

MicroRNA-centered regulatory gene network in
Arabidopsis

Xin Zhao

Beijing, China

B.S., Beihang University, 2010

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy
for the Degree of Doctor of Philosophy

Department of Biology

University of Virginia

December, 2014

Abstract

MicroRNAs (miRNAs) are a class of sequence-specific, trans-acting small RNA molecules that regulate gene expression at the post-transcription level in both plants and animals. They impact a substantial portion of the transcriptome and are required for many developmental processes and responses to environmental challenges. Although a number of miRNA-mediated gene circuits have been studied in detail in the model plant *Arabidopsis thaliana*, miRNA as a class is underrepresented in genome scale studies. Consequently, current transcriptional regulatory networks (TRNs) in *Arabidopsis* are confined to transcription factors (TFs) and their target genes. This limitation prompted me to incorporate miRNAs into the TRNs to extend our understanding on gene regulation. In this project, I mapped the interactions among TFs, miRNAs, and their target genes to construct the first miRNA-centered regulatory gene network in *Arabidopsis*.

At the time when this project was initiated, the promoter structure remained unknown for most plant miRNAs, which are transcribed by RNA Polymerase II (Pol II). Therefore, the first portion of my project focused on the identification and analysis of the proximal promoter of *Arabidopsis* miRNAs. I analyzed genome wide Pol II chromatin immunoprecipitation (ChIP) data and discovered unique Pol II binding pattern at miRNAs loci. Next, I developed a pattern-based promoter prediction method, which allowed precise prediction of transcription start sites (TSSs) and promoter regions for 167 miRNA genes in *Arabidopsis*. Thus, this work helped to elucidate how miRNAs are regulated by TFs and provided the fundamental building blocks for the miRNA-centered regulatory network.

Given that miRNAs function post-transcriptionally, obtaining a comprehensive set of miRNA targets genes in *Arabidopsis* will greatly facilitate the study of this important class of regulator. By integrating results from computational prediction, degradome sequencing analysis, and previous experiments, I identified 2189 miRNA-target interactions (MTIs) among 1381 target genes for 264 miRNAs. By processing genome-wide binding information for 35 TFs, I systematically identified transcriptional regulation and eventually constructed a miRNA-centered regulatory gene network in *Arabidopsis*, which consists of 1701 genes (TFs, miRNAs, and miRNA targets genes) and 6424 transcriptional and post-transcriptional regulations.

Topological analysis revealed that miRNAs act similar as TFs, preferentially connecting together more TFs than other genes. I demonstrated the expression of miRNAs is highly dependent on the combination of upstream TFs, suggesting that expression dynamics of miRNAs serves as a signal integration mechanism to facilitate crosstalk between different TFs and hence the biological pathways they regulate. I extracted the sub-network of miRNAs regulated by six different light-signaling TFs and found highly overlapped regulations by TFs functioning in circadian clock, flower development and polarity identity determination. These findings revealed potential heavy crosstalk mediated by miRNAs and provided new directions for molecular and genetic studies.

As a sample to functionally study miRNA in a network background, I collaborated with other researchers to investigate miRNAs regulated by two TFs: *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 7 (SPL7)* and *ELONGATED HYPOCOTYL 5 (HY5)*, which mediate copper and light signaling, respectively. Through genome-wide ChIP-seq analysis,

I elucidated the *SPL7* regulon. By comparing it with that of *HY5*, I found that *SPL7* and *HY5* act coordinately to transcriptionally regulate *MIR408*, which results in differential expression of miR408 as well as its target genes in response to changing light and copper conditions. We demonstrate that this regulation is tied to copper allocation to the chloroplast and plastocyanin level. Finally, we found that constitutively activated miR408 rescues distinct developmental defects of the *hy5*, *spl7*, and *hy5 spl7* mutants. These findings revealed a previously uncharacterized light-copper crosstalk mediated by a *HY5-SPL7-MIR408* network.

In summary, I constructed a miRNA-centered network in *Arabidopsis* incorporating both transcriptional and post-transcriptional regulations. Analysis of this network has provided insights into the design principles and specific gene circuits. These results should be instrumental to functional studies aiming at elucidating how miRNAs function in the context of regulatory networks. Such knowledge should provide further mechanistic insight into highly coordinated and adaptable control of gene batteries that underpin plant development and responses to environmental challenges.

Acknowledgements

Pursuing a Ph.D. in biology at the University of Virginia has been the most challenging and bold decision I have made, but only during this process of surviving and thriving, I had the most unforgettable and memorable four and half years of my life. It's never too much to express my sincere gratitude to the people who have always encouraged and supported me. This work wouldn't have been accomplished without their help.

My sincerest thanks go to my advisor and mentor, Dr. Lei Li. I want to thank him for opening the door of scientific research for me, in which I found both joy and frustrations, but more importantly I found the motivation to become a better me. Throughout the entire process, Lei has offered his best patience and belief in me, and became one of the most respectable and reliable people with whom that I like to share everything. Lei's creative ideas and critical thinking gave birth to this project, while his crave for the unknown and rigorous requirement provided the foundation and direction of this work. I feel very fortunate to be a part of Lei's lab, and will always be grateful for what I've learnt from Lei.

I would also like to express my greatest gratitude to my dissertation committee: Dr. Martin Wu, Dr. Aaron J. Mackey, Dr. Christopher D. Deppmann, Dr. Keith G. Kozminski, and Dr. Stefan Bekiranov. I appreciate Martin for a great rotation experience, and for serving as my first reader. His advices on my proposal and dissertation helped me realize several potential problems of this dissertation in both writing and experimental design. I want thank Dr. Keith G. Kozminski and Dr. Aaron J. Mackey for their thoughtful ideas on understanding miRNA promoter structures and handling noisy ChIP-seq data, which helped me

established the story of my first and third publications. I appreciate Dr. Stefan Bekiranov for accepting to serve on my committee without any hesitation, and sharply pointed out all the possible bioinformatics analysis I can do to improve my project. I want to thank Dr. Christopher D. Deppmann for the critical questions on transcription factors which forced me to re-establish my TF knowledgebase and re-evaluate my TF part of the network. Moreover, for the relaxed atmosphere he brought to every committee meeting. It's my honor to have such a great committee with rigorous standard but supportive advices. I couldn't have made this without their kindness and encourage.

Meanwhile, I am also grateful to all my past colleagues, labmates, and collaborators, Dr. Chengjun Wu, Dr. Huiyong Zhang, Dr. Xiaozeng Yang, Chun Su, Dr. Alexander F. Koepfel, Dr. Zhang Wang, Tiantian Ren, Dr. Yimiao Tang, Dr. Xiangzheng Liao, and Dr. Shiqing Gao. They always offered me their sincerest help and suggestions, and made it a wonderful time to work with them. My specialty in computer science meant collaboration is the most indispensable part of this project. Therefore, I want to especially thank Dr. Huiyong Zhang and Dr. Yimiao Tang for their innovative ideas in data analysis and rich experiences in experimental designs, without which I wouldn't have finished this work.

The Department of Biology and UVa has always been my strong backing by offering me teaching assistantship and dissertation fellowship. I felt lucky to have three great DGSs, Dr. Keith G. Kozminski, Dr. Dorothy Schafer, and Dr. Barry Condron, who were always ready to answer my questions and figure out a solution. I have to appreciate Dr. Mark Kopeny, Dr. David Kittlesen, and lovely Ms. Joanne Chaplin and Ms. Kay Christopher for the unforgettable three-year TA experience.

At last but not least, I wish to express my sincerest wishes and deepest gratitude to my parents for their undivided support and interest, which has inspired me and encouraged me throughout my entire life. It's never too much to express my love for them, and tell them how lucky I was to be their son. They are the only reason that made all the things possible. This work is dedicated to them.

Table of Contents

Abstract	I
Acknowledgements	IV
Table of Contents.....	VII
Chapter 1. Introduction	1
Biogenesis of miRNAs.....	2
Regulatory roles of miRNAs.....	7
Regulatory gene networks.....	10
A microRNA-centered gene network.....	12
References	17
Chapter 2. Identification and Analysis of the Proximal Promoters of microRNA Genes in <i>Arabidopsis</i>	31
Abstract	32
Introduction.....	33
Materials and methods	37
Results.....	42
Discussion.....	49
References.....	53
Supplementary materials.....	71
Chapter 3. Comparative Analysis of microRNA Promoters in <i>Arabidopsis</i> and Rice.....	78
Abstract	79
Introduction.....	80
Results and discussion.....	82
Methods.....	87
References.....	90
Chapter 4. Construction and Analysis of a microRNA-centered Regulatory Gene Network in <i>Arabidopsis</i>	98
Abstract	99
Introduction.....	101
Results.....	105
Discussion.....	113
Materials and methods	116
References.....	121
Supplemental materials	139
Chapter 5. microRNA408 Is Critical for the HY5-SPL7 Gene Network That Mediates Coordinated Response to Light and Copper.....	141
Abstract	142
Introduction.....	143
Results.....	146
Discussion.....	164
Methods.....	173
References.....	182
Supplementary materials.....	215

Chapter 1. Introduction

Biogenesis of miRNAs

Small RNAs (sRNAs), which are a regulatory class of RNAs that appear in all organisms, including eukaryotes, bacteria, and Archaea, and some viruses, are mainly represented by the 21 to 24 nucleotides. The study of the biosynthesis and function of small regulatory RNAs is relatively new, but it has already revolutionized how we think about genetics and molecular biology. The 2006 Nobel Prize for Physiology or Medicine was awarded to Craig Mello and Andrew Fire for their discovery that double-stranded RNA serves as an intermediate in RNA-based gene regulation. We now appreciate that many different types of cellular sRNAs are processed by the Dicer family of RNase III enzymes (Dicer-like or DCL in plant) in double-stranded RNAs or self-complementary RNA duplexes. Via these sRNAs, transcriptional and post-transcriptional regulations collaborate to generate sophisticated temporal and spatial gene expression dynamics. The identification and functional analysis of sRNAs have tremendously expanded our capacities to understand complex biological phenomena.

In plants, microRNAs (miRNAs) represent a major class of trans-acting, sequence-specific endogenous sRNAs that modulate the expression of target genes at the post-transcriptional level. Most plants possess hundreds of miRNA genes (*MIR*) (Nozawa et al., 2012), which are mainly found in the intergenic regions of the genome (Reinhart et al., 2002). Research has unveiled the mysteries behind miRNA biogenesis, which is conventionally divided into four major steps: *MIR* transcription, pri-miRNA processing, nuclear export, and the assembly of the RNA-induced silencing complex (RISC) (Figure 1).

Similar to protein-coding genes, the vast majority of plant *MIRs* exist as independent transcriptional units (Griffiths-Jones et al., 2008) and are transcribed by RNA polymerase II (Pol II) to produce the primary transcripts, pri-miRNAs (Kim et al., 2011; Xie et al., 2005). Once transcribed, the 5' 7-methylguanosine cap and 3' polyadenylate tail are added to the pri-miRNA in the nucleus for stabilization (Jones-Rhoades and Bartel, 2004b; Xie et al., 2005; Zhang et al., 2005). In sequence, features of the stem-loop structure of pri-miRNAs cause the initial cleavage by DCL (Margis et al., 2006) near the stem to generate precursor miRNA (pre-miRNA) (Bologna et al., 2009; Cuperus et al., 2010b; Mateos et al., 2010; Song et al., 2010; Werner et al., 2010). Similarly directed by DCL, the subsequent cleavage of pre-miRNA along the stem may yield one or more complementary miRNA/miRNA* duplexes (Liu et al., 2012). Four members of the DCL family correspond to the distinct sizes of the small RNAs that they generate. 21-nucleotide miRNAs, predominately enriched in plants (Chen et al., 2010; Cuperus et al., 2010a), are processed by DCL1 and DCL4, while 22-nucleotide miRNAs are generated from DCL2, and 24-nucleotide miRNAs are generated from DCL3 (Akbergenov et al., 2006; Cuperus et al., 2010a; Deleris et al., 2006; Xie et al., 2005; Xie et al., 2004). In addition, a complex scenario in which multiple DCLs sequentially process the same pri-miRNA has been elucidated (Wu et al., 2010). RNA binding proteins (e.g. TOUGH [TGH], SERRATE [SE], and HYL1), phospho-regulation proteins (e.g. C-TERMINAL DOMAIN PHOSPHATASE-LIKE1 [CPL1], and DAWDLE [DDL]), and others (SICKLE [SIC], MODIFIER OF SNC1,2 [MOS2]) are also believed crucial in DCL recruitment or in mediating accurate miRNA processing (Fang and Spector, 2007; Kurihara et al., 2006; Lobbes et al., 2006; Machida et al., 2011; Manavella et al.,

2012; Qin et al., 2010; Ren et al., 2012b; Yang et al., 2006; Yu et al., 2008; Zhan et al., 2012).

