Usharani, R., Radjacommare, R., Wang, H. T., Tsai, H. L., . . . Fu, H. (2010). Cross-species divergence of the major recognition pathways of ubiquitylated substrates for ubiquitin/26S proteasome-mediated proteolysis. *FEBS J., 277*(3), 796-816. doi:10.1111/j.1742-4658.2009.07531.x

Fehr, W. R., Caviness, C. E., Burmood, D. T., & Pennington, J. S. (1971). Stage of development descriptions for soybean, *Glycine max* (L.) Merrill *Crop Sci, 11*, 929-931.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genet, 2*(3), 155-168. doi:10.1002/cfg.82

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol., 48*(1-2), 155-171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nature Biotech, 18*(11), 1157-1161.

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

Finkelstein, R. (2013). Abscisic Acid synthesis and response. *Arabidopsis Book, 11*, e0166. doi:10.1199/tab.0166

Finkelstein, R., Reeves, W., Ariizumi, T., & Steber, C. (2008). Molecular aspects of seed dormancy. *Annu Rev Plant Biol, 59*, 387-415. doi:10.1146/annurev.arplant.59.032607.092740

Finkelstein, R. R., Gampala, S. S., & Rock, C. D. (2002). Abscisic acid signaling in seeds and seedlings. *Plant Cell, 14 Suppl*, S15-45.

Fisher, K. M. (2008). Bayesian reconstruction of ancestral expression of the LEA gene families reveals propagule-derived desiccation tolerance in resurrection plants. *Am J Bot, 95*, 506-515. doi:95/4/506 [pii]

10.3732/ajb.95.4.506

Fouquet, R., Martin, F., Fajardo, D. S., Gault, C. M., Gomez, E., Tseung, C. W., . . . Settles, A. M. (2011). Maize *rough endosperm3* encodes an RNA splicing factor required for endosperm cell differentiation and has a nonautonomous effect on embryo development. *Plant Cell, 23*(12), 4280-4297. doi:10.1105/tpc.111.092163

Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2009). Integrated omics approaches in plant systems biology. *Curr Opin Chem Biol, 13*, 532-538. doi:S1367-5931(09)00142-2 [pii]

10.1016/j.cbpa.2009.09.022

Gechev, T. S., Dinakar, C., Benina, M., Toneva, V., & Bartels, D. (2012). Molecular mechanisms of desiccation tolerance in resurrection plants. *Cell Mol Life Sci, 69*(19), 3175-3186. doi:10.1007/s00018-012-1088-0

Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., . . . Jackson, S. A. (2009). Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol, 151*, 1167-1174. doi:10.1104/pp.109.137935

Gissot, L., Polge, C., Jossier, M., Girin, T., Bouly, J. P., Kreis, M., & Thomas, M. (2006). AKINbetagamma contributes to SnRK1 heterotrimeric complexes and interacts with two proteins implicated in plant pathogen resistance through its KIS/GBD sequence. *Plant Physiol., 142*(3), 931-944. doi:10.1104/pp.106.087718

Glaab, E., Garibaldi, J. M., & Krasnogor, N. (2009). ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinf., 10*(1), 358.

Goffard, N., & Weiller, G. (2006). Extending MapMan: application to legume genome arrays. *Bioinformatics, 22*, 2958-2959. doi:btl517 [pii]

10.1093/bioinformatics/btl517

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., . . . Putnam, N. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res., 40*(D1), D1178-D1186.

Goyer, A., Collakova, E., de la Garza, R. D., Quinlivan, E. P., Williamson, J., Gregory, J. F., . . . Hanson, A. D. (2005). 5-Formyltetrahydrofolate is an inhibitory but well tolerated metabolite in *Arabidopsis leaves*. *J Biol Chem, 280*(28), 26137-26142. doi:10.1074/jbc.M503106200

Grabarek, Z. (2006). Structural basis for diversity of the EF-hand calcium-binding proteins. *J. Mol. Biol., 359*(3), 509-525. doi:10.1016/j.jmb.2006.03.066

Graeber, K., Nakabayashi, K., Miatton, E., Leubner-Metzger, G., & Soppe, W. J. (2012). Molecular mechanisms of seed dormancy. *Plant Cell Environ., 35*(10), 1769-1786. doi:10.1111/j.1365-3040.2012.02542.x

Graham, I. A. (2008). Seed storage oil mobilization. *Annu Rev Plant Biol, 59*, 115-142. doi:10.1146/annurev.arplant.59.032607.092938

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics, 27*(7), 1017-1018. doi:10.1093/bioinformatics/btr064

Grant, D., Nelson, R. T., Cannon, S. B., & Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res., 38*, D843-D846.

Grant, G. A. (2006). The ACT domain: A small molecule binding domain and its role as a common regulatory element. *J. Biol. Chem., 281*(45), 33825-33829.

Grene, R., Vasquez-Robinet, C., & Bohnert, H. J. (2011). Molecular biology and physiological genomics of dehydration stress. In U. Luttge, E. Beck, & D. Bartels (Eds.), *Plant Desiccation Tolerance* (Vol. 215, pp. 255-287): Springer.

Guerriero, G., Martin, N., Golovko, A., Sundstrom, J. F., Rask, L., & Ezcurra, I. (2009). The RY/Sph element mediates transcriptional repression of maturation genes from late maturation to early seedling growth. *New Phytol., 184*(3), 552-565. doi:10.1111/j.1469-8137.2009.02977.x

Guillen, G., Diaz-Camino, C., Loyola-Torres, C. A., Aparicio-Fabre, R., Hernandez-Lopez, A., Diaz-Sanchez, M., & Sanchez, F. (2013). Detailed analysis of putative genes encoding small proteins in legume genomes. *Front. Plant Sci., 4*, 208. doi:10.3389/fpls.2013.00208

Guo, X., Hou, X., Fang, J., Wei, P., Xu, B., Chen, M., . . . Chu, C. (2013). The rice *GERMINATION DEFECTIVE 1*, encoding a B3 domain transcriptional repressor, regulates seed germination and seedling development by integrating GA and carbohydrate metabolism. *Plant J., 75*(3), 403-416. doi:10.1111/tpj.12209

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology, 8*(2), R24.

Gutierrez, L., Van Wuytswinkel, O., Castelain, M., & Bellini, C. (2007). Combined networks regulating seed maturation. *Trends Plant Sci, 12*, 294-300. doi:10.1016/j.tplants.2007.06.003

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell, 154*(1), 240-251.

Hadden, D. A., Phillipson, B. A., Johnston, K. A., Brown, L. A., Manfield, I. W., El-Shami, M., . . . Baker, A. (2006). Arabidopsis PEX19 is a dimeric protein that binds the peroxin PEX10. *Mol. Membr. Biol., 23*(4), 325-336. doi:10.1080/09687860600738221

Harding, E. W., Tang, W., Nichols, K. W., Fernandez, D. E., & Perry, S. E. (2003). Expression and maintenance of embryogenic potential is enhanced through constitutive expression of AGAMOUS-Like 15. *Plant Physiol., 133*(2), 653-663. doi:10.1104/pp.103.023499

Hasse, D., Mikkat, S., Hagemann, M., & Bauwe, H. (2009). Alternative splicing produces an H-protein with better substrate properties for the P-protein of glycine decarboxylase. *FEBS J., 276*(23), 6985-6991. doi:10.1111/j.1742-4658.2009.07406.x

He, Z. S., Xie, R., Zou, H. S., Wang, Y. Z., Zhu, H. B., & Yu, G. Q. (2007). Structure and alternative splicing of a heat shock transcription factor gene, MsHSF1, in *Medicago sativa*. *Biochem. Biophys. Res. Comm., 364*(4), 1056-1061. doi:10.1016/j.bbrc.2007.10.131

Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., & Dimopoulos, G. (2005). Bayesian coclustering of Anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proc Natl Acad Sci U S A, 102*(47), 16939-16944. doi:10.1073/pnas.0408393102

Henriquez‐Valencia, C., Moreno, A. A., Sandoval‐Ibañez, O., Mitina, I., Blanco‐Herrera, F., Cifuentes‐Esquivel, N., & Orellana, A. (2015). bZIP17 and bZIP60 Regulate the Expression of BiP3 and Other Salt Stress Responsive Genes in an UPR‐Independent Manner in Arabidopsis thaliana. *Journal of cellular biochemistry, 116*(8), 1638-1645.