In animals, Exportin 5 is responsible for the transportation of pre-miRNAs to cytoplasm before being cleaved into mature miRNAs (Zeng and Cullen, 2004). In the model plant *Arabidopsis thaliana*, the miRNA/miRNA* duplexes are first methylated by HEN1 (Yu et al., 2008) and then are exported by the Exportin 5 homolog HASTY (HST) (Ren et al., 2012b). Previous research has revealed that HST affects miRNA accumulation in a tissue type and *MIR* gene-dependent manner (Park et al., 2005), indicating the existence of as yet unknown export mechanisms of miRNA. The interaction between miRNA and its target is mediated by RISC that contains ARGONAUTE (AGO). Once exported, the less thermodynamically stable 5'-end of the guide strand (miRNA) facilitates the correct strand selection of the RISC (Eamens et al., 2009). The removal of the passenger strand (miRNA*) requires the disassociation of specific AGO1-associated proteins, such as *HSP90*, and SQUINT (SQN) (Iki et al., 2012), which might then trigger the conformational changes of *AGO1*. Moreover, the cooperative roles of SMALL RNA DEGRADING NUCLEASE1 (SDN1) and HEN1 SUPPRESSOR1 (HESO1) in degrading HEN1 methylated miRNAs have demonstrated the additional accumulation-limiting mechanism of miRNAs (Ramachandran and Chen, 2008; Ren et al., 2012a; Zhao et al., 2012). Eventually, the miRNAs loaded in RISC serve as templates for recognizing complementary sequences within their target mRNAs and repress them through direct target cleavage (Llave et al., 2002; Reinhart et al., 2002), translational inhibition (Beauclair et al., 2010; Brodersen et al., 2008), DNA

methylation (Wu et al., 2010), and possibly mRNA destabilization (Rogers and Chen, 2013).

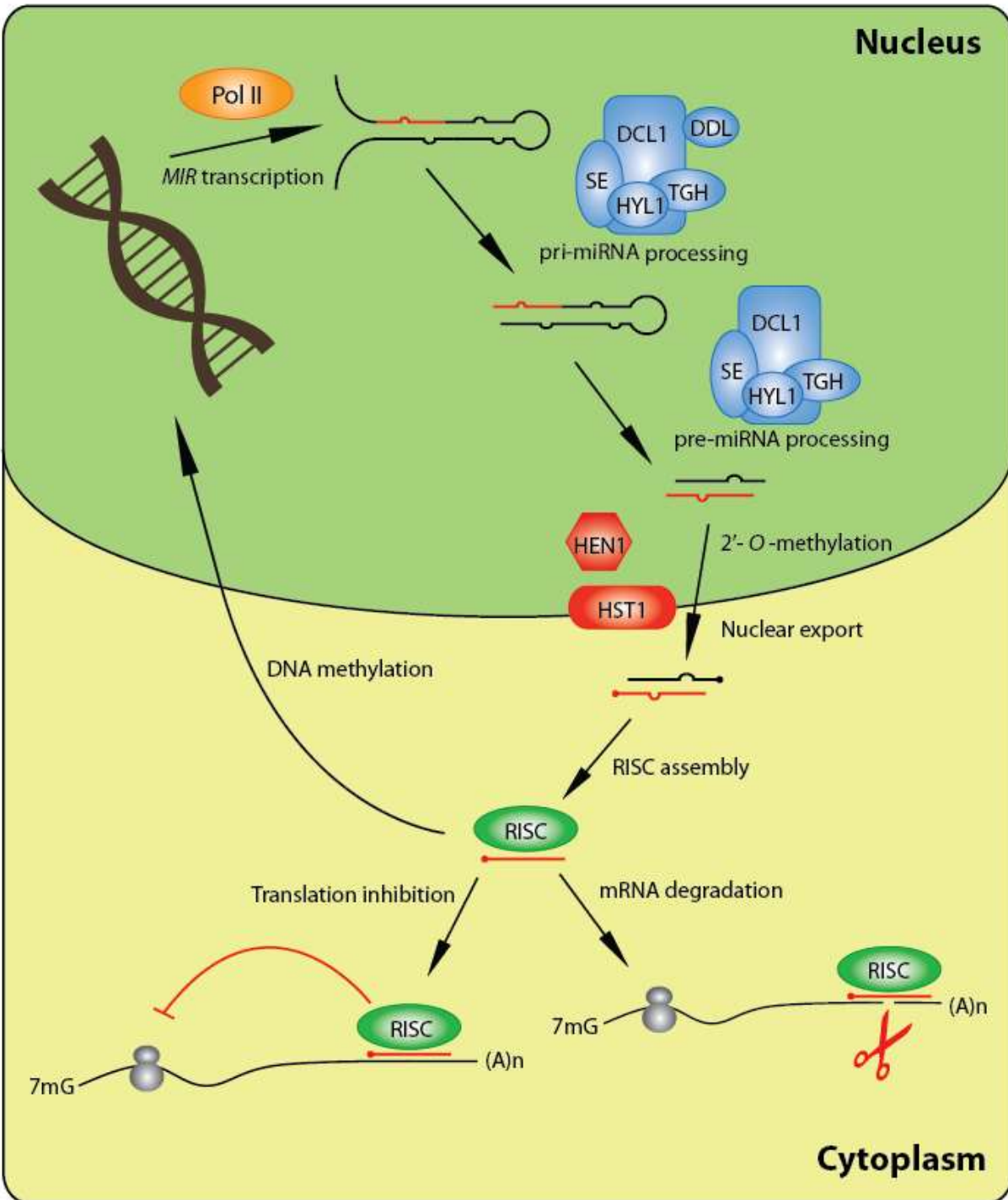


Figure1. miRNA biogenesis in plants

MIR genes are transcribed by RNA polymerase II (Pol II) to produce primary transcripts termed pri-miRNAs. Two sequential cleavages mediated by DCL1 yield one or several miRNA:miRNA* duplexes. Before transport out of the nucleus by HST1, 3' overhangs of the duplex are methylated by HEN1 for stabilization purpose. Once exported, the guide strand is selected and incorporated into the RISC complex which can subsequently direct target mRNA cleavage, translation inhibition, and DNA methylation.

Regulatory roles of miRNAs

Unlike animal miRNAs, which normally regulate hundreds of target genes through relaxed sequence complementarity in the 3'UTR, the miRNAs in plants bind to the coding regions of mRNAs with a much higher sequence complementarity and tend to have only a handful of targets (Jones-Rhoades and Bartel, 2004a; Rhoades et al., 2002). However, it has been shown that the target genes of plant miRNAs often share a regulatory function, which places miRNAs in a pivotal position in the gene regulation programs in plants. Previous research has revealed the participation of miRNAs in governing plant behavior throughout the life cycle, as well as involvement in stimulus and stress responses (Rubio-Somoza and Weigel, 2011). For example, the antagonistic activities of miR156 and miR172 in *Arabidopsis* facilitate the progression through different developmental phases (Wang et al., 2009; Wu et al., 2009). miR164 and miR319 were found to be involved in the control of senescence (Kim et al., 2009; Schommer et al., 2008) and miR165/166 and miR390-TAS3 were demonstrated to be intimately associated with plant's abaxial/adaxial polarity. Moreover, our recent study of *MIR408* demonstrated its critical role in regulating seedling development under changing light and copper conditions by targeting the transcripts of copper-containing proteins (Zhang and Li, 2013).

The recognition and validation of target genes have always been the very first step towards the full understanding of miRNA function. The transient cleavage of target mRNAs as well as the efficient degradation of cleavage products makes it difficult to identify miRNA-target pairs *in vivo*. Thus, numerous target prediction approaches have been developed to

depict the post-transcriptional regulation of miRNAs. The most commonly applied methods have taken advantage of the stringent base pairing between miRNAs and target mRNAs and filter potential miRNA-target pairs in a score-based manner (Dai and Zhao, 2011). Other methods used to predict the target genes have considered the minimal free energy of miRNA-target hybridization (Bonnet et al., 2010), evolutionary conservation of miRNA target sites, and 5'-end uncapped cleavage products of target mRNAs (Addo-Quaye et al., 2008; German et al., 2008). However, a recent analysis unveiled problems in the existing methods (Ding et al., 2012) and suggested a potential improvement by combining multiple methods. Taken together, these findings suggest that the current methods for genome-wide miRNA target identification are inadequate.

Despite the wealth of knowledge about miRNA biogenesis, the transcriptional regulation of *MIR* genes has not been fully elucidated. The 5' cap and Poly (A) tail of pri-miRNA transcript suggest regulatory properties similar to those of protein coding genes. The TATA box core promoter element and other known *cis*-regulatory motifs were found to be overrepresented in experimentally validated *MIR* promoter sequences (Megraw et al., 2006; Xie et al., 2005). In *Arabidopsis*, however, the transcriptional start sites (TSSs) of only a handful of *MIR* genes were experimentally identified (Xie et al., 2005). Although an increasing number of miRNAs have been retrieved from fast-growing deep sequencing (Yang and Li, 2011), the current deficit in the knowledge about *MIR* TSSs prevents the identification of the promoter regions of the *MIR* gene, and hence hinders the comprehensive study of the transcription factor (TF) recruitment and its transcriptional regulation.

TF is a class of proteins that activate or repress the expression of genes by binding to the proximal promoter and enhancer regions of genes. TFs usually recognize a few specific sequence motifs that vary from 6 to 15 nucleotides in length (Stormo, 2000) and function in singular, collaborative or competitive modes (Umetani et al., 2001). More than 2000 TF genes have been identified in *Arabidopsis* (Riechmann et al., 2000). Much progress has been made in the genome-wide identification of the binding sites of TFs since the advancement of microarray and sequencing technologies, which permit the high throughput analysis of large-scale chromatin immunoprecipitation data (hereafter referred to as ChIP-chip and ChIP-seq) in diverse cell types and conditions. Many computational methods have been developed to identify the binding sites of TFs with known TFBSs. The most common method of identifying TFBSs uses a position-specific scoring matrix (PSSM). Combined with ChIP, the identification of confident TFBSs in miRNA promoters will be a fundamental contribution to the understanding of the transcriptional regulation of *MIR* genes.

Recent research has begun to characterize the regulation mechanisms of *MIR* transcriptions and their biological functions. For example, *MIR398b* and *c* are regulated by *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE7 (SPL7)* to regulate copper homeostasis in response to low copper conditions in *Arabidopsis* (Yamasaki et al., 2009). *MIR172* has been shown to be bound by *SHORT VEGETATIVE PHASE (SVP)*, revealing its potential temperature-sensitive regulatory function (Cho et al., 2012). *FUSCA3* directly binds to *MIR156a* and *c*, and regulates the accumulation of different *MIR156* family members, which in turn control TFs involved in developmental phase changes (Wang and Perry,

2013). APETALA2 is recruited by members of the *MIR156* and *MIR172* families and accommodates the flowering of *Arabidopsis* by activating *MIR156* and repressing *MIR172* (Yant et al., 2010). However, because there are over three hundred identified *MIR* genes in *Arabidopsis*, many more regulatory relationships between TFs and miRNAs await discovery, as particularly indicated by the fast-expanding collection of ChIP-chip/seq data of *Arabidopsis* TFs.

Regulatory gene networks

Reductionism, which focuses on individual cellular components and their functions, has achieved unparalleled success in biological research for the past century (Barabasi and Oltvai, 2004). No longer satisfied with defining single cellular constituent, researchers now attempt to determine how they interact with each other and contribute to an organism's continuous adaptation to changing environmental conditions. The development of systems biology, as well as the latest high throughput screening/sequencing technologies, has made it possible to study complex gene clusters at the network level. The rapid progress in the study of transcription regulatory networks (TRNs) in the past decade has enabled the elucidation of the regulation of protein-coding genes in *Saccharomyces cerevisiae* (Harbison et al., 2004), *Caenorhabditis elegans* (Deplancke et al., 2006; Vermeirssen et al., 2007), *Drosophila melanogaster* (Sandmann et al., 2007), and mammals (Boyer et al., 2005; Carro et al., 2010; Li et al., 2007). Despite that further improvement is needed, these TRNs have already provided considerable insight into the regulation of gene expression in

response to changing environments and the evolutionary origins that lead to the formation of these networks.

TRNs, at a highly abstract level, are usually depicted as a series of nodes connected by edges (Bolouri and Davidson, 2002; Lee et al., 2002; Maslov and Sneppen, 2002; Milo et al., 2002; Shen-Orr et al., 2002; Thieffry et al., 1998). Nodes are individual genes, whereas directed edges represent transcriptional regulations between TFs and their target genes (Mangan and Alon, 2003). Recent research has revealed simple principles that are common to the architecture of most networks (Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2003). Unlike in a random network, where the node degrees (i.e., the number of edges linked to it) follow an elegant Poisson distribution in which most nodes share roughly the same degree, the degrees in real networks are characterized by a highly asymmetrical power-law distribution (Barabasi and Albert, 1999). A network with this feature is scale-free (Bollobas, 1985). The existence of a few nodes that make many more connections with others (i.e., “hubs”) are believed to indicate a hierarchical structure, whereby hubs serve as major regulators in controlling biological processes, while the majority of genes with fewer connections protect the networks from random perturbations, making them more robust (Blais and Dynlacht, 2005).

First found in the bacterium *Escherichia coli* (Milo et al., 2002; Shen-Orr et al., 2002) and the yeast *Saccharomyces cerevisiae* (Lee et al., 2002), the overrepresentation of motifs in TRNs is another characteristic that distinguishes a real network from a random one. Defined as recurrent and statistically significant sub-graphs or patterns, network motifs, with their distinct functions in modulating gene expression, help us understand how

regulatory networks manage to complete different tasks. For instance, auto-regulation is able to adjust the response time of gene circuits (Alon, 2007) and alter cell-cell variations in protein levels (Kaern et al., 2005). Single-input modules can generate a temporal order program in a group of genes with shared function (Kalir et al., 2001; Laub et al., 2000; McAdams and Shapiro, 2003; Ronen et al., 2002; Spellman et al., 1998; Zaslaver et al., 2004). Feedback loops play an important role in developmental networks by making irreversible decisions that transduce signals into cell-fate decisions (Alon, 2007; Davidson et al., 2002; Levine and Davidson, 2005; Longabaugh et al., 2005). In addition, feed-forward loops induce stable transcriptional responses and filter out noise (Alon, 2007; Mangan and Alon, 2003). Each real network is depicted by a unique set of distinct motifs; however, recent studies in yeast and *E. coli* showed duplicate regulatory genes are randomly distributed across motif types, thus motifs do not share common ancestry. Therefore, the convergent evolution towards the same motif types indicates the underlying biological relevance of network motifs (Conant and Wagner, 2003; Hinman et al., 2003).

A microRNA-centered gene network

Recent studies have demonstrated that miRNAs play a pivotal role in the metabolism and development of plants. Although the function of individual miRNAs has been heavily studied over the years, advancements in theories and techniques of systems biology have made it possible to decipher miRNA functionality in the context of a gene network. Therefore, the purpose of my dissertation is to develop an understanding of the regulatory

function of miRNA genes by constructing and analyzing a miRNA-centered regulatory gene network in *Arabidopsis*.

In this dissertation, the miRNA-centered regulatory gene network is defined as consisting of three components—miRNAs, their upstream TFs, and downstream target genes—and three interactions—TF-miRNA interactions (TMIs), miRNA-target interactions (MTIs), and TF-miRNA target interactions (TTIs). Chapter 2 of this dissertation describes an attempt to identify pri-miRNAs, which are difficult to characterize because of their transient nature. In collaboration with Dr. Huiyong Zhang (College of Life Sciences, Henan Agricultural University), we designed and performed a ChIP-chip experiment for *Arabidopsis thaliana* in order to profile systematically the RNA Pol II binding features at *MIR* loci. By characterizing the unique Pol II binding pattern and comparing it with protein coding genes, I developed a novel computational prediction method to pinpoint the TSSs of 167 *MIR* genes, which represents 87% of all intergenic *Arabidopsis MIR* genes. The predictions were validated by experimentally supported TSS annotation and average free energy (AFE) profiling near TSSs. Furthermore, by implementing position weight matrices (PWM) for 99 *cis*-elements, I systematically scanned the regulatory regions upstream of the TSSs and discovered eleven and ten *cis*-elements that were statistically over- and under-represented in *MIR* promoters, respectively. The identification of *Arabidopsis MIR* TSSs yielded insight into the promoter structure of plant *MIR* and provided the foundation for analyzing TMIs.

The results of my study of *MIR* promoters in *Arabidopsis* motivated me to extend this knowledge to other plant species. As described in Chapter 3, by analyzing full-length

cDNA sequences related to miRNAs in rice, I mapped TSSs for 62 and 55 miRNAs in *Arabidopsis* and rice, respectively. The AFE profiles in the vicinity of TSSs and over-represented *cis*-elements were studied in both species. This work revealed structural differences between *Arabidopsis* and rice miRNA promoters, thus providing a new perspective on comparing miRNA structure and studying miRNA regulation in plants.

Chapter 4 of this dissertation describes the construction and subsequent analysis of miRNA-centered regulatory gene network in *Arabidopsis*. First, experimentally validated MTIs, computational target prediction methods, and degradome sequencing data analysis were integrated to yield 2189 high-confident MTIs. Whole genome ChIP data for 35 TFs were analyzed to systematically determine the transcriptional regulatory interactions that link TFs to miRNAs and miRNA targets. The results that 900 transcriptional regulatory interactions between 35 TFs and 218 miRNAs (TMIs), and 3351 interactions between 35 TFs and 951 miRNA targets (TTIs) were gathered, respectively. Hence, a hierarchical network centered on *Arabidopsis* miRNAs that consists of 1701 genes and 6424 interactions was then built. The topology analysis of the network suggested that connectivity profiles of miRNAs and TFs were similar and revealed the over-representation of the feed-forward loop (FFL), in which *MIR* genes may play a crucial role. In addition, combinatorial regulation at *MIR* loci was observed, indicating the central role of miRNAs in facilitating crosstalk among diverse TFs. Finally, a *MIR408*-centered sub-network was extracted and used as the sample in an experimental investigation of miRNA function in the network context.

In Chapter 5, a work once again collaborated with Dr. Zhang is described, in which the *MIR408* sub-network was functionally examined to reveal a previously uncharacterized light-copper crosstalk. Dr. Zhang first demonstrate in *Arabidopsis* an interaction between the two TFs, *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 7 (SPL7)* and *ELONGATED HYPOCOTYL 5 (HY5)*, which mediates copper and light signaling, respectively. By conducting whole genome ChIP-sequencing and RNA-sequencing analyses, I elucidated the *SPL7* regulon and compared it with that of *HY5*. I found that the two TFs co-regulate many genes, such as those involved in anthocyanin accumulation and photosynthesis. Specifically, we showed that *SPL7* and *HY5* act in coordination to regulate *MIR408* transcriptionally, which resulted in a differential expression of miR408 as well as its target genes, in response to changing light and copper conditions. Finally, Dr. Zhang found that constitutively activated miR408 rescued distinct developmental defects of the *hy5*, *spl7*, and *hy5 spl7* mutants. These findings revealed a previously uncharacterized light-copper crosstalk mediated by a *HY5-SPL7-MIR408* network. The integration of transcriptional and post-transcriptional regulations is critical for governing proper metabolism and development in response to combined copper and light signaling.

In summary, my dissertation research constructed the first miRNA-centered regulatory gene network in *Arabidopsis*. Moreover, by extracting a light signaling-related sub-network and specifically studying the regulatory function of *MIR408*, my dissertation has built a working model for the functional analysis of plant miRNAs. Extension of this research method to other miRNA-centered sub-networks will help us understand previously uncharacterized functions of plant miRNAs, provide further insight into the mechanics of

the highly coordinated and adaptable control of gene batteries that underpins the development and responses of plants to environmental challenges, and ultimately build a knowledge base for genetically modifying plant, which can assist us in facing the challenges of global climate change and food scarcity.

References

Addo-Quaye, C., Eshoo, T.W., Bartel, D.P., and Axtell, M.J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol* 18, 758-762.

Akbergenov, R., Si-Ammour, A., Blevins, T., Amin, I., Kutter, C., Vanderschuren, H., Zhang, P., Gruissem, W., Meins, F., Jr., Hohn, T., *et al.* (2006). Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Res* 34, 462-471.

Albert, R., and Barabasi, A.L. (2002). Statistical mechanics of complex networks. *Rev Mod Phys* 74, 47-97.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet* 8, 450-461.

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509-512.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101-113.

Beauclair, L., Yu, A., and Bouche, N. (2010). microRNA-directed cleavage and translational repression of the copper chaperone for superoxide dismutase mRNA in Arabidopsis. *Plant J* 62, 454-462.

Blais, A., and Dynlacht, B.D. (2005). Constructing transcriptional regulatory networks. *Gene Dev* 19, 1499-1511.

Bollobas, B.I. (1985). *Random graphs* (London, Academic Press).

Bologna, N.G., Mateos, J.L., Bresso, E.G., and Palatnik, J.F. (2009). A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *EMBO Journal* 28, 3646-3656.

Bolouri, H., and Davidson, E.H. (2002). Modeling transcriptional regulatory networks. *Bioessays* 24, 1118-1129.

Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* 26, 1566-1568.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.R., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947-956.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science* 320, 1185-1190.

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X.D., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., *et al.* (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* *463*, 318-U368.

Chen, H.M., Chen, L.T., Patel, K., Li, Y.H., Baulcombe, D.C., and Wu, S.H. (2010). 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc Natl Acad Sci U S A* *107*, 15269-15274.

Cho, H.J., Kim, J.J., Lee, J.H., Kim, W., Jung, J.H., Park, C.M., and Ahn, J.H. (2012). SHORT VEGETATIVE PHASE (SVP) protein negatively regulates miR172 transcription via direct binding to the pri-miR172a promoter in Arabidopsis. *FEBS Lett* *586*, 2332-2337.

Conant, G.C., and Wagner, A. (2003). Convergent evolution of gene circuits. *Nat Genet* *34*, 264-266.

Cuperus, J.T., Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Burke, R.T., Takeda, A., Sullivan, C.M., Gilbert, S.D., Montgomery, T.A., and Carrington, J.C. (2010a). Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat Struct Mol Biol* *17*, 997-1003.

Cuperus, J.T., Montgomery, T.A., Fahlgren, N., Burke, R.T., Townsend, T., Sullivan, C.M., and Carrington, J.C. (2010b). Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. *Proc Natl Acad Sci U S A* *107*, 466-471.

Dai, X., and Zhao, P.X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 39, W155-159.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., *et al.* (2002). A genomic regulatory network for development. *Science* 295, 1669-1678.

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K.D., Carrington, J.C., and Voinnet, O. (2006). Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. *Science* 313, 68-71.

Deplancke, B., Mukhopadhyay, A., Ao, W.Y., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S., Hope, I.A., *et al.* (2006). A gene-centered *C. elegans* protein-DNA interaction network. *Cell* 125, 1193-1205.

Ding, J., Li, D., Ohler, U., Guan, J., and Zhou, S. (2012). Genome-wide search for miRNA-target interactions in *Arabidopsis thaliana* with an integrated approach. *BMC Genomics* 13 *Suppl 3*, S3.

Dorogovtsev, S.N., and Mendes, J.F.F. (2003). *Evolution of networks : from biological nets to the Internet and WWW* (Oxford, Oxford University Press).

Eamens, A.L., Smith, N.A., Curtin, S.J., Wang, M.B., and Waterhouse, P.M. (2009). The *Arabidopsis thaliana* double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA* 15, 2219-2235.

Fang, Y., and Spector, D.L. (2007). Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living Arabidopsis plants. *Curr Biol* 17, 818-823.

German, M.A., Pillay, M., Jeong, D.H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta, K., German, R., *et al.* (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* 26, 941-946.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36, D154-158.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99-104.

Hinman, V.F., Nguyen, A.T., Cameron, R.A., and Davidson, E.H. (2003). Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc Natl Acad Sci U S A* 100, 13356-13361.

Iki, T., Yoshikawa, M., Meshi, T., and Ishikawa, M. (2012). Cyclophilin 40 facilitates HSP90-mediated RISC assembly in plants. *EMBO J* 31, 267-278.

Jones-Rhoades, M.W., and Bartel, D.P. (2004a). Computational identification of plant MicroRNAs and their targets, including a stress-induced miRNA. *Molecular Cell* 14, 787-799.

Jones-Rhoades, M.W., and Bartel, D.P. (2004b). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* 14, 787-799.

Kaern, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* 6, 451-464.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., and Alon, U. (2001). Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* 292, 2080-2083.

Kim, J.H., Woo, H.R., Kim, J., Lim, P.O., Lee, I.C., Choi, S.H., Hwang, D., and Nam, H.G. (2009). Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in *Arabidopsis*. *Science* 323, 1053-1057.

Kim, Y.J., Zheng, B., Yu, Y., Won, S.Y., Mo, B., and Chen, X. (2011). The role of Mediator in small and long noncoding RNA production in *Arabidopsis thaliana*. *EMBO J* 30, 814-822.

Kurihara, Y., Takashi, Y., and Watanabe, Y. (2006). The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. *RNA* 12, 206-212.

Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M., and Shapiro, L. (2000). Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 290, 2144-2148.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. *Proc Natl Acad Sci U S A* 102, 4936-4942.

Li, J., Liu, Z.J.J., Pan, Y.C.C., Liu, Q., Fu, X., Cooper, N.G.F., Li, Y.X., Qiu, M.S., and Shi, T.L. (2007). Regulatory module network of basic/helix-loop-helix transcription factors in mouse brain. *Genome Biol* 8, R244.

Liu, C., Axtell, M.J., and Fedoroff, N.V. (2012). The helicase and RNaseIIIa domains of *Arabidopsis* Dicer-Like1 modulate catalytic parameters during microRNA biogenesis. *Plant Physiol* 159, 748-758.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* 297, 2053-2056.

Lobbes, D., Rallapalli, G., Schmidt, D.D., Martin, C., and Clarke, J. (2006). SERRATE: a new player on the plant microRNA scene. *EMBO Rep* 7, 1052-1058.