Heo, J. B., & Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science, 331*(6013), 76-79.

Higashi, Y., Hirai, M. Y., Fujiwara, T., Naito, S., Noji, M., & Saito, K. (2006). Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. *Plant J., 48*(4), 557-571. doi:10.1111/j.1365-313X.2006.02900.x

Hill, L. M., Morley-Smith, E. R., & Rawsthorne, S. (2003). Metabolism of sugars in the endosperm of developing seeds of oilseed rape. *Plant Physiol, 131*(1), 228-236. doi:10.1104/pp.010868

Hill, L. M., & Rawsthorne, S. (2000). Carbon supply for storage-product synthesis in developing seeds of oilseed rape. *Biochem Soc Trans, 28*, 667-669.

Hirayama, T., & Shinozaki, K. (2010). Research on plant abiotic stress responses in the post-genome era: past, present and future. *Plant J., 61*(6), 1041-1052. doi:10.1111/j.1365-313X.2010.04124.x

Ho, J. W., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., & Park, P. J. (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *Bmc Genomics, 12*(1), 134.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res., 31*(13), 3429-3431.

Holdsworth, M. J., Bentsink, L., & Soppe, W. J. (2008). Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. *New Phytol, 179*, 33-54. doi:NPH2437 [pii]

10.1111/j.1469-8137.2008.02437.x

Hoppmann, V., Thorstensen, T., Kristiansen, P. E., Veiseth, S. V., Rahman, M. A., Finne, K., . . . Aasland, R. (2011). The CW domain, a new histone recognition module in chromatin proteins. *EMBO J., 30*(10), 1939-1952. doi:10.1038/emboj.2011.108

Hoth, S., Morgante, M., Sanchez, J. P., Hanafey, M. K., Tingey, S. V., & Chua, N. H. (2002). Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. *J. Cell Sci., 115*(Pt 24), 4891-4900.

Hou, A., Liu, K., Catawatcharakul, N., Tang, X., Nguyen, V., Keller, W. A., . . . Cui, Y. (2005). Two naturally occurring deletion mutants of 12S seed storage proteins in *Arabidopsis thaliana*. *Planta, 222*(3), 512-520. doi:10.1007/s00425-005-1555-z

Howell, S. H. (2013). Endoplasmic reticulum stress responses in plants. *Annual review of plant biology, 64*, 477-499.

Hsieh, M. H., & Goodman, H. M. (2002). Molecular characterization of a novel gene family encoding ACT domain repeat proteins in Arabidopsis. *Plant Physiol., 130*(4), 1797-1806.

Huang, C. Y., Chung, C. I., Lin, Y. C., Hsing, Y. I. C., & Huang, A. H. C. (2009). Oil bodies and oleosins in physcomitrella possess characteristics representative of early trends in evolution. *Plant Physiol., 150*(3), 1192-1203. doi:10.1104/pp.109.138123

Huang, H.-Y., Chien, C.-H., Jen, K.-H., & Huang, H.-D. (2006). RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Research, 34*(suppl 2), W429-W434.

Hubé, F., & Francastel, C. (2015). Mammalian introns: when the junk generates molecular diversity. *International journal of molecular sciences, 16*(3), 4429-4452.

Ietswaart, R., Wu, Z., & Dean, C. (2012). Flowering time control: another window to the connection between antisense RNA and chromatin. *Trends in Genetics, 28*(9), 445-453.

Ikeda, Y., Banno, H., Niu, Q. W., Howell, S. H., & Chua, N. H. (2006). The *ENHANCER OF SHOOT REGENERATION 2* gene in Arabidopsis regulates *CUP-SHAPED COTYLEDON 1* at the transcriptional level and controls cotyledon development. *Plant Cell Physiol., 47*(11), 1443-1456. doi:10.1093/pcp/pcl023

Illing, N., Denby, K. J., Collett, H., Shen, A., & Farrant, J. M. (2005). The signature of seeds in resurrection plants: a molecular and physiological comparison of desiccation tolerance in seeds and vegetative tissues. *Integr Comp Biol, 45*, 771-787. doi:45/5/771 [pii]

10.1093/icb/45.5.771

Immink, R. G., Posé, D., Ferrario, S., Ott, F., Kaufmann, K., Valentim, F. L., . . . Schmid, M. (2012). Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. *Plant physiology, 160*(1), 433-449.

Ito, T., Matsui, Y., Ago, T., Ota, K., & Sumimoto, H. (2001). Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions. *EMBO J., 20*(15), 3938-3946. doi:10.1093/emboj/20.15.3938

Iwata, Y., & Koizumi, N. (2012). Plant transducers of the endoplasmic reticulum unfolded protein response. *Trends in plant science, 17*(12), 720-727.

Iyer, V. V., Sriram, G., Fulton, D. B., Zhou, R., Westgate, M. E., & Shanks, J. V. (2008). Metabolic flux maps comparing the effect of temperature on protein and oil biosynthesis in developing soybean cotyledons. *Plant Cell Environ, 31*(4), 506-517. doi:10.1111/j.1365-3040.2008.01781.x

Jabnoune, M., Secco, D., Lecampion, C., Robaglia, C., Shu, Q., & Poirier, Y. (2013). A rice cis-natural antisense RNA acts as a translational enhancer for its cognate mRNA and contributes to phosphate homeostasis and plant fitness. *Plant Cell, 25*(10), 4166-4182. doi:10.1105/tpc.113.116251

James, A. B., Syed, N. H., Bordage, S., Marshall, J., Nimmo, G. A., Jenkins, G. I., . . . Nimmo, H. G. (2012). Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *Plant Cell, 24*(3), 961-981. doi:10.1105/tpc.111.093948

Jeong, S., Volny, M., & Lukowitz, W. (2012). Axis formation in Arabidopsis - transcription factors tell their side of the story. *Curr Opin Plant Biol, 15*, 4-9. doi:10.1016/j.pbi.2011.10.007

Jia, F., & Rock, C. D. (2013). MIR846 and MIR842 comprise a cistronic MIRNA pair that is regulated by abscisic acid by alternative splicing in roots of Arabidopsis. *Plant Mol. Biol., 81*(4-5), 447-460. doi:10.1007/s11103-013-0015-6

Jia, H., Suzuki, M., & McCarty, D. R. (2014). Regulation of the seed to seedling developmental phase transition by the LAFL and VAL transcription factor networks. *WIREs Dev. Biol. , 3*, 135-145.

Jiang, T., Zhang, X. F., Wang, X. F., & Zhang, D. P. (2011). Arabidopsis 3-ketoacyl-CoA thiolase-2 (KAT2), an enzyme of fatty acid beta-oxidation, is involved in ABA signal transduction. *Plant Cell Physiol., 52*(3), 528-538.

Jin, J., Liu, J., Wang, H., Wong, L., & Chua, N.-H. (2013). PLncDB: plant long non-coding RNA database. *Bioinformatics, 29*(8), 1068-1071.

Jones, M. A., Williams, B. A., McNicol, J., Simpson, C. G., Brown, J. W. S., & Harmer, S. L. (2012). Mutation of Arabidopsis *SPLICEOSOMAL TIMEKEEPER LOCUS1* causes circadian clock defects. *Plant Cell, 24*(10), 4066-4082. doi:10.1105/tpc.112.104828

Jorgensen, R. A., & Dorantes-Acosta, A. E. (2012). Conserved peptide upstream open reading frames are associated with regulatory genes in angiosperms. *Front. Plant Sci., 3*, 191. doi:10.3389/fpls.2012.00191

Junker, A., Monke, G., Rutten, T., Keilwagen, J., Seifert, M., Thi, T. M., . . . Baumlein, H. (2012). Elongation-related functions of LEAFY COTYLEDON1 during the development of *Arabidopsis thaliana*. *Plant J., 71*(3), 427-442. doi:10.1111/j.1365-313X.2012.04999.x

Kagale, S., & Rozwadowski, K. (2011). EAR motif-mediated transcriptional repression in plants: an underlying mechanism for epigenetic regulation of gene expression. *Epigenetics, 6*(2), 141-146.

Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., . . . Hattori, T. (2005). Indirect ABA-dependent regulation of seed storage protein genes by FUSCA3 transcription factor in Arabidopsis. *Plant Cell Physiol., 46*(2), 300-311. doi:10.1093/pcp/pci031

Kakumanu, A., Ambavaram, M. M., Klumas, C., Krishnan, A., Batlang, U., Myers, E., . . . Pereira, A. (2012). Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. *Plant Physiol, 160*, 846-867. doi:pp.112.200444 [pii]

10.1104/pp.112.200444

Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., . . . Brown, J. W. S. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res., 40*(6), 2454-2469. doi:10.1093/nar/gkr932

Kanno, Y., Jikumaru, Y., Hanada, A., Nambara, E., Abrams, S. R., Kamiya, Y., & Seo, M. (2010). Comprehensive hormone profiling in developing Arabidopsis seeds: examination of the site of ABA biosynthesis, ABA transport and hormone interactions. *Plant Cell Physiol., 51*(12), 1988-2001. doi:10.1093/pcp/pcq158

Kaufmann, K., Muino, J. M., Østerås, M., Farinelli, L., Krajewski, P., & Angenent, G. C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors

followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nat Protoc, 5*(3), 457-472.

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

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol, 14*(4), R36.

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic acids research, 35*(1), 125-131.

Kim, H. S., Abbasi, N., & Choi, S. B. (2013). Bruno‐like proteins modulate flowering time via 3′ UTR‐dependent decay of SOC1 mRNA. *New Phytologist, 198*(3), 747-756.

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., & Fiehn, O. (2009). FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry, 81*(24), 10038-10048. doi:10.1021/ac9019522

Klie, S., Krueger, S., Krall, L., Giavalisco, P., Flugge, U. I., Willmitzer, L., & Steinhauser, D. (2011). Analysis of the compartmentalized metabolome - a validation of the non-aqueous fractionation technique. *Front Plant Sci, 2*, 55. doi:10.3389/fpls.2011.00055

Koncz, C., deJong, F., Villacorta, N., Szakonyi, D., & Koncz, Z. (2012). The spliceosome-activating complex: molecular mechanisms underlying the function of a pleiotropic regulator. *Frontiers in plant science, 3*, 9.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research, 35*(suppl 2), W345-W349.

Koo, A. J., Thireault, C., Zemelis, S., Poudel, A. N., Zhang, T., Kitaoka, N., . . . Howe, G. A. (2014). Endoplasmic reticulum-associated inactivation of the hormone jasmonoyl-L-isoleucine by multiple members of the cytochrome P450 94 family in Arabidopsis. *J. Biol. Chem., 289*(43), 29728-29738. doi:10.1074/jbc.M114.603084

Kriechbaumer, V., Wang, P. W., Hawes, C., & Abell, B. M. (2012). Alternative splicing of the auxin biosynthesis gene YUCCA4 determines its subcellular compartmentation. *Plant J., 70*(2), 292-302. doi:10.1111/j.1365-313X.2011.04866.x

Kueger, S., Steinhauser, D., Willmitzer, L., & Giavalisco, P. (2012). High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *Plant J, 70*, 39-50. doi:10.1111/j.1365-313X.2012.04902.x

Kung, J. T., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics, 193*(3), 651-669. doi:10.1534/genetics.112.146704

Kunz, H. H., Scharnewski, M., Feussner, K., Feussner, I., Flugge, U. I., Fulda, M., & Gierth, M. (2009). The ABC transporter PXA1 and peroxisomal beta-oxidation are vital for metabolism in mature leaves of Arabidopsis during extended darkness. *Plant Cell, 21*(9), 2733-2749.

Kuroda, H., Yanagawa, Y., Takahashi, N., Horii, Y., & Matsui, M. (2012). A comprehensive analysis of interaction and localization of Arabidopsis SKP1-like (ASK) and F-box (FBX) proteins. *PLoS One, 7*(11), e50009. doi:10.1371/journal.pone.0050009

Kwon, C. S., Hibara, K., Pfluger, J., Bezhani, S., Metha, H., Aida, M., . . . Wagner, D. (2006). A role for chromatin remodeling in regulation of *CUC* gene expression in the Arabidopsis cotyledon boundary. *Development, 133*(16), 3223-3230. doi:10.1242/dev.02508

Kwong, R. W., Bui, A. Q., Lee, H., Kwong, L. W., Fischer, R. L., Goldberg, R. B., & Harada, J. J. (2003). LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell, 15*(1), 5-18.

Labusch, C., Shishova, M., Effendi, Y., Li, M., Wang, X., & Scherer, G. F. (2013). Patterns and timing in expression of early auxin-induced genes imply involvement of phospholipases A (pPLAs) in the regulation of auxin responses. *Mol. Plant, 6*(5), 1473-1486. doi:10.1093/mp/sst053

Lamberto, I., Percudani, R., Gatti, R., Folli, C., & Petrucco, S. (2010). Conserved alternative splicing of Arabidopsis transthyretin-like determines protein localization and S-allantoin synthesis in peroxisomes. *Plant Cell, 22*(5), 1564-1574. doi:10.1105/tpc.109.070102

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., . . . Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res., 40*(Database issue), D1202-D1210. doi:10.1093/nar/gkr1090

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*. doi:R25

10.1186/gb-2009-10-3-r25

Lau, S., Slane, D., Herud, O., Kong, J., & Juergens, G. (2012). Early embryogenesis in flowering plants: Setting up the basic body pattern. In S. S. Merchant (Ed.), *Annu Rev Plant Biol* (Vol. 63, pp. 483-506).

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol, 15*(2), R29.

Le, B. H., Cheng, C., Bui, A. Q., Wagmaister, J. A., Henry, K. F., Pelletier, J., . . . Goldberg, R. B. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc. Natl. Acad. Sci. U. S. A., 107*(18), 8063-8070. doi:1003530107 [pii]

10.1073/pnas.1003530107

Le, B. H., Wagmaister, J. A., Kawashima, T., Bui, A. Q., Harada, J. J., & Goldberg, R. B. (2007). Using genomics to study legume seed development. *Plant Physiol, 144*, 562-574. doi:10.1104/pp.107.100362

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., . . . Kitano, H. (2009). The Systems Biology Graphical Notation. *Nat. Biotechnol., 27*(8), 735-741. doi:10.1038/nbt.1558

Lee, H., Fischer, R. L., Goldberg, R. B., & Harada, J. J. (2003). Arabidopsis LEAFY COTYLEDON1 represents a functionally specialized subunit of the CCAAT binding transcription factor. *Proc. Natl. Acad. Sci. U. S. A., 100*(4), 2152-2156. doi:10.1073/pnas.0437909100

Li, W. F., Lin, W. D., Ray, P., Lan, P., & Schmidt, W. (2013). Genome-wide detection of condition-sensitive alternative splicing in Arabidopsis roots. *Plant Physiol., 162*(3), 1750-1763. doi:10.1104/pp.113.217778

Li, X., Zhuo, J. J., Jing, Y., Liu, X., & Wang, X. F. (2012). Expression of a GALACTINOL SYNTHASE gene is positively associated with desiccation tolerance of *Brassica napus* seeds during development. *J Plant Physiol, 168*, 1761-1770. doi:10.1016/j.jplph.2011.04.006

Li, Y., Beisson, F., Pollard, M., & Ohlrogge, J. (2006). Oil content of Arabidopsis seeds: the influence of seed anatomy, light and plant-to-plant variation. *Phytochemistry, 67*(9), 904-915. doi:S0031-9422(06)00114-2 [pii]

10.1016/j.phytochem.2006.02.015

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., . . . Ohlrogge, J. (2013). Acyl-lipid metabolism. *Arabidopsis Book, 11*, e0161.

Liberles, J. S., Thorolfson, M., & Martinez, A. (2005). Allosteric mechanisms in ACT domain containing enzymes involved in amino acid metabolism. *Amino Acids, 28*(1), 1-12.

Liberman, L. M., Sozzani, R., & Benfey, P. N. (2012). Integrative systems biology: an attempt to describe a simple weed. *Curr Opin Plant Biol, 15*, 162-167. doi:S1369-5266(12)00005-2 [pii]

10.1016/j.pbi.2012.01.004

Lie, C., Kelsom, C., & Wu, X. (2012). WOX2 and STIMPY-LIKE/WOX8 promote cotyledon boundary formation in Arabidopsis. *Plant J., 72*(4), 674-682. doi:10.1111/j.1365-313X.2012.05113.x

Lightfoot, D. J., Malone, K. M., Timmis, J. N., & Orford, S. J. (2008). Evidence for alternative splicing of MADS-box transcripts in developing cotton fibre cells. *Mol. Genet. Genomics, 279*(1), 75-85. doi:10.1007/s00438-007-0297-y

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics, 27*(13), i275-i282.

Lindsey, K., Casson, S., & Chilley, P. (2002). Peptides: new signalling molecules in plants. *Trends in plant science, 7*(2), 78-83.

Lister, R., Gregory, B. D., & Ecker, J. R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol., 12*(2), 107-118. doi:10.1016/j.pbi.2008.11.004

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., . . . Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell Online, 24*(11), 4333-4345.

Liu, J. X., & Howell, S. H. (2016). Managing the protein folding demands in the endoplasmic reticulum of plants. *New Phytologist*.

Liu, Q., Kasuga, M., Sakuma, Y., Abe, H., Miura, S., Yamaguchi-Shinozaki, K., & Shinozaki, K. (1998). Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis. *Plant Cell, 10*(8), 1391-1406.

Liu, R., Loraine, A. E., & Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC bioinformatics, 15*(1), 364.

Liu, X., Zhang, H., Zhao, Y., Feng, Z., Li, Q., Yang, H. Q., . . . He, Z. H. (2013). Auxin controls seed dormancy through stimulation of abscisic acid signaling by inducing ARF-mediated ABI3 activation in Arabidopsis. *Proc. Natl. Acad. Sci. U S A, 110*(38), 15485-15490.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2^{-DDCT} method. *Methods, 25*(4), 402-408.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., . . . Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ., 37*(5), 1250-1258.

Lonien, J., & Schwender, J. (2009). Analysis of metabolic flux phenotypes for two *Arabidopsis* mutants with severe impairment in seed storage lipid synthesis. *Plant Physiol, 151*(3), 1617-1634. doi:10.1104/pp.109.144121

Lorkovic, Z. J., Lopato, S., Pexa, M., Lehner, R., & Barta, A. (2004). Interactions of Arabidopsis RS domain containing cyclophilins with SR proteins and U1 and U11 small nuclear ribonucleoprotein-specific proteins suggest their involvement in pre-mRNA splicing. *J. Biol. Chem., 279*(32), 33890-22898.

Lu, Q. S., Paz, J. D., Pathmanathan, A., Chiu, R. S., Tsai, A. Y., & Gazzarrini, S. (2010). The C-terminal domain of FUSCA3 negatively regulates mRNA and protein levels, and mediates sensitivity to the hormones abscisic acid and gibberellic acid in Arabidopsis. *Plant J., 64*(1), 100-113. doi:10.1111/j.1365-313X.2010.04307.x

Lu, X. D., Li, Y., Su, Y. P., Liang, Q. J., Meng, H. Y., Li, S., . . . Zhang, C. Y. (2012). An Arabidopsis gene encoding a C2H2-domain protein with alternatively spliced

transcripts is essential for endosperm development. *J. Exp. Bot., 63*(16), 5935-5944. doi:10.1093/jxb/ers243

Lu, Y., Savage, L. J., Larson, M. D., Wilkerson, C. G., & Last, R. L. (2011). Chloroplast 2010: A Database for Large-Scale Phenotypic Screening of Arabidopsis Mutants. *Plant Physiol, 155*, 1589-1600. doi:10.1104/pp.110.170118

Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., . . . Cheng, C. (2011). Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res., 21*(2), 276-285.

Lynch, T., Erickson, B. J., & Finkelstein, R. R. (2012). Direct interactions of ABA-insensitive(ABI)-clade protein phosphatase(PP)2Cs with calcium-dependent protein kinases and ABA response element-binding bZIPs may contribute to turning off ABA response. *Plant Mol. Biol., 80*(6), 647-658. doi:10.1007/s11103-012-9973-3

Ma, F., Jazmin, L. J., Young, J. D., & Allen, D. K. (2014). Isotopically nonstationary 13C flux analysis of changes in *Arabidopsis thaliana* leaf metabolism due to high light acclimation. *Proc. Natl. Acad. Sci. U. S. A., 111*(47), 16967-16972. doi:10.1073/pnas.1319485111

Mancini, E., Sanchez, S. E., Romanowski, A., Schlaen, R. G., Sanchez‐Lamas, M., Cerdán, P. D., & Yanovsky, M. J. (2016). Acute Effects of Light on Alternative Splicing in Light‐Grown Plants. *Photochemistry and photobiology, 92*(1), 126-133.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., . . . Gonzales, N. R. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research, 39*(suppl 1), D225-D229.

Marden, J. H. (2008). Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity, 100*(2), 111-120. doi:10.1038/sj.hdy.6800904

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics, 12*(10), 671-682.

Martin-Trillo, M., Grandio, E. G., Serra, F., Marcel, F., Rodriguez-Buey, M. L., Schmitz, G., . . . Cubas, P. (2011). Role of tomato *BRANCHED1-like* genes in the control of shoot branching. *Plant J., 67*(4), 701-714. doi:10.1111/j.1365-313X.2011.04629.x

Mastrangelo, A. M., Marone, D., Laido, G., De Leonardis, A. M., & De Vita, P. (2012). Alternative splicing: Enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci., 185*, 40-49. doi:10.1016/j.plantsci.2011.09.006

Matsukura, S., Mizoi, J., Yoshida, T., Todaka, D., Ito, Y., Maruyama, K., . . . Yamaguchi-Shinozaki, K. (2010). Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic

stress-responsive genes. *Mol. Genet. Genomics, 283*(2), 185-196. doi:10.1007/s00438-009-0506-y

Matsumura, H., Kitajima, H., Akada, S., Abe, J., Minaka, N., & Takahashi, R. (2009). Molecular cloning and linkage mapping of cryptochrome multigene family in soybean. *Plant Genome, 2*(3), 271-281. doi:10.3835/plantgenome2009.06.0018

Mazzucotelli, E., Mastrangelo, A. A., Crosatti, C., Guerra, D., Stanca, A. M., & Cattivelli, L. (2008). Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription. *Plant Sci., 174*(4), 420-431. doi:10.1016/j.plantsci.2008.02.005

McLeay, R. C., & Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics, 11*, 165. doi:10.1186/1471-2105-11-165

Meinke, D. W. (1995). Molecular genetics of plant embryogenesis. *Annual review of plant biology, 46*(1), 369-394.

Meinke, D. W., Chen, J., & Beachy, R. N. (1981). Expression of storage-protein genes during soybean seed development. *Planta, 153*, 130-139.

Mendes, A., Kelly, A. A., van Erp, H., Shaw, E., Powers, S. J., Kurup, S., & Eastmond, P. J. (2013). bZIP67 regulates the omega-3 fatty acid content of Arabidopsis seed oil by activating fatty acid desaturase3. *Plant Cell, 25*(8), 3104-3116. doi:10.1105/tpc.113.116343

Mercer, T. R., & Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology, 20*(3), 300-307.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., & Gottardo, R. (2011). An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS One, 6*(2), e16432.

Metzger, M. B., Hristova, V. A., & Weissman, A. M. (2012). HECT and RING finger families of E3 ubiquitin ligases at a glance. *J. Cell Sci., 125*, 531-537.