Longabaugh, W.J., Davidson, E.H., and Bolouri, H. (2005). Computational representation of developmental genetic regulatory networks. *Dev Biol* 283, 1-16.

Machida, S., Chen, H.Y., and Adam Yuan, Y. (2011). Molecular insights into miRNA processing by *Arabidopsis thaliana* SERRATE. *Nucleic Acids Res* 39, 7828-7836.

Manavella, P.A., Hagmann, J., Ott, F., Laubinger, S., Franz, M., Macek, B., and Weigel, D. (2012). Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. *Cell* *151*, 859-870.

Mangan, S., and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci USA* *100*, 11980-11985.

Margis, R., Fusaro, A.F., Smith, N.A., Curtin, S.J., Watson, J.M., Finnegan, E.J., and Waterhouse, P.M. (2006). The evolution and diversification of Dicers in plants. *FEBS Lett* *580*, 2442-2450.

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* *296*, 910-913.

Mateos, J.L., Bologna, N.G., Chorostecki, U., and Palatnik, J.F. (2010). Identification of microRNA processing determinants by random mutagenesis of Arabidopsis MIR172a precursor. *Curr Biol* *20*, 49-54.

McAdams, H.H., and Shapiro, L. (2003). A bacterial cell-cycle regulatory network operating in time and space. *Science* *301*, 1874-1877.

Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA promoter element discovery in Arabidopsis. *RNA* *12*, 1612-1619.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* *298*, 824-827.

Nozawa, M., Miura, S., and Nei, M. (2012). Origins and evolution of microRNA genes in plant species. *Genome Biol Evol* 4, 230-239.

Park, M.Y., Wu, G., Gonzalez-Sulser, A., Vaucheret, H., and Poethig, R.S. (2005). Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A* 102, 3691-3696.

Qin, H., Chen, F., Huan, X., Machida, S., Song, J., and Yuan, Y.A. (2010). Structure of the Arabidopsis thaliana DCL4 DUF283 domain reveals a noncanonical double-stranded RNA-binding fold for protein-protein interaction. *RNA* 16, 474-481.

Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. *Science* 321, 1490-1492.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. *Genes Dev* 16, 1616-1626.

Ren, G., Chen, X., and Yu, B. (2012a). Uridylation of miRNAs by hen1 suppressor1 in Arabidopsis. *Curr Biol* 22, 695-700.

Ren, G., Xie, M., Dou, Y., Zhang, S., Zhang, C., and Yu, B. (2012b). Regulation of miRNA abundance by RNA binding protein TOUGH in Arabidopsis. *Proc Natl Acad Sci U S A* 109, 12817-12821.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513-520.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C.Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., *et al.* (2000). Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* 290, 2105-2110.

Rogers, K., and Chen, X.M. (2013). Biogenesis, Turnover, and Mode of Action of Plant MicroRNAs. *Plant Cell* 25, 2383-2399.

Ronen, M., Rosenberg, R., Shraiman, B.I., and Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A* 99, 10555-10560.

Rubio-Somoza, I., and Weigel, D. (2011). MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci* 16, 258-264.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E.M. (2007). A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Gene Dev* 21, 436-449.

Schommer, C., Palatnik, J.F., Aggarwal, P., Chetelat, A., Cubas, P., Farmer, E.E., Nath, U., and Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol* 6, e230.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31, 64-68.

Song, L., Axtell, M.J., and Fedoroff, N.V. (2010). RNA secondary structural determinants of miRNA precursor processing in Arabidopsis. *Curr Biol* 20, 37-41.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9, 3273-3297.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16, 16-23.

Thieffry, D., Huerta, A.M., Perez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20, 433-440.

Umetani, M., Mataka, C., Minegishi, N., Yamamoto, M., Hamakubo, T., and Kodama, T. (2001). Function of GATA transcription factors in induction of endothelial vascular cell adhesion molecule-1 by tumor necrosis factor-alpha. *Arterioscler Thromb Vasc Biol* 21, 917-922.

Vermeirssen, V., Barrasa, M.I., Hidalgo, C.A., Babon, J.A.B., Sequerra, R., Doucette-Stamm, L., Barabasi, A.L., and Walhout, A.J.M. (2007). Transcription factor modularity in a gene-centered *C. elegans* core neuronal protein-DNA interaction network. *Genome Research* 17, 1061-1071.

Wang, F., and Perry, S.E. (2013). Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. *Plant Physiol* *161*, 1251-1264.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana. *Cell* *138*, 738-749.

Werner, S., Wollmann, H., Schneeberger, K., and Weigel, D. (2010). Structure determinants for accurate processing of miR172a in Arabidopsis thaliana. *Curr Biol* *20*, 42-48.

Wu, G., Park, M.Y., Conway, S.R., Wang, J.W., Weigel, D., and Poethig, R.S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* *138*, 750-759.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. *Mol Cell* *38*, 465-475.

Xie, Z.X., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. *Plant Physiol* *138*, 2145-2154.

Xie, Z.X., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* *2*, 642-652.

Yamasaki, H., Hayashi, M., Fukazawa, M., Kobayashi, Y., and Shikanai, T. (2009). SQUAMOSA Promoter Binding Protein-Like7 Is a Central Regulator for Copper Homeostasis in Arabidopsis. *Plant Cell* 21, 347-361.

Yang, L., Liu, Z., Lu, F., Dong, A., and Huang, H. (2006). SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis. *Plant J* 47, 841-850.

Yang, X., and Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 27, 2614-2615.

Yant, L., Mathieu, J., Dinh, T.T., Ott, F., Lanz, C., Wollmann, H., Chen, X., and Schmid, M. (2010). Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. *Plant Cell* 22, 2156-2170.

Yu, B., Bi, L., Zheng, B., Ji, L., Chevalier, D., Agarwal, M., Ramachandran, V., Li, W., Lagrange, T., Walker, J.C., *et al.* (2008). The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. *Proc Natl Acad Sci U S A* 105, 10073-10078.

Zaslaver, A., Mayo, A.E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M.G., and Alon, U. (2004). Just-in-time transcription program in metabolic pathways. *Nat Genet* 36, 486-491.

Zeng, Y., and Cullen, B.R. (2004). Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res* 32, 4776-4785.

Zhan, X., Wang, B., Li, H., Liu, R., Kalia, R.K., Zhu, J.K., and Chinnusamy, V. (2012). Arabidopsis proline-rich protein important for development and abiotic stress tolerance is involved in microRNA biogenesis. *Proc Natl Acad Sci U S A* *109*, 18198-18203.

Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P., and Anderson, T.A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell Res* *15*, 336-360.

Zhang, H., and Li, L. (2013). SQUAMOSA promoter binding protein-like7 regulated microRNA408 is required for vegetative development in Arabidopsis. *Plant J* *74*, 98-109.

Zhao, Y., Yu, Y., Zhai, J., Ramachandran, V., Dinh, T.T., Meyers, B.C., Mo, B., and Chen, X. (2012). The Arabidopsis nucleotidyl transferase HESO1 uridylates unmethylated small RNAs to trigger their degradation. *Curr Biol* *22*, 689-694.

Chapter 2. Identification and Analysis of the Proximal Promoters of microRNA Genes in *Arabidopsis*¹

(This section of work was done in collaboration with Dr. Huiyong Zhang)

¹Formatted as a co-author manuscript published as:

Zhao X, Zhang H, Lei Li. 2013. *Genomics* 101 (2013) 187–194

Abstract

Endogenous microRNAs (miRNAs) modulate gene expression at the post-transcription level. In plants, a vast majority of *MIR* genes are thought to be transcribed by RNA Polymerase II (Pol II). However, promoter organization is currently unknown for most plant *MIR* genes. This deficiency prevents a comprehensive understanding of miRNA-mediated gene networks. In this study, collaborated with Dr. Zhang, we performed Pol II chromatin immunoprecipitation (ChIP) analysis in *Arabidopsis* using a genome tiling microarray. Distinct Pol II binding was found at most *MIR* loci, which allowed prediction of the transcription start sites (TSSs) for 167 *MIR* genes in *Arabidopsis* that was validated by average free energy profiling. By employing 99 position weight matrices (PWM), I systematically scanned the regulatory regions upstream of the TSSs. I discovered eleven and ten *cis*-elements that are statistically over- and under-represented in *MIR* promoters, respectively. Thus, analysis of Pol II binding provides a new perspective for studying miRNAs in plants.

Introduction

Following the initial discovery in *Caenorhabditis elegans* (Lee et al., 1993) and (Wightman et al., 1993), miRNAs are recognized as a conspicuous class of regulatory small RNA molecules (Bartel, 2004; Voinnet, 2009). The 20 to 24 nucleotide-long mature miRNAs are encoded by endogenous *MIR* genes and processed from much longer primary transcripts known as pri-miRNAs via stem-loop structured intermediates called pre-miRNAs (Bartel, 2004; Yang et al., 2012). In higher plants, pri-miRNA and pre-miRNA processing is carried out in the nucleus mainly by the endonuclease DICER-LIKE1 (Papp et al., 2003). Mature miRNAs are then transported to the cytoplasm and integrated into the RNA-induced silencing complex (RISC) (Khvorova et al., 2003; Schwarz et al., 2003). After integration into RISC, miRNAs interact with their cognate target mRNA through base pairing. In plants, such interactions typically lead to repression of gene expression through cleavage of the target transcripts (Llave et al., 2002; Reinhart et al., 2002) and translational attenuation (Brodersen et al., 2008). Recently, down regulation of gene expression by miRNA-directed DNA methylation at the target loci has also been reported (Wu et al., 2010).

As trans-acting regulators, temporal and spatial control of the abundance of individual miRNAs is immediately relevant to our understanding of any biological process that involves miRNAs. In contrast to the numerous reports that attest to the importance of miRNA-mediated regulation of gene expression, our knowledge on the regulation of *MIR* genes themselves is sparse. Several studies indicate that plant *MIR* genes are transcribed by Pol II, similar to what is found in animals (Cai et al., 2004; Lee et al., 2004). By sequencing

the 5' transcript ends, Xie et al. (Xie et al., 2005) mapped the TSSs for 52 *MIR* genes. This list was expanded upon through computational prediction of the core promoters (Zhou et al., 2007). In addition to the TATA box motifs located upstream of the TSSs (Xie et al., 2005), Megraw et al. (2006) identified other transcription factor binding motifs in the promoter of *MIR* genes in *Arabidopsis*. They showed that within the 800 nucleotides region upstream of TSSs, sequences resembling the binding sites for the transcription factors AtMYC2, ARF, SORLREP3, and LFY were overrepresented relative to protein-coding gene promoters and randomly sampled genomic sequences (Megraw et al., 2006).

While the previous studies are instrumental in establishing our working understanding of *MIR* gene transcription in plants, there are several major issues that need to be addressed. Binding of Pol II to the identified *MIR* promoter regions has not been demonstrated in plants. In addition, few of the predicted *cis*-elements have been functionally tested. Further, the number of annotated miRNA genes has since increased and the newly identified genes need to be subject to examination for promoter regions and regulatory sequences. Compared to the previously known miRNAs, most of the new miRNA genes have narrower phylogenetic distribution and exhibit weaker expression level and more prominent tissue-specific expression (Fahlgren et al., 2007; Rajagopalan et al., 2006; Yang et al., 2011). Based on these observations, it has been argued that continuous gene birth and death allows beneficial miRNAs to be maintained while deleterious ones avoided (Axtell and Bowman, 2008; Chen and Rajewsky, 2007; Fahlgren et al., 2007; Rajagopalan et al., 2006). Identification of the promoter regions of these *MIR* genes and comparison to those of the conserved are thus highly desirable to fully elucidate miRNA based gene regulation.

MIR genes that encode transcripts which are processed into identical or near identical mature miRNAs are grouped in paralogous families (Meyers et al., 2006). In contrast to the small but abundant miRNA families in animals, plants have fewer but larger families. For example, in *Arabidopsis*, the miR169 family contains at least 14 members (Kozomara and Griffiths-Jones, 2011; Yang et al., 2011). *MIR* genes of the same family, although encoding identical mature miRNAs, can differ considerably in gene structure and regulatory sequences. Thus, paralogous *MIR* genes may be differentially expressed at different developmental stages or in response to various environmental stimuli. On the other hand, many families contain highly similar members, suggesting recent expansion via tandem gene duplication and segmental duplication events (Li and Mao, 2007). Therefore, the promoter regions of the paralogous members may contain shared as well as unique motifs. For example, there are six *MIR395* genes (*MIR395a-f*) in *Arabidopsis*. When the promoter of individual family member was used to drive GFP expression, it was found that some family members share the same tissue- and cell-specific patterns of GFP expression while additional GFP expression observed for individual members (Kawashima et al., 2009). Globally cataloging the DNA motifs shared within a paralogous family or unique to individual members thus will help to identify their function and trace their evolution.