Mizoi, J., Ohori, T., Moriwaki, T., Kidokoro, S., Todaka, D., Maruyama, K., . . . Yamaguchi-Shinozaki, K. (2013). GmDREB2A;2, a canonical DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN2-type transcription factor in soybean, is posttranslationally regulated and mediates dehydration-responsive element-dependent gene expression. *Plant Physiol., 161*(1), 346-361. doi:10.1104/pp.112.204875

Mochizuki, N., Brusslan, J. A., Larkin, R., Nagatani, A., & Chory, J. (2001). Arabidopsis genomes uncoupled 5 (GUN5) mutant reveals the involvement of Mg-chelatase H subunit in plastid-to-nucleus signal transduction. *Proc. Natl. Acad. Sci. U. S. A., 98*(4), 2053-2058. doi:10.1073/pnas.98.4.2053

Moco, S., Schneider, B., & Vervoort, J. (2009). Plant Micrometabolomics: The Analysis of Endogenous Metabolites Present in a Plant Cell or Tissue. *J Proteome Res, 8*, 1694-1703. doi:10.1021/pr800973r

Monke, G., Altschmied, L., Tewes, A., Reidt, W., Mock, H. P., Baumlein, H., & Conrad, U. (2004). Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. *Planta, 219*(1), 158-166. doi:10.1007/s00425-004-1206-9

Monke, G., Seifert, M., Keilwagen, J., Mohr, M., Grosse, I., Hahnel, U., . . . Altschmied, L. (2012). Toward the identification and regulation of the *Arabidopsis thaliana* ABI3 regulon. *Nucleic Acids Res., 40*(17), 8240-8254. doi:10.1093/nar/gks594

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol., 9 Suppl 1*, S4. doi:10.1186/gb-2008-9-s1-s4

Mouriz, A., Lopez-Gonzalez, L., Jarillo, J. A., & Pineiro, M. (2015). PHDs govern plant development. *Plant Signal Behav., 10*(7), e993253. doi:10.4161/15592324.2014.993253

Muller, M., & Munne-Bosch, S. (2011). Rapid and sensitive hormonal profiling of complex plant samples by liquid chromatography coupled to electrospray ionization tandem mass spectrometry. *Plant Methods, 7*, 37. doi:10.1186/1746-4811-7-37

Munier-Jolain, N. G., Munier-Jolain, N. M., Roche, R., Ney, B., & Duthion, C. (1998). Seed growth rate in grain legumes - I. Effect of photoassimilate availability on seed growth rate. *J Exp Bot, 49*, 1963-1969. doi:10.1093/jexbot/49.329.1963

Nakajima, S., Ito, H., Tanaka, R., & Tanaka, A. (2012). Chlorophyll b reductase plays an essential role in maturation and storability of Arabidopsis seeds. *Plant Physiol., 160*(1), 261-273.

Nambara, E., & Marion-Poll, A. (2005). Abscisic acid biosynthesis and catabolism. *Annu. Rev. Plant Biol., 56*, 165-185. doi:10.1146/annurev.arplant.56.032604.144046

Ni, M., Tepperman, J. M., & Quail, P. H. (1998). PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. *Cell, 95*(5), 657-667.

Nishizawa, A., Yabuta, Y., & Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol, 147*, 1251-1263. doi:10.1104/pp.108.122465

Oh, E., Yamaguchi, S., Kamiya, Y., Bae, G., Chung, W. I., & Choi, G. (2006). Light activates the degradation of PIL5 protein to promote seed germination through gibberellin in Arabidopsis. *Plant J., 47*(1), 124-139. doi:10.1111/j.1365-313X.2006.02773.x

Onate-Sanchez, L., & Vicente-Carbajosa, J. (2008). DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. *BMC Res. Notes, 1*, 93. doi:1756-0500-1-93 [pii]

10.1186/1756-0500-1-93

Palusa, S. G., Ali, G. S., & Reddy, A. S. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J., 49*(6), 1091-1107.

Palusa, S. G., & Reddy, A. S. N. (2010). Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. *New Phytologist, 185*(1), 83-89. doi:10.1111/j.1469-8137.2009.03065.x

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics, 10*(10), 669-680.

Parra-Rojas, J., Moreno, A. A., Mitina, I., & Orellana, A. (2015). The Dynamic of the Splicing of bZIP60 and the Proteins Encoded by the Spliced and Unspliced mRNAs Reveals Some Unique Features during the Activation of UPR in Arabidopsis thaliana. *PLoS One, 10*(4), e0122936.

Paz, I., Akerman, M., Dror, I., Kosti, I., & Mandel-Gutfreund, Y. (2010). SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Research*, gkq444.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Pelechano, V., & Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nature Reviews Genetics, 14*(12), 880-893.

Penfield, S., Graham, S., & Graham, I. A. (2005). Storage reserve mobilization in germinating oilseeds: Arabidopsis as a model system. *Biochem Soc Trans, 33*, 380-383. doi:BST0330380 [pii]

10.1042/BST0330380

Penfield, S., Josse, E. M., & Halliday, K. J. (2010). A role for an alternative splice variant of PIF6 in the control of Arabidopsis primary seed dormancy. *Plant Mol. Biol., 73*(1-2), 89-95. doi:10.1007/s11103-009-9571-1

Penfield, S., Pinfield-Wells, H. M., & Graham, I. A. (2006). Storage reserve mobilisation and seedling establishment in Arabidopsis. *Arabidopsis Book, 4*, e0100. doi:10.1199/tab.0100

Peremyslov, V., Mockler, T. C., Filichkin, S. A., Fox, S. E., Jaiswal, P., Makarova, K. S., . . . Dolja, V. V. (2011). Expression, splicing, and evolution of the myosin gene family in plants. *Plant Physiol., 155*(3), 1191-1204. doi:10.1104/pp.110.170720

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

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol, 33*(3), 290-295.

Pillitteri, L. J., Bemis, S. M., Shpak, E. D., & Torii, K. U. (2007). Haploinsufficiency after successive loss of signaling reveals a role for ERECTA-family genes in Arabidopsis ovule development. *Development, 134*(17), 3099-3109. doi:10.1242/dev.004788

Piva, F., Giulietti, M., Nocchi, L., & Principato, G. (2009). SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics, 25*(9), 1211-1213.

Ponjavic, J., Ponting, C. P., & Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome research, 17*(5), 556-565.

Puyaubert, J., Denis, L., & Alban, C. (2008). Dual targeting of Arabidopsis holocarboxylase synthetase1: A small upstream open reading frame regulates translation initiation and protein targeting. *Plant Physiol., 146*(2), 478-491. doi:10.1104/pp.107.111534

Qi, Y., Tsuda, K., Joe, A., Sato, M., Nguyen, L. V., Glazebrook, J., . . . Katagiri, F. (2010). A putative RNA-binding protein positively regulates salicylic acid-mediated immunity in Arabidopsis. *Molecular plant-microbe interactions, 23*(12), 1573-1583.

Qin, F., Sakuma, Y., Tran, L. S., Maruyama, K., Kidokoro, S., Fujita, Y., . . . Yamaguchi-Shinozaki, K. (2008). Arabidopsis DREB2A-interacting proteins function as RING E3 ligases and negatively regulate plant drought stress-responsive gene expression. *Plant Cell, 20*(6), 1693-1707. doi:10.1105/tpc.107.057380

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., . . . Yang, A. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature, 499*(7457), 172-177.

Reddy, A. S., Marquez, Y., Kalyna, M., & Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *The Plant Cell, 25*(10), 3657-3683.

Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry, 81*.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, gkv007.

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics, 27*(17), 2325-2329. doi:10.1093/bioinformatics/btr355

Rogers, M. F., Thomas, J., Reddy, A. S., & Ben-Hur, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol., 13*(1), R4.

Rolletschek, H., Radchuk, R., Klukas, C., Schreiber, F., Wobus, U., & Borisjuk, L. (2005). Evidence of a key role for photosynthetic oxygen release in oil storage in

developing soybean seeds. *New Phytol, 167*, 777-786. doi:10.1111/j.1469-8137.2005.01473.x

Roman, A., Andreu, V., Hernandez, M. L., Lagunas, B., Picorel, R., Martinez-Rivas, J. M., & Alfonso, M. (2012). Contribution of the different omega-3 fatty acid desaturase genes to the cold response in soybean. *J. Exp. Bot., 63*(13), 4973-4982. doi:10.1093/jxb/ers174

Rosloski, S. M., Singh, A., Jali, S. S., Balasubramanian, S., Weigel, D., & Grbic, V. (2013). Functional analysis of splice variant expression of MADS AFFECTING FLOWERING 2 of *Arabidopsis thaliana*. *Plant Mol. Biol., 81*(1-2), 57-69. doi:10.1007/s11103-012-9982-2

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., . . . Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. *Plant J, 73*, 143-153. doi:10.1111/tpj.12026

Ruberti, C., Kim, S.-J., Stefano, G., & Brandizzi, F. (2015). Unfolded protein response in plants: one master, many questions. *Current opinion in plant biology, 27*, 59-66.

Ruhl, C., Stauffer, E., Kahles, A., Wagner, G., Drechsel, G., Ratsch, G., & Wachter, A. (2012). Polypyrimidine tract binding protein homologs from Arabidopsis are key regulators of alternative splicing with implications in fundamental developmental processes. *Plant Cell, 24*(11), 4360-4375. doi:10.1105/tpc.112.103622

Ruuska, S. A., Schwender, J., & Ohlrogge, J. B. (2004). The capacity of green oilseeds to utilize photosynthesis to drive biosynthetic processes. *Plant Physiol, 136*(1), 2700-2709.

Ryu, J. Y., Kim, J.-Y., & Park, C.-M. (2015). Adaptive thermal control of stem gravitropism through alternative RNA splicing in Arabidopsis. *Plant signaling & behavior, 10*(11), e1093715.

Saavedra, X., Modrego, A., Rodriguez, D., Gonzalez-Garcia, M. P., Sanz, L., Nicolas, G., & Lorenzo, O. (2010). The nuclear interactor PYL8/RCAR3 of *Fagus sylvatica* FsPP2C1 is a positive regulator of abscisic acid signaling in seeds and stress. *Plant Physiol., 152*(1), 133-150. doi:10.1104/pp.109.146381

Saha, R., Suthers, P. F., & Maranas, C. D. (2011). *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS One, 6*(7), e21784. doi:10.1371/journal.pone.0021784

PONE-D-11-02644 [pii]

Saldanha, A. J. (2004). Java Treeview - extensible visualization of microarray data. *Bioinformatics, 20*(17), 3246-3248.

Sanchez, R., & Zhou, M. M. (2011). The PHD finger: a versatile epigenome reader. *Trends Biochem. Sci., 36*(7), 364-372. doi:10.1016/j.tibs.2011.03.005

Sanchez, S. E., Petrillo, E., Beckwith, E. J., Zhang, X., Rugnone, M. L., Hernando, C. E., . . . Yanovsky, M. J. (2010). A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature, 468*(7320), 112-116. doi:10.1038/nature09470

Sano, N., Permana, H., Kumada, R., Shinozaki, Y., Tanabata, T., Yamada, T., . . . Kanekatsu, M. (2012). Proteomic analysis of embryonic proteins synthesized from long-lived mRNAs during germination of rice seeds. *Plant Cell Physiol, 53*, 687-698. doi:pcs024 [pii]

10.1093/pcp/pcs024

Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., & Lepiniec, L. (2008). Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *Plant J., 54*(4), 608-620. doi:10.1111/j.1365-313X.2008.03461.x

Santosh, B., Varshney, A., & Yadava, P. K. (2015). Non‐coding RNAs: biological functions and applications. *Cell biochemistry and function, 33*(1), 14-22.

Schlaen, R. G., Mancini, E., Sanchez, S. E., Perez-Santángelo, S., Rugnone, M. L., Simpson, C. G., . . . Yanovsky, M. J. (2015). The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms. *Proceedings of the National Academy of Sciences, 112*(30), 9382-9387.

Schlueter, J. A., Lin, J. Y., Schlueter, S. D., Vasylenko-Sanders, I. F., Deshpande, S., Yi, J., . . . Shoemaker, R. C. (2007). Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *Bmc Genomics, 8*, 330. doi:1471-2164-8-330 [pii]

10.1186/1471-2164-8-330

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., . . . Lohmann, J. U. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet., 37*(5), 501-506.

Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc., 3*(6), 1101-1108.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., . . . Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature, 463*, 178-183. doi:nature08670 [pii]

10.1038/nature08670

Schneider, A., Aghamirzaie, D., Elmarakeby, H., Poudel, A. N., Koo, A. J., Heath, L. S., . . . Collakova, E. (2015). Potential targets of VIVIPAROUS1/ABI3‐LIKE1 (VAL1) repression in developing Arabidopsis thaliana embryos. *The Plant Journal*.

Schubert, D., Primavesi, L., Bishopp, A., Roberts, G., Doonan, J., Jenuwein, T., & Goodrich, J. (2006). Silencing by plant Polycomb-group genes requires dispersed trimethylation of histone H3 at lysine 27. *EMBO J., 25*(19), 4638-4649. doi:10.1038/sj.emboj.7601311

Schultz, T. F., & Quatrano, R. S. (1997). Characterization and expression of a rice RAD23 gene. *Plant Mol. Biol., 34*(3), 557-562.

Schwartz, S., Hall, E., & Ast, G. (2009). SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Research, 37*(suppl 2), W189-W192.

Schweighofer, A., Hirt, H., & Meskiene, I. (2004). Plant PP2C phosphatases: emerging functions in stress signaling. *Trends Plant Sci., 9*(5), 236-243. doi:10.1016/j.tplants.2004.03.007

Schwender, J., Goffman, F., Ohlrogge, J. B., & Shachar-Hill, Y. (2004). Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature, 432*(7018), 779-782.

Schwender, J., & Ohlrogge, J. B. (2002). Probing in vivo metabolism by stable isotope labeling of storage lipids and proteins in developing *Brassica napus* embryos. *Plant Physiol, 130*(1), 347-361.

Schwender, J., Shachar-Hill, Y., & Ohlrogge, J. B. (2006). Mitochondrial metabolism in developing embryos of *Brassica napus*. *J Biol Chem, 281*(45), 34040-34047. doi:10.1074/jbc.M606266200

Severing, E. I., van Dijk, A. D., & van Ham, R. C. (2011). Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. *BMC plant biology, 11*(1), 82.

Severing, E. I., van Dijk, A. D. J., Morabito, G., Busscher-Lange, J., Immink, R. G. H., & van Ham, R. (2012). Predicting the impact of alternative splicing on plant MADS domain protein function. *PLOS One, 7*(1), e30524. doi:e30524

10.1371/journal.pone.0030524

Severing, E. I., van Dijk, A. D. J., Stiekema, W. J., & van Ham, R. (2009). Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics, 10*, 154. doi:154

10.1186/1471-2164-10-154

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research, 13*(11), 2498-2504.

Sharma, N., Bender, Y., Boyle, K., & Fobert, P. R. (2013). High-level expression of sugar inducible gene2 (*HSI2*) is a negative regulator of drought stress tolerance in Arabidopsis. *BMC Plant Biol., 13*, 170. doi:10.1186/1471-2229-13-170

Sharov, A. A., & Ko, M. S. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA research, 16*(5), 261-273.

Shu, K., Zhang, H., Wang, S., Chen, M., Wu, Y., Tang, S., . . . Xie, Q. (2013). ABI4 regulates primary seed dormancy by regulating the biogenesis of abscisic acid and gibberellins in Arabidopsis. *PLoS Genet., 9*(6), e1003577.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Richards, S. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research, 15*(8), 1034-1050.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol., 3*, Article3. doi:10.2202/1544-6115.1027

Smyth, G. K. (2005). Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor*, 397--420.