The goal of the current study is to comprehensively identify and analyze the proximal promoter regions of *MIR* genes in *Arabidopsis*. Toward this goal, Dr. Zhang performed Pol II ChIP followed with a whole genome tiling microarray analysis. Based on the Pol II binding profiles, I designed a computational method to reliably predict the TSSs and hence the proximal promoter regions of 167 *MIR* genes. I show this dataset is useful in identifying

cis-regulatory elements that are necessary for fully understanding the regulation of *MIR* genes and elucidating the miRNA networks.

Materials and methods

Plant materials and growth conditions

The plant used in this work was *Arabidopsis thaliana* ecotype Col-0. The seeds were placed on Murashige and Skoog (Sigma-Aldrich) agar plates containing 1% sucrose and incubated at 4 °C for two days after which they were exposed to continuous white light (170 $\mu\text{mol sec}^{-1} \text{m}^{-2}$) at 22 °C for four days. The seedlings were then incubated either under light or in the dark and harvested eight hours thereafter.

ChIP analyses

Chromatin isolation was performed using four-day-old whole seedlings grown under continuous white light or undergone dark-transition as previously described (Bowler et al., 2004). The resuspended chromatin pellet was sonicated at 4 °C with a Diagenode Bioruptor set at high intensity for 10 min (30 sec on, 30 sec off intervals). Chromatin was immunoprecipitated with a polyclonal anti-RNA polymerase II antibody (Santa Cruz), washed, reverse cross-linked, amplified, and hybridized to the Affymetrix At35b_MR_v04 genome tiling microarray using the manufacturer supplied protocol (Affymetrix). An aliquot of untreated sonicated chromatin was reverse cross-linked and used as a total input DNA control for microarray hybridization. Four biological replicates were hybridized to the tiling microarrays.

For ChIP-qPCR analysis, an equal amount of sonicated chromatin was incubated with IgG as a control in parallel to immunoprecipitation by the Pol II antibody. Relative abundance

of regions of interest in immunoprecipitated DNA was measured by qPCR using the ABI 7500 system and the Power SYBR Green PCR master mix (Applied Biosystems). Three independent qPCR assays were performed on immunoprecipitated DNA prepared using either IgG or the Pol II specific antibody and compared with the corresponding input DNA.

Assigning transcription level to protein-coding and MIR genes

The *Arabidopsis* genome and annotation data were extracted from the TAIR 10 release of The *Arabidopsis* Information Resource (Lamesch et al., 2012). Protein-coding genes were selected if they do not overlap with any other genes in both the 2 kb upstream and 2 kb downstream regions. Of the 3953 protein-coding genes meeting this criterion, 2000 were randomly selected for further analysis. RNA-Seq data from 11-day *Arabidopsis* seedlings (GSE30814) were used and processed through the Tophat-Cufflinks pipeline to rank the transcription level for these genes. *MIR* genes (miRBase release 17) (Kozomara and Griffiths-Jones, 2011) with validated TSSs (Xie et al., 2005) and full-length cDNA support were collected. After excluding those embedded in intron of host genes, 59 *MIR* genes were eventually selected. Their transcription levels were ranked based on qRT-PCR data specifically interrogating the pri-miRNAs as previously reported (Bielewicz et al., 2012).

Microarray data analysis

Raw microarray data was processed using Cisgenome with default parameters (Ji et al., 2008; Ji and Wong, 2005). Log₂ transformation was applied during normalization and only probes perfectly matched to the genome were used for intensity computation. Signal intensity from moving average statistics (MA statistic in TileMap) was used for pattern

making. For each *MIR* gene, a Pol II binding profile in the – 1000 to 1000 bp region relative to the TSS was drawn using a sliding window approach (window size = 100 bp, step = 5 bp). Within each window, the average signal intensity of all probes was calculated and set as the signal intensity for the current position (midpoint of the window). After scanning the entire region, the resultant series of points were lined up, the signal intensity for each point set as the value for each position, and plotted against genome coordinates to make the binding profile. For profiling multiple genes, TSSs were used to align the genes and then the same procedure carried out to average the combined dataset.

TSS prediction

The graphic characters among Pol II binding profiles of the 59 *MIR* genes with known TSSs were utilized to identify similar Pol II binding pattern for other genes. For each pre-miRNA, the region with extensive declination of Pol II binding was first identified using a sliding window approach (window size = 100 bp, step = 5 bp) to calculate the average Pol II signal intensity for a region. The midpoint of the first window with an average Pol II signal intensity 0.2 lower than the two windows (200 bp) upstream and downstream was designated as the valley. Overall, valleys were observed for 167 *MIR* genes (including 51 of the 59 with known TSSs). To predict TSSs for the 167 *MIR* genes, I developed a three-step procedure. First, the position 500 bp upstream the valley was set as the start point. Second, I searched within the 300 bp flanking sequences of the start point for TATA box like motifs. To this end, Motif Matcher (<http://users.soe.ucsc.edu/~kent/improbizer/motifMatcher.html>) was used to conduct a search based on PWM for TATA box from 345 experimentally verified plant promoters, which were collected from PlantProm DB (Shahmuradov et al.,

2003) and *MIR* genes with identified TATA boxes (Xie et al., 2005). Third, the same region was scanned using a PWM based on 236 experimentally verified transcription initiation motifs for dicots collected from PlantProm DB. If a TATA box was found 25 bp upstream to an initiation motif, the 5th nucleotide in the initiation motif was set as the refined TSS. If only a TATA box motif was found within the flanking region, position 25 bp downstream of the TATA box was set as the TSS. Otherwise the original start point was used as approximation for the TSS.

Average Free Energy profiling

I used dinucleotide parameters in DNA melting based on previously proposed models (Allawi and SantaLucia, 1997; SantaLucia, 1998) to calculate the Average Free Energy (AFE). The 2000 selected protein-coding genes, 59 *MIR* genes with known TSSs, and 167 *MIR* genes with predicted TSSs were aligned within each group with the TSSs set at the + 1 position. The overall sequences in the – 1000 to 1000 regions relative to the TSSs were scanned by calculating the mean value of free energy in DNA melting at each position. A previously described method was employed to reduce noise (Morey et al., 2011). In brief, the dinucleotide parameters were averaged over a 15 bp sliding window with one nucleotide step. After that, the mean value assigned to the midpoint of each window was used to generate the AFE profile over all the sequences.

Analysis of *cis*-elements

To identify putative *cis*-regulatory elements, a previously described method was followed (Zhou et al., 2007). Briefly, PWM for 99 transcription factor binding sites were built based

on experimentally validated data derived from the *Arabidopsis thaliana* Promoter Binding Element Database (<http://exon.cshl.org/cgi-bin/atprobe/atprobe.pl>) and the *Arabidopsis* Gene Regulatory Information Server dataset (Davuluri et al., 2003). Threshold used for specific matrix was set as the lowest score from using the matrix against all validated binding site variants. For the 2000 selected protein-coding genes, 167 *MIR* genes, and 2000 random genome sequences (1 kb long each), I used the 99 PWM to scan all the $-1,000$ to $+1$ bp regions. For each *cis*-element, proportion of sequences found to contain at least one copy of the *cis*-element was calculated for all three datasets (Ppc for protein-coding genes, PmiRNA for *MIR* genes, and Prandom for random sequences). Posterior Probability for all four possibilities ($P_{miRNA} > P_{pc}$, $P_{miRNA} > P_{random}$, $P_{miRNA} < P_{pc}$, and $P_{miRNA} < P_{random}$) was calculated using 10,000 times Monte Carlo simulation in Matlab. For specific *cis*-element, if posterior probability of $(P_{miRNA} > P_{random}) > 0.85$, the *cis*-element was considered to be enriched in *MIR* promoters. If posterior probability of $(P_{miRNA} < P_{random}) > 0.85$, the *cis*-element was considered to under-represented in *MIR* promoters.

Accession number

The original CHIP-chip data have been deposited in the National Institutes of Health Gene Expression Omnibus database under the accession number GSE35608.

Results

Profiling genome-wide Pol II binding sites in *Arabidopsis*

To obtain *in vivo* Pol II binding sites at the genome scale in *Arabidopsis*, Dr. Zhang performed ChIP experiments using a commercial antibody of *Arabidopsis* origin that is specific for the N-terminus of the largest subunit of Pol II. Two independent experiments were performed in young seedlings either grown under continuous white light or undergone light-to-dark transition. Pol II-immunoprecipitated DNA was then hybridized to an Affymetrix genome tiling microarray that interrogates ~ 97% of the nuclear genome. After data processing as previously described (Zhang et al., 2011), a global profile of Pol II binding signal was generated for both biological samples. For light-grown seedlings, I identified a total of 11,689 high-confidence Pol II-bound regions with a total length of approximately 7.8 Mb, or 6.5% of the sequenced nuclear genome. For seedlings that have undergone dark-transition, a total of 8217 Pol II-bound loci were identified that cover approximately 7.0 Mb or 5.9% of the genome.

To validate the identified Pol II occupancy, I performed two sets of experiments. First, I examined the distribution of Pol II binding activity across the five chromosomes in both samples using a sliding window approach. I found that the global Pol II binding pattern correlates in general with the gene density (Figure S1), suggesting that the detected Pol II binding reflects the transcriptional activity. As an example, analysis of chromosome 1 is illustrated in Figure 1A. Second, I randomly selected 13 Pol II-occupied regions identified from either the light or dark-transition samples and performed quantitative PCR analysis

following the ChIP assay (ChIP-qPCR). As shown in Figure 1B, specific Pol II binding was confirmed for all 13 examined loci, attesting to the reliability of the microarray analysis.

Characteristic Pol II binding around the TSSs of *MIR* genes

To characterize Pol II binding at the gene level, I first selected 2000 protein-coding genes that contain reliable information on the TSSs. I grouped these genes into three equal-sized subsets based on their ranked transcription level (see Materials and Methods). I then determined Pol II binding profiles in the light sample, which is consistent with conditions under which the expression data were obtained, by plotting the probe intensity from tiling microarray against the distance from the TSSs. I found that protein-coding genes with different transcription levels possess distinct Pol II binding profiles around the TSSs (Figure 2A). Genes ranked in the top one-third in terms of transcription level collectively show a strong Pol II binding peak at the TSSs. The profile then gradually increases in the gene bodies (Figure 2A). By contrast, for genes ranked in the middle and bottom one-third, the overall Pol II binding profiles are distinct from the high expression genes. For these genes, no strong Pol II binding peak at the TSSs was observed while even weaker binding was found in the gene body (Figure 2A). Further, for all three subsets, the average levels of Pol II binding in the gene body are highly consistent with the transcription level (Figure 2A). Such pronounced patterns suggest that the detected binding activity is primarily from Pol II in the pre-initiating and the elongating states (Phatnani and Greenleaf, 2006).

Our next goal is to quantitatively examine the distribution of Pol II binding in the miRNA

loci. To this end, I compiled 59 *MIR* genes with TSSs either validated experimentally (Xie et al., 2005) or supported by full-length cDNA collections in *Arabidopsis*. These *MIR* genes were also divided into three subsets based on ranked expression level. Plotting the Pol II binding profiles revealed similar pattern for highly expressed *MIR* genes as protein-coding genes with Pol II binding peaking at the TSSs followed by a gradually increasing profile (Figure 2B). For the middle- and bottom-ranked *MIR* genes, despite that the Pol II pattern fluctuates more due to the small dataset, the basic features of Pol II profiles remain the same as the protein-coding genes. These include relatively higher Pol II signal in upstream regions, decreased Pol II binding in the gene body, and a correlation of expression level with Pol II binding in the gene body (Figure 2B). These results demonstrate that the identified Pol II binding is relevant to the expression of *MIR* genes and useful to study their transcriptional regulation.

It can be observed that a “valley” locates around the 500 bp position downstream of TSSs in the Pol II binding profile for highly expressed protein-coding genes (Figure 2A). Intriguingly, this valley is more profound for *MIR* genes regardless of the expression level (Figure 2B). Indeed, when the 59 *MIR* genes with known TSSs were individually examined in the two biological samples, I found that 51 (86%) genes possess a Pol II binding valley around 500 bp downstream the TSSs in at least one of the samples. To test whether this pattern is robust for individual genes, I compared the Pol II binding profiles under the light and dark-transition conditions. Some representative examples are illustrated in Figure 3. This analysis revealed that overall Pol II binding profiles for individual genes, even those in the same family, are different in terms of the peak position and shape (Figure 3). However,

the characters of Pol II binding near the TSSs observed under different growth conditions were in general conserved for the same *MIR* gene (Figure 3). These results indicate that extensive declination of Pol II binding downstream of TSS is characteristic for *MIR* genes with known TSSs in *Arabidopsis*.

The above observation prompted us to examine all the 232 annotated *MIR* genes in *Arabidopsis*. Of these, 21 are embedded within the intron of another gene. As they are likely co-transcribed with their host gene and controlled by the host gene promoter (Baskerville and Bartel, 2005), they were excluded from further analysis. Additionally, there are 20 *MIR* genes that are either too close to or overlap with other genes and 7 *MIR* genes that have poor Pol II signal. These genes were also excluded as their Pol II binding is indistinguishable from the background. For the remaining 191 *MIR* genes, I was able to identify a total of 167 (87%) displaying the characteristic Pol II binding pattern (Table S1), a proportion identical to the *MIR* genes with known TSSs. Together these results indicate that strong and unique Pol II binding is associated with a majority of the *MIR* genes transcribed as independent units.

Predicting TSSs for *MIR* genes based on Pol II binding pattern

To harvest further information in the Pol II binding profile, I sought to identify the promoter regions for the 167 *MIR* genes with discernable Pol II binding pattern. To this end, I developed a method to predict the TSSs for these *MIR* genes, which is motivated by the observation that 51 of the 59 *MIR* genes exhibit spatial correlation between the Pol II binding valley and the known TSSs. To utilize the Pol II binding profile for predicting TSS

of individual *MIR* genes, a three-step procedure was followed. First, I used the base of the valley as the start point and set the position 500 bp upstream of the valley as an approximation for the TSS (Figure 4A). Then, I searched within the local sequence context for TATA box like motifs based on the previous observation that TATA box is present in the core promoter of most *MIR* genes (Xie et al., 2005). Finally, I searched approximately 25 bp downstream of the identified TATA boxes for sequences similar to the weak consensus motif around known TSSs. After each step, the prediction was refined in case these motifs were found to arrive at predicted TSSs for all 167 *MIR* genes (Table S1).

I performed two sets of analyses to evaluate the predicting power of Pol II binding profile for mapping TSSs of *MIR* genes. In the first set of analyses, I utilized the 51 *MIR* genes with known TSSs as the benchmark to determine the accuracy of the predicted TSSs. For these genes, I found that the absolute distance between the predicted TSSs and the actual TSSs is 32 bp on average. Further, I applied false discovery rate (FDR) control to the null hypothesis that the predicted TSSs are more than 200 bp away from the known TSSs and found the FDR only to be 2.0%. Next, I calculated the distance from the TSSs to the first nucleotide of the pre-miRNAs for the 59 *MIR* genes with known TSSs and genes with predicted TSSs. I found that the distribution of this measurement is essentially identical between the two groups (Figure 4B). Thus, this set of experiments proved that the predicted TSSs are physically close to the true TSSs.

The second set of experiments aimed at examining whether the DNA structural features of the predicted TSSs are the same as the known TSSs. To this end, I generated AFE profiles on the basis of free energy change in DNA melting (Morey et al., 2011) in the vicinity of

TSS for protein-coding genes and *MIR* genes. Similar to the previous report (Morey et al., 2011), I found that protein-coding genes with known TSSs show an AFE profiles with a significant difference between upstream and downstream regions and a sharp spike immediately upstream of the TSSs (Figure 5A). Such an AFE profile is consistent with the general regulatory landscape in which the upstream promoter region is less stable while the downstream region relatively more stable. As previous reported (Morey et al., 2011), the spike found ubiquitously at approximately the -35 bp region was found to coincide with several AT-rich tetramers.

Similar to the protein-coding genes, I found a spike upstream of TSS in the AFE profiles for the 59 *MIR* genes with known TSSs (Figure 5B). Interestingly, the decrease in the AFE profile downstream of the TSS is less profound for *MIR* genes (Figure 5B), which is consistent with the absence of open reading frames in the *MIR* genes. For the 167 *MIR* genes with predicted TSSs, I found that all features are observed including the sharp AFE spike immediately preceding the predicted TSSs (Figure 5C). Taken together, these results attest to the effectiveness of identifying TSSs from the Pol II binding profiles and generate accurate and reliable TSSs for 167 *MIR* genes in *Arabidopsis*.

Analyzing the *cis*-regulatory motifs in the *MIR* promoters

The reliably predicted TSSs enabled us to precisely pinpoint the proximal promoter region for each of the 167 *MIR* genes. As it was shown for *MIR* genes that 90% of predicted *cis*-elements fall within 800 bp from the TSSs (Megraw et al., 2006), I used DNA fragment corresponding to the 1 kb upstream region from the TSSs as approximation for miRNA

promoters to comprehensively identify putative *cis*-regulatory elements. Previously, 99 position weight matrices (PWM) derived from known transcription factor binding sites were used to search 52 *MIR* promoters in *Arabidopsis* (Megraw et al., 2006). Based on posterior probability against random genomic sequences, it was reported that four *cis*-elements, TATA box, AtMYC2, ARF, and SORLREP3 were most enriched in *MIR* promoters (Megraw et al., 2006). Following the same PWM procedure, I analyzed all 167 promoters (Table S2). I found from the expanded dataset that three of the four motifs (except the ARF motif) were indeed over-represented in *MIR* promoters. Additionally, I found eight more *cis*-elements (G-box, SORLIP1, RY-repeat, LTRE, EveningElement, TELO-box, DRE-like, and AtMYB2) that also show significant enrichment in *MIR* promoters based on high posterior probability ($P(P_{miRNA} > P_{random}) > 0.85$; Figure 6A).

Further, I were able to identify ten under-represented *cis*-elements (GATA box, LFY motif, T-box, GCC-box, RAV1-B, Bellringer BS3, CArG, HSEs, Ibox and CCA1) in the *MIR* promoters that were not previously reported ($P(P_{miRNA} < P_{random}) > 0.85$; Figure 6B). Since the exact method was used in the current and the previous studies (Megraw et al., 2006), these results demonstrate the importance of comprehensive and accurate promoter information in interpreting the *cis*-regulatory motifs of *MIR* genes. Interestingly, compared to protein-coding genes, approximately half of the 21 motifs also show significant difference in their frequency ($P(P_{pc} > P_{random}) > 0.85$ or $P(P_{pc} < P_{random}) > 0.85$; Table S2), suggesting that *MIR* genes may preferentially use certain *cis*-elements to control their expression.

Discussion

MIR genes are mainly transcribed by RNA Pol II (Lee et al., 2004). The resulting primary transcript is capped at the 5' end and polyadenylated at the 3' end (Cai et al., 2004), similar to mRNAs. Because the abundance of pri-miRNAs ultimately determines the level of mature miRNAs present in the cell, temporal and spatial control of the transcription of individual *MIR* genes is thus critical to miRNA-based gene regulation. Mapping the genomic regions upstream of the stem – loop-structured pre-miRNAs through nucleosome positioning and Pol II ChIP analysis has been carried out for human cells (Corcoran et al., 2009; Ozsolak et al., 2008; Wang et al., 2010). These studies indicate that many characteristics of *MIR* promoters, including the relative frequencies of CpG islands, TATA box, TFIIB recognition, initiator elements and other chromatin signatures, are similar to those of protein-coding genes (Corcoran et al., 2009; Davis-Dusenbery and Hata, 2010; Ozsolak et al., 2008; Wang et al., 2010).

In our current study performed in young seedlings of *Arabidopsis*, I found that global Pol II binding pattern for *MIR* genes generally agrees with that of protein-coding genes (Figures 1, 2 and 3). However, I noticed three features in the global Pol II binding profiles that differentiate *MIR* genes from the protein-coding genes under our experimental conditions. First, though a Pol II binding peak at the TSS for highly transcribed protein-coding genes and *MIR* genes was observed, the declination of Pol II signals downstream of the TSS is more profound for *MIR* genes at all transcription levels (Figure 2). A global Pol II signal valley was found at a position approximately 500 bp downstream of TSSs. I speculate that

such a distinct pattern is generated due to the different structure of *MIR* genes compared to protein-coding genes although further experiments are required to fully explain this phenomenon. Practically speaking, the highly similar but unique Pol II binding profile for *MIR* genes allowed us to reliably predict the TSSs and hence the promoter region for 167 *MIR* genes in *Arabidopsis* (Figure 4; Table S1).

Second, I found that the structural features of DNA in the vicinity of TSS are different for protein-coding and *MIR* genes. As shown in the AFE profiles, protein-coding genes exhibit a free energy change of about 1.5 kcal/mol when the immediate upstream and downstream regions of the TSS are compared (Figure 5A). Such an AFE profile indicates that the promoter region is thermodynamically less stable than the 5' untranslated and coding regions. However, for *MIR* genes the AFE difference upstream and downstream of the TSS is much milder (Figs. 5B, C). This finding is consistent with the fact that translation is omitted and transcription thus the primary mechanism in controlling the expression of *MIR* genes. In support of this notion, I found that *MIR* genes exhibit generally higher Pol II binding than protein-coding genes (Figure 2). Recent studies in plants revealed that new *MIR* genes are continuously appearing in evolution (Fahlgren et al., 2007; Rajagopalan et al., 2006; Yang et al., 2011). It was argued that genetic changes resulting in beneficial miRNAs are maintained while deleterious or nonproductive changes are purged or allowed to drift (Axtell and Bowman, 2008; Chen and Rajewsky, 2007). Therefore, DNA structural features could be an important determinant in the evolution of young *MIR* genes in plants that have not yet been adequately investigated.

Third, I found that *MIR* promoters have distinctive *cis*-element composition. Employing 99 PWM derived from known *cis*-regulatory elements (Megraw et al., 2006), I systematically scanned the 167 putative *MIR* promoters. I found eleven and ten *cis*-elements are over- and under-represented, tested against randomly sampled genomic sequences (Figure 6). Compared to protein-coding genes, about half of the 21 motifs also show significant difference in their frequency ($P(P_{miRNA} > P_{pc}) > 0.85$ or $P(P_{miRNA} < P_{pc}) > 0.85$; Table S2). For example, TATA box is the most abundant motifs in *MIR* promoters and shows a high posterior probability of enrichment relative to both protein-coding and random sequences (Figure 6A; Table S2). This is probably one of the reasons that lead to our accurate prediction of TSSs. Another conspicuous motif in *MIR* promoters is the G-box, which is implicated in environment sensing and responses, although it has a somewhat lower posterior probability relative to protein-coding sequences (Table S2).

In addition to providing testable candidates for functional studies, our analysis of the *MIR* promoters represents a new step toward reconstituting the miRNA networks. Our results demonstrate that global Pol II binding profile is a useful tool in the dissection of *MIR* promoters in *Arabidopsis*. As Pol II is highly conserved, our method should be easily applicable to other plant species. Identification and analysis of *cis*-regulatory elements of *MIR* genes provides important temporal and spatial measurements regarding transcription initiation, and therefore are useful to illustrate the regulatory networks in a broad range of plant species. Given the crucial roles of miRNAs in plant development and responses to environmental challenges, a comparative approach (Warthmann et al., 2008) should prove

fruitful in identifying adaptable miRNA gene batteries and tracing their evolution to help us understand the physiological diversity and successful adaptation across plant species.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.ygeno.2012.12.004>.

References

- Allawi, H.T., and SantaLucia, J., Jr. (1997). Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* *36*, 10581-10594.
- Axtell, M.J., and Bowman, J.L. (2008). Evolution of plant microRNAs and their targets. *Trends Plant Sci* *13*, 343-349.
- Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* *116*, 281-297.
- Baskerville, S., and Bartel, D.P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* *11*, 241-247.
- Bielewicz, D., Dolata, J., Zielezinski, A., Alaba, S., Szarzynska, B., Szczesniak, M.W., Jarmolowski, A., Szweykowska-Kulinska, Z., and Karlowski, W.M. (2012). mirEX: a platform for comparative exploration of plant pri-miRNA expression data. *Nucleic Acids Res* *40*, D191-197.
- Bowler, C., Benvenuto, G., Laflamme, P., Molino, D., Probst, A.V., Tariq, M., and Paszkowski, J. (2004). Chromatin techniques for plant cells. *Plant J* *39*, 776-789.
- Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science* *320*, 1185-1190.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10, 1957-1966.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 8, 93-103.

Corcoran, D.L., Pandit, K.V., Gordon, B., Bhattacharjee, A., Kaminski, N., and Benos, P.V. (2009). Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One* 4, e5279.

Davis-Dusenbery, B.N., and Hata, A. (2010). Mechanisms of control of microRNA biogenesis. *J Biochem* 148, 381-392.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4, 25.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., *et al.* (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS One* 2, e219.

Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., and Wong, W.H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26, 1293-1300.

Ji, H., and Wong, W.H. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* 21, 3629-3636.

Kawashima, C.G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y.N., Saito, K., Takahashi, H., and Dalmay, T. (2009). Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. *Plant J* 57, 313-321.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115, 209-216.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39, D152-157.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., *et al.* (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40, D1202-1210.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843-854.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23, 4051-4060.

Li, A., and Mao, L. (2007). Evolution of plant microRNA gene families. *Cell Res* 17, 212-218.

- Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297, 2053-2056.
- Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA promoter element discovery in Arabidopsis. *RNA* 12, 1612-1619.
- Meyers, B.C., Souret, F.F., Lu, C., and Green, P.J. (2006). Sweating the small stuff: microRNA discovery in plants. *Curr Opin Biotechnol* 17, 139-146.
- Morey, C., Mookherjee, S., Rajasekaran, G., and Bansal, M. (2011). DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes. *Plant Physiol* 156, 1300-1315.
- Ozsolak, F., Poling, L.L., Wang, Z., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X., Song, J.S., and Fisher, D.E. (2008). Chromatin structure analyses identify miRNA promoters. *Genes Dev* 22, 3172-3183.
- Papp, I., Mette, M.F., Aufsatz, W., Daxinger, L., Schauer, S.E., Ray, A., van der Winden, J., Matzke, M., and Matzke, A.J. (2003). Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol* 132, 1382-1390.
- Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 20, 2922-2936.
- Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* 20, 3407-3425.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. *Genes Dev* *16*, 1616-1626.

SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* *95*, 1460-1465.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z.S., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* *115*, 199-208.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M., and Solovyev, V.V. (2003). PlantProm: a database of plant promoter sequences. *Nucleic Acids Res* *31*, 114-117.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* *136*, 669-687.

Wang, G., Wang, Y., Shen, C., Huang, Y.W., Huang, K., Huang, T.H., Nephew, K.P., Li, L., and Liu, Y. (2010). RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation. *PLoS One* *5*, e13798.

Warthmann, N., Chen, H., Ossowski, S., Weigel, D., and Herve, P. (2008). Highly specific gene silencing by artificial miRNAs in rice. *PLoS One* *3*, e1829.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* *75*, 855-862.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. *Mol Cell* *38*, 465-475.

- Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. *Plant Physiol* 138, 2145-2154.
- Yang, X., Zhang, H., and Li, L. (2011). Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. *Genomics* 98, 40-46.
- Yang, X., Zhang, H., and Li, L. (2012). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. *Plant J* 70, 421-431.
- Zhang, H., He, H., Wang, X., Yang, X., Li, L., and Deng, X.W. (2011). Genome-wide mapping of the HY5-mediated gene networks in Arabidopsis that involve both transcriptional and post-transcriptional regulation. *Plant J* 65, 346-358.
- Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol* 3, e37.

Figures

Figure 1. Identification and confirmation of Pol II binding along the *Arabidopsis* genome.

- (A) Distribution of gene density (top track) and Pol II signal (bottom track) from the light sample along chromosome 1. The tracks were generated using 50 Kb sliding windows with 1 Kb step. Within each window, number of annotated genes and average Pol II binding signal were calculated and aligned to the chromosomal coordinate.
- (B) ChIP-qPCR confirmation of Pol II binding on selected loci. From microarray data, 13 Pol II-occupied regions were randomly selected. ChIP-qPCR was performed using either IgG or the Pol II specific antibody and normalized against the input genomic DNA. Error bars indicate standard deviation derived from three independent qPCR experiments.

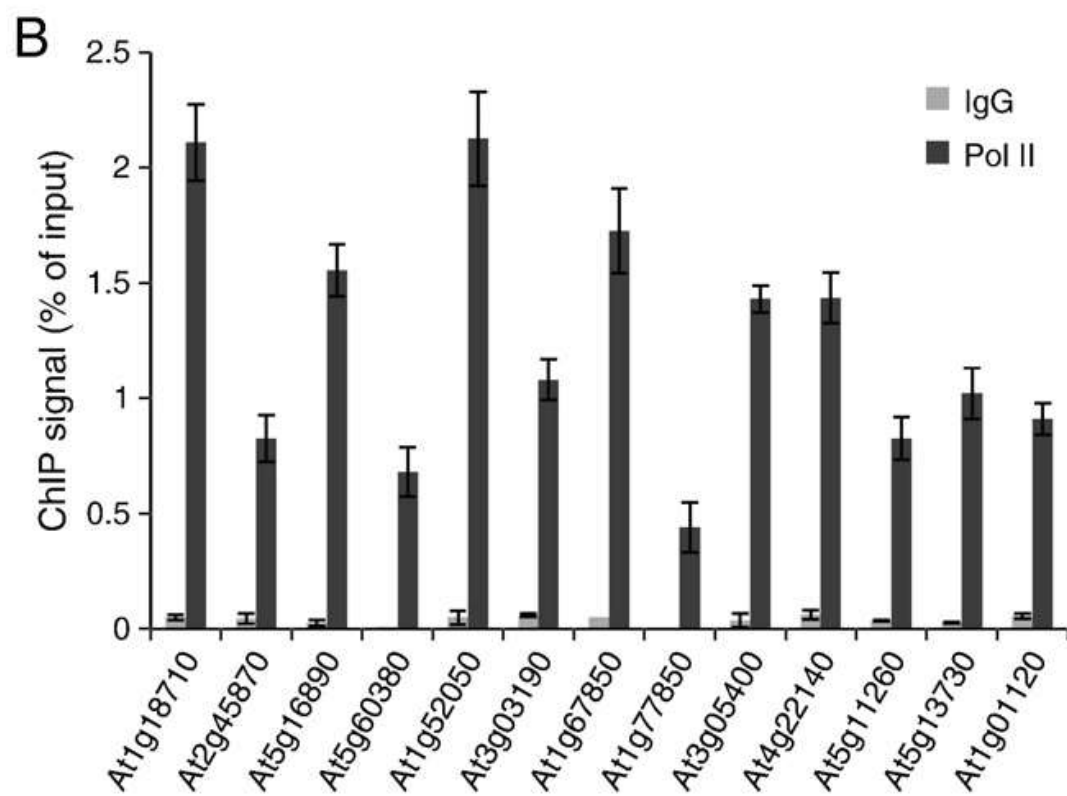
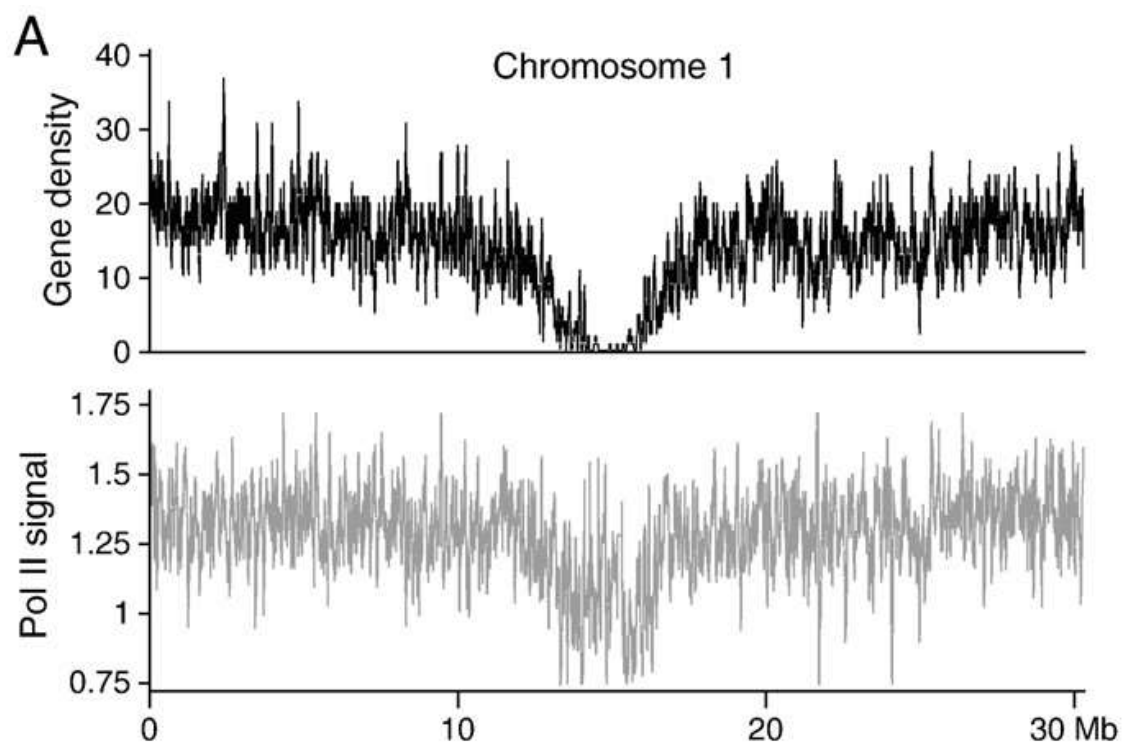


Figure 2. Pol II binding profiles in the vicinity of TSSs of protein-coding and *MIR* genes in *Arabidopsis*.

(A) A total of 2000 protein-coding genes with known TSSs were randomly selected and divided into top one-third (Top), middle one-third (Middle), and bottom one-third (Bottom) based on their ranked transcription levels. Genes in each rank were then aligned at the TSS, which is designated the + 1 position. Within each rank, the average log₂-transformed Pol II ChIP signal from the light sample was calculated for the -1000 to the 1000 regions and plotted against the relative position from the TSS.

(B) A total of 59 *MIR* genes with known TSSs were selected and similarly analyzed.

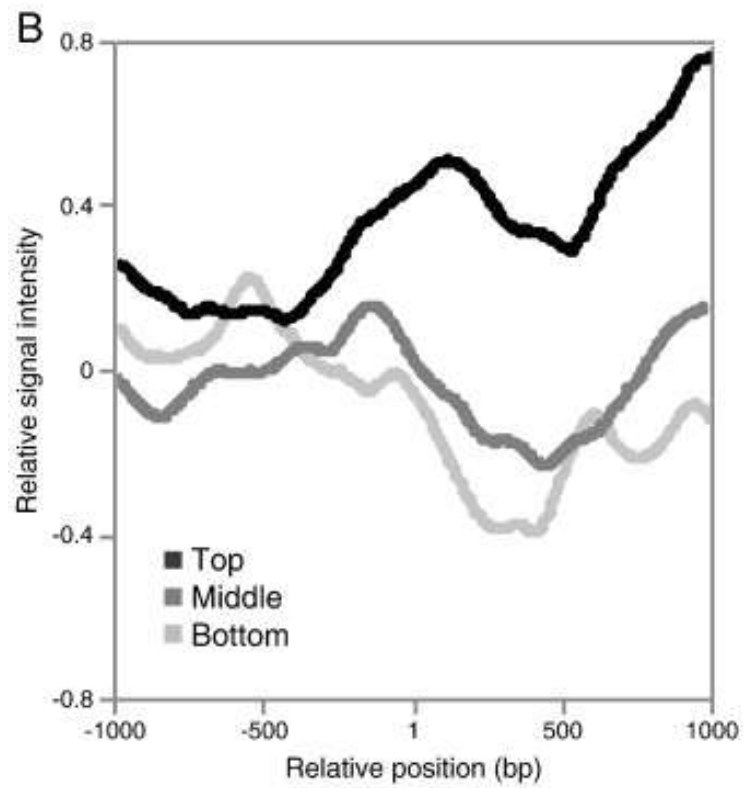
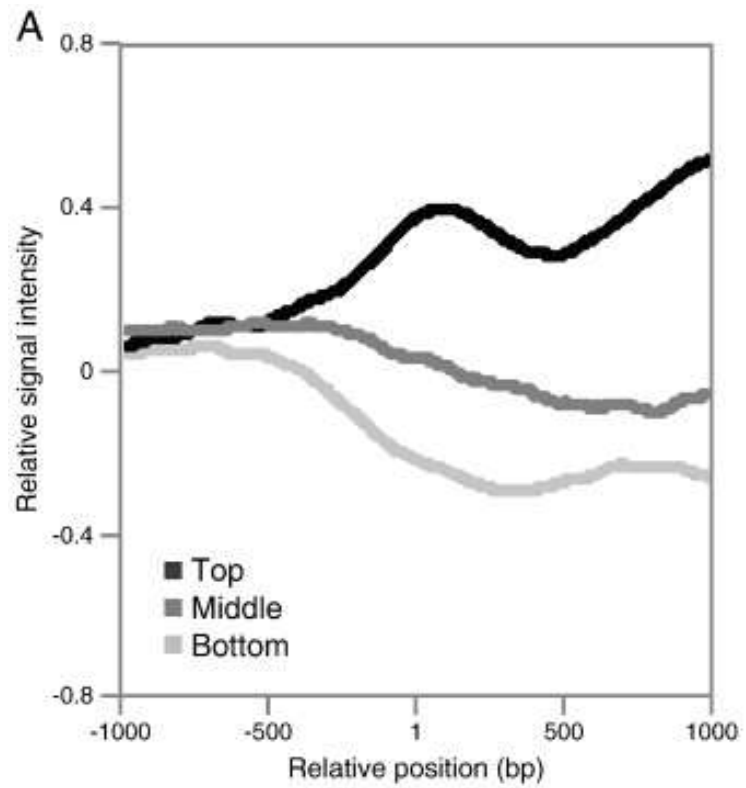
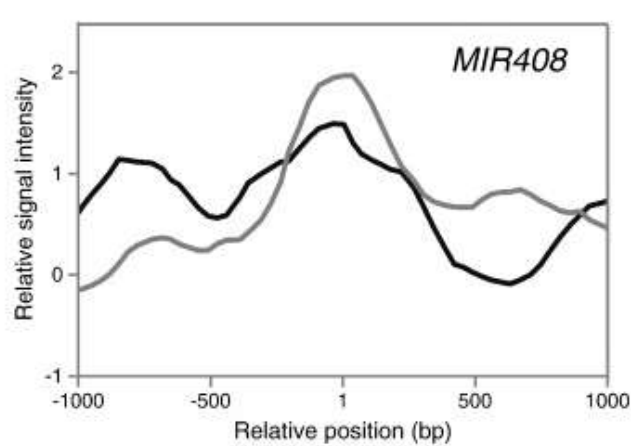
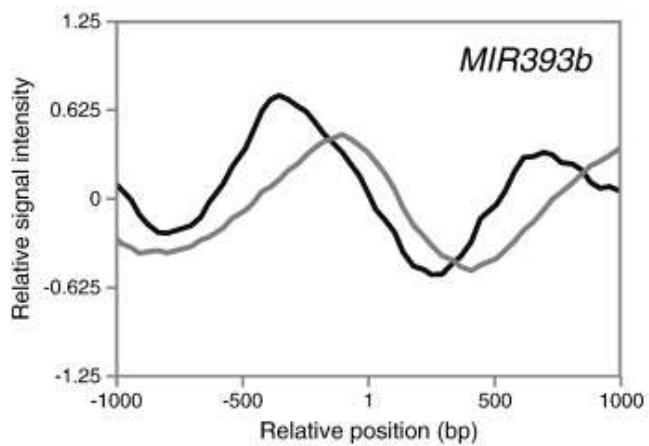
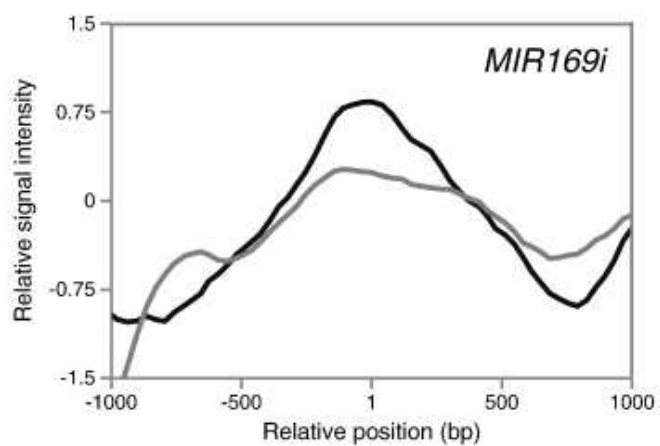
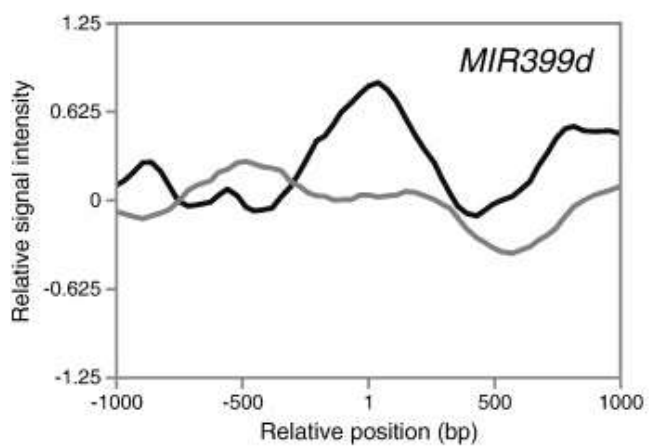
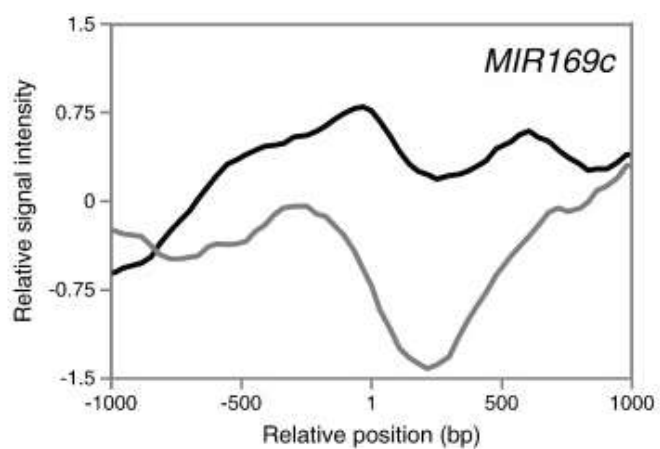
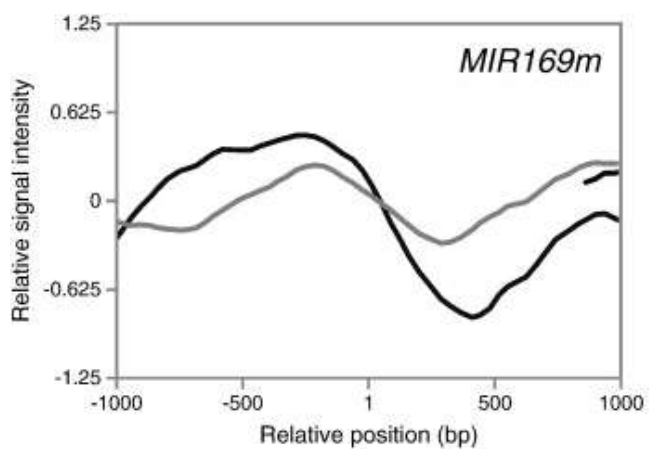


Figure 3. Comparison of the Pol II binding profile of individual *MIR* genes in different growth conditions.

Pol II ChIP signal from the light and dark-transition samples was calculated and plotted separately in the -1000 to the 1000 bp region for individual *MIR* genes with known TSSs. Shown are six representative *MIR* genes: *MIR169c*, *MIR169i*, *MIR169m*, *MIR393b*, *MIR399d*, and *MIR408*.



■ Light ■ Dark-transition

Figure 4. Prediction of TSSs for *MIR* genes based on Pol II binding pattern.

- (A) TSSs were predicted from the Pol II binding profile in a three-step-procedure. In step 1, positions 500 bp upstream of the Pol II signal valley were used as approximations for the TSSs. In step 2, local sequences flanking the putative TSSs were scanned for the TATA box motifs. In step 3, sequence approximately 25 bp downstream of the identified TATA boxes were searched for the weak consensus motif found at known TSSs. After steps 2 and 3, the predicted TSSs were refined.
- (B) Distance measured in nucleotides between the TSSs and the first nucleotide of the pre-miRNAs was calculated. For the 59 known and 167 predicted TSSs, the proportion of *MIR* genes having a given distance were respectively calculated and plotted in 200 bp intervals

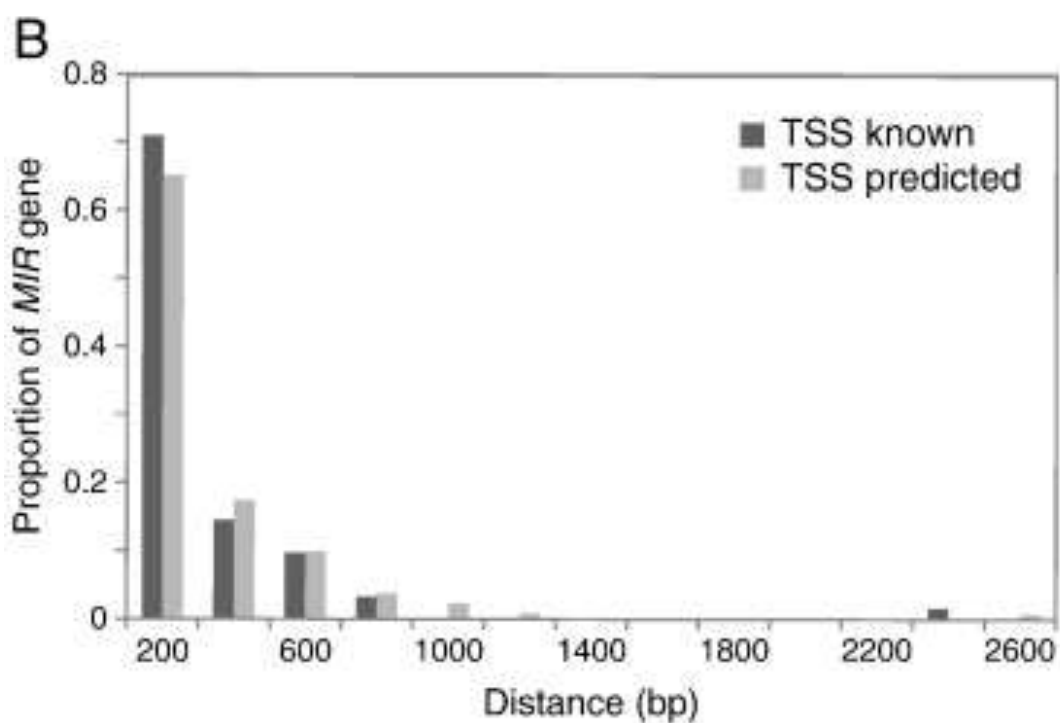
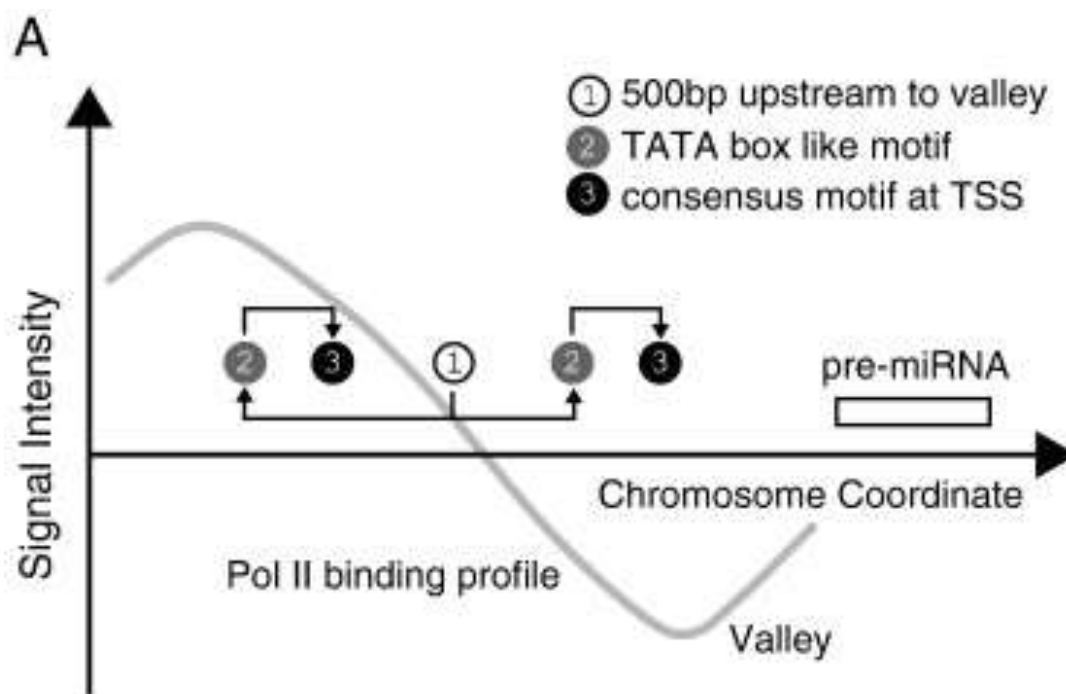


Figure 5. AFE profiles in the vicinity of known and predicted TSSs.

- (A) AFE profiles for 2000 protein-coding *Arabidopsis* genes with known TSSs. For calculating the AFE, sequences in the – 1000 to the 1000 bp region were aligned with the TSS set as the + 1 position. To obtain an average profile, the mean value of free energy change in DNA melting based on dinucleotide parameters was calculated at each position and smoothed using a previously reported sliding window approach [36].
- (B) AFE profiles for 59 *MIR* genes with experimentally validated TSSs.
- (C) AFE profiles for 167 *MIR* genes with predicted TSSs.

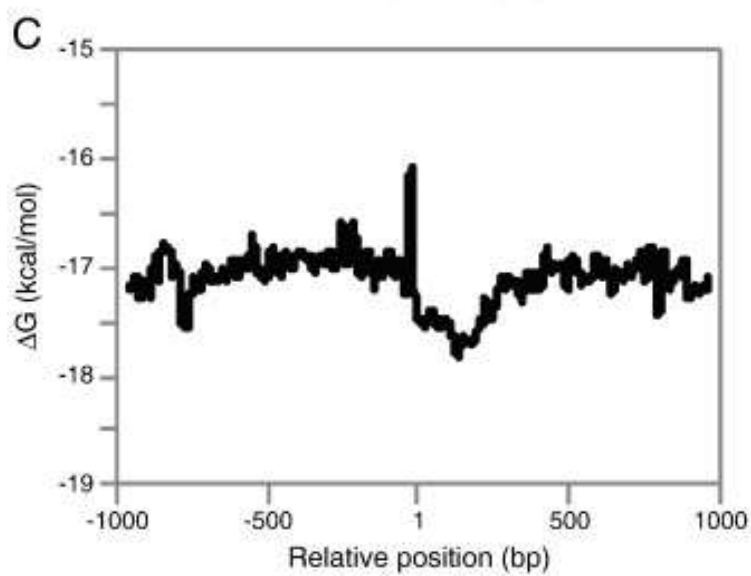