Song, Z.-T., Sun, L., Lu, S.-J., Tian, Y., Ding, Y., & Liu, J.-X. (2015). Transcription factor interaction with COMPASS-like complex regulates histone H3K4 trimethylation for specific gene expression in plants. *Proceedings of the National Academy of Sciences, 112*(9), 2900-2905.

Sriram, G., Fulton, D. B., Iyer, V. V., Peterson, J. M., Zhou, R., Westgate, M. E., . . . Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional 13C labeling, two-dimensional [13C, 1H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiol, 136*, 3043-3057.

Srivastava, R., Deng, Y., Shah, S., Rao, A. G., & Howell, S. H. (2013). BINDING PROTEIN is a master regulator of the endoplasmic reticulum stress sensor/transducer bZIP28 in Arabidopsis. *The Plant Cell, 25*(4), 1416-1429.

Staiger, D., Korneli, C., Lummer, M., & Navarro, L. (2013). Emerging role for RNA-based regulation in plant immunity. *New Phytologist, 197*(2), 394-404. doi:10.1111/nph.12022

Stauffer, E., Westermann, A., Wagner, G., & Wachter, A. (2010). Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant J., 64*(2), 243-255. doi:10.1111/j.1365-313X.2010.04321.x

Stone, S. L. (2014). The role of ubiquitin and the 26S proteasome in plant abiotic stress signaling. *Front. Plant Sci., 5*, 135. doi:10.3389/fpls.2014.00135

Streitner, C., Koster, T., Simpson, C. G., Shaw, P., Danisman, S., Brown, J. W. S., & Staiger, D. (2012). An hnRNP-like RNA-binding protein affects alternative splicing by in vivo interaction with transcripts in *Arabidopsis thaliana*. *Nucleic Acids Res., 40*(22), 11240-11255. doi:10.1093/nar/gks873

Strijbis, K., van den Burg, J., Visser, W. F., van den Berg, M., & Distel, B. (2012). Alternative splicing directs dual localization of *Candida albicans* 6-phosphogluconate dehydrogenase to cytosol and peroxisomes. *FEMS Yeast Res., 12*(1), 61-68.

Sugliani, M., Brambilla, V., Clerkx, E. J. M., Koornneef, M., & Soppe, W. J. J. (2010). The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *Plant Cell, 22*(6), 1936-1946. doi:10.1105/tpc.110.074674

Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., & Sun, H. (2013). iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics, 14*(Suppl 2), S7.

Suzuki, M., Wang, H. H., & McCarty, D. R. (2007). Repression of the LEAFY COTYLEDON 1/B3 regulatory network in plant embryo development by VP1/ABSCISIC ACID INSENSITIVE 3-LIKE B3 genes. *Plant Physiol., 143*(2), 902-911. doi:10.1104/pp.106.092320

Swaminathan, K., Peterson, K., & Jack, T. (2008). The plant B3 superfamily. *Trends Plant Sci., 13*(12), 647-655. doi:10.1016/j.tplants.2008.09.006

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., & Brown, J. W. S. (2012). Alternative splicing in plants–coming of age. *Trends in plant science, 17*(10), 616-623.

Tan, J. L., Wang, C. Y., Xiang, B., Han, R. H., & Guo, Z. F. (2013). Hydrogen peroxide and nitric oxide mediated cold- and dehydration-induced myo-inositol phosphate synthase that confers multiple resistances to abiotic stresses. *Plant Cell Environ, 36*, 288-299. doi:10.1111/j.1365-3040.2012.02573.x

Terasawa, H., Noda, Y., Ito, T., Hatanaka, H., Ichikawa, S., Ogura, K., . . . Inagaki, F. (2001). Structure and ligand recognition of the PB1 domain: a novel protein module binding to the PC motif. *EMBO J., 20*(15), 3947-3956. doi:10.1093/emboj/20.15.3947

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., . . . Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J, 37*(6), 914-939. doi:10.1111/j.1365-313X.20041.02016.x

Thomas, J., Palusa, S. G., Prasad, K., Ali, G. S., Surabhi, G. K., Ben-Hur, A., . . . Reddy, A. S. N. (2012). Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. *Plant J., 72*(6), 935-946. doi:10.1111/tpj.12004

Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., & Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res., 16*(7), 885-889.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol, 31*(1), 46-53. doi:10.1038/nbt.2450

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology, 31*(1), 46-53.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc, 7*(3), 562-578. doi:10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol, 28*(5), 511-515. doi:10.1038/nbt.1621

Tsai, A. Y., & Gazzarrini, S. (2012). AKIN10 and FUSCA3 interact to control lateral organ development and phase transitions in Arabidopsis. *Plant J., 69*(5), 809-821. doi:10.1111/j.1365-313X.2011.04832.x

Tsukagoshi, H., Morikami, A., & Nakamura, K. (2007). Two B3 domain transcriptional repressors prevent sugar-inducible expression of seed maturation genes in Arabidopsis seedlings. *Proc. Natl. Acad. Sci. U. S. A., 104*(7), 2543-2547. doi:10.1073/pnas.0607940104

Tsukagoshi, H., Saijo, T., Shibata, D., Morikami, A., & Nakamura, K. (2005). Analysis of a sugar response mutant of Arabidopsis identified a novel B3 domain protein that functions as an active transcriptional repressor. *Plant Physiol., 138*(2), 675-685. doi:10.1104/pp.104.057752

Ueda, M., & Laux, T. (2012). The origin of the plant body axis. *Curr Opin Plant Biol, 15*, 578-584. doi:10.1016/j.pbi.2012.08.001

Ueda, M., Zhang, Z., & Laux, T. (2011). Transcriptional activation of Arabidopsis axis patterning genes WOX8/9 links zygote polarity to embryo development. *Dev. Cell, 20*(2), 264-270. doi:10.1016/j.devcel.2011.01.009

Umezawa, T., Sugiyama, N., Takahashi, F., Anderson, J. C., Ishihama, Y., Peck, S. C., & Shinozaki, K. (2013). Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in *Arabidopsis thaliana*. *Sci Signal, 6*(270), rs8. doi:10.1126/scisignal.2003509

UniProt, C. (2015). UniProt: a hub for protein information. *Nucleic Acids Res, 43*(Database issue), D204-212. doi:10.1093/nar/gku989

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2000). Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc. Natl. Acad. Sci. U. S. A., 97*(21), 11632-11167. doi:10.1073/pnas.190309197

Urano, K., Kurihara, Y., Seki, M., & Shinozaki, K. (2010). 'Omics' analyses of regulatory networks in plant abiotic stress responses. *Curr Opin Plant Biol, 13*, 132-138. doi:S1369-5266(09)00182-4 [pii]

10.1016/j.pbi.2009.12.006

Valadkhan, S., & Jaladat, Y. (2010). The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics, 10*(22), 4128-4141.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., . . . Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods, 5*(9), 829-834.

Varbanova, M., Yamaguchi, S., Yang, Y., McKelvey, K., Hanada, A., Borochov, R., . . . Pichersky, E. (2007). Methylation of gibberellins by Arabidopsis GAMT1 and GAMT2. *Plant Cell, 19*(1), 32-45. doi:10.1105/tpc.106.044602

Veerappan, V., Chen, N., Reichert, A. I., & Allen, R. D. (2014). HSI2/VAL1 PHD-like domain promotes H3K27 trimethylation to repress the expression of seed maturation genes and complex transgenes in Arabidopsis seedlings. *BMC Plant Biol., 14*, 293. doi:10.1186/s12870-014-0293-4

Veerappan, V., Wang, J., Kang, M., Lee, J., Tang, Y., Jha, A. K., . . . Allen, R. D. (2012). A novel *HSI2* mutation in Arabidopsis affects the PHD-like domain and leads to derepression of seed-specific gene expression. *Planta, 236*(1), 1-17. doi:10.1007/s00425-012-1630-1

Verdier, J., Lalanne, D., Pelletier, S., Torres-Jerez, I., Righetti, K., Bandyopadhyay, K., . . . Buitink, J. (2013). A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of *Medicago truncatula* seeds. *Plant Physiol.*

Walley, J., Xiao, Y., Wang, J.-Z., Baidoo, E. E., Keasling, J. D., Shen, Z., . . . Dehesh, K. (2015). Plastid-produced interorgannellar stress signal MEcPP potentiates induction of the unfolded protein response in endoplasmic reticulum. *Proceedings of the National Academy of Sciences, 112*(19), 6212-6217.

Wang, B.-B., & Brendel, V. (2004). The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol, 5*(12), R102.

Wang, B. B., & Brendel, V. (2006). Molecular characterization and phylogeny of U2AF35 homologs in plants. *Plant Physiol., 140*(2), 624-636.

Wang, F., & Perry, S. E. (2013). Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. *Plant Physiol., 161*(3), 1251-1264. doi:10.1104/pp.112.212282

Wang, H., Chung, P. J., Liu, J., Jang, I. C., Kean, M. J., Xu, J., & Chua, N. H. (2014). Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res., 24*(3), 444-453. doi:10.1101/gr.165555.113

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., . . . Perou, C. M. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research*, gkq622.

Wang, L., Hua, D., He, J., Duan, Y., Chen, Z., Hong, X., & Gong, Z. (2011). Auxin Response Factor2 (ARF2) and its regulated homeodomain gene HB33 mediate abscisic acid response in Arabidopsis. *PLoS Genet., 7*(7), e1002172. doi:10.1371/journal.pgen.1002172

Wang, L., Li, P., & Brutnell, T. P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. *Brief Funct Genomics, 9*, 118-128. doi:elp057 [pii]

10.1093/bfgp/elp057

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research, 41*(6), e74-e74.

Wang, Y. X., & Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *J Theor Biol, 362*, 53-61. doi:10.1016/j.jtbi.2014.03.040

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

Warpeha, K. M., Upadhyay, S., Yeh, J., Adamiak, J., Hawkins, S. I., Lapik, Y. R., . . . Kaufman, L. S. (2007). The GCR1, GPA1, PRN1, NF-Y signal chain mediates both blue light and abscisic acid responses in Arabidopsis. *Plant Physiol., 143*(4), 1590-1600. doi:10.1104/pp.106.089904

Watkins, J. F., Sung, P., Prakash, L., & Prakash, S. (1993). The *Saccharomyces cerevisiae* DNA repair gene RAD23 encodes a nuclear protein containing a ubiquitin-like domain required for biological function. *Mol. Cell Biol., 13*(12), 7757-7765.

Werneke, J. M., Chatfield, J. M., & Ogren, W. L. (1989). Alternative messenger-RNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis *Plant Cell, 1*(8), 815-825.

Weselake, R. J., Taylor, D. C., Rahman, M. H., Shah, S., Laroche, A., McVetty, P. B. E., & Harwood, J. L. (2009). Increasing the flow of carbon into seed oil. *Biotechnol Adv, 27*(6), 866-878. doi:10.1016/j.biotechadv.2009.07.001

Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & development, 23*(13), 1494-1504.

Wilusz, J. E., & Wilusz, J. (2014). Nonsense-mediated RNA decay: at the 'cutting edge'of regulated snoRNA production. *Genes & development, 28*(22), 2447-2449.

Wiszniewski, A. A., Smith, S. M., & Bussell, J. D. (2012). Conservation of two lineages of peroxisomal (Type I) 3-ketoacyl-CoA thiolases in land plants, specialization of the genes in *Brassicaceae*, and characterization of their expression in I. *J. Exp. Bot., 63*(17), 6093-6103.

Wolf, J. B. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA‐seq tutorial. *Molecular ecology resources, 13*(4), 559-572.

Wu, M. F., Tian, Q., & Reed, J. W. (2006). Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. *Development, 133*(21), 4211-4218. doi:10.1242/dev.02602

Xia, X. (2012). Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica, 2012*.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., . . . Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res., 42*(D1), D98-D103.

Xing, D., Wang, Y., Hamilton, M., Ben-Hur, A., & Reddy, A. S. (2015). Transcriptome-Wide Identification of RNA Targets of Arabidopsis SERINE/ARGININE-RICH45 Uncovers the Unexpected Roles of This RNA Binding Protein in RNA Processing. *The Plant Cell, 27*(12), 3294-3308.

Xu, F., Xu, S. H., Wiermer, M., Zhang, Y. L., & Li, X. (2012). The cyclin L homolog MOS12 and the MOS4-associated complex are required for the proper splicing of plant resistance genes. *Plant J., 70*(6), 916-928.

Xu, J., & Chua, N. H. (2009). Arabidopsis *decapping 5* is required for mRNA decapping, P-body formation, and translational repression during postembryonic development. *Plant Cell, 21*(10), 3270-3279. doi:10.1105/tpc.109.070078

Yang, C., Bratzel, F., Hohmann, N., Koch, M., Turck, F., & Calonje, M. (2013). VAL- and AtBMI1-mediated H2Aub initiate the switch from embryonic to postgerminative growth in Arabidopsis. *Curr. Biol., 23*(14), 1324-1329. doi:10.1016/j.cub.2013.05.050

Yang, L., Froberg, J. E., & Lee, J. T. (2014). Long noncoding RNAs: fresh perspectives into the RNA world. *Trends in biochemical sciences, 39*(1), 35-43.

Yang, Z., & Ohlrogge, J. B. (2009). Turnover of fatty acids during natural senescence of Arabidopsis, Brachypodium, and switchgrass and in Arabidopsis beta-oxidation mutants. *Plant Physiol., 150*(4), 1981-1989.

Yazawa, K., & Kamada, H. (2007). Identification and characterization of carrot HAP factors that form a complex with the embryo-specific transcription factor C-LEC1. *J. Exp. Bot., 58*(13), 3819-2388. doi:10.1093/jxb/erm238

Yoshida, T., Fujita, Y., Maruyama, K., Mogami, J., Todaka, D., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2015). Four Arabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. *Plant Cell Environ., 38*(1), 35-49. doi:10.1111/pce.12351

Zhang, S., Jin, G., Zhang, X. S., & Chen, L. (2007). Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics, 7*(16), 2856-2869. doi:10.1002/pmic.200700095

Zhang, W., Han, Z., Guo, Q., Liu, Y., Zheng, Y., Wu, F., & Jin, W. (2014). Identification of Maize Long Non-Coding RNAs Responsive to Drought Stress. *PloS one, 9*(6), e98958.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., & Jacobsen, S. E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol., 5*(5), e129. doi:10.1371/journal.pbio.0050129

Zhang, X., Lu, X., Shi, Q., Xu, X. Q., Leung, H. C., Harris, L. N., . . . Wong, W. H. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics, 7*, 197. doi:10.1186/1471-2105-7-197

Zhang, X. N., & Mount, S. M. (2009). Two Alternatively Spliced Isoforms of the Arabidopsis SR45 Protein Have Distinct Roles during Normal Plant Development. *Plant Physiol., 150*(3), 1450-1458. doi:10.1104/pp.109.138180

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137.

Zhang, Y., Mayba, O., Pfeiffer, A., Shi, H., Tepperman, J. M., Speed, T. P., & Quail, P. H. (2013). A quartet of PIF bHLH factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in Arabidopsis. *PLoS genetics, 9*(1), e1003244.

Zhou, Y., Tan, B., Luo, M., Li, Y., Liu, C., Chen, C., . . . Huang, S. (2013). HISTONE DEACETYLASE19 interacts with HSL1 and participates in the repression of seed maturation genes in Arabidopsis seedlings. *Plant Cell, 25*(1), 134-148. doi:10.1105/tpc.112.096313

Zhu, S. Y., Yu, X. C., Wang, X. J., Zhao, R., Li, Y., Fan, R. C., . . . Zhang, D. P. (2007). Two calcium-dependent protein kinases, CPK4 and CPK11, regulate abscisic acid signal transduction in Arabidopsis. *Plant Cell, 19*(10), 3019-3036. doi:10.1105/tpc.107.050666

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C. T., Bader, G. D., & Morris, Q. (2013). GeneMANIA prediction server 2013 update. *Nucleic Acids Res., 41*(Web Server issue), W115-W122. doi:10.1093/nar/gkt533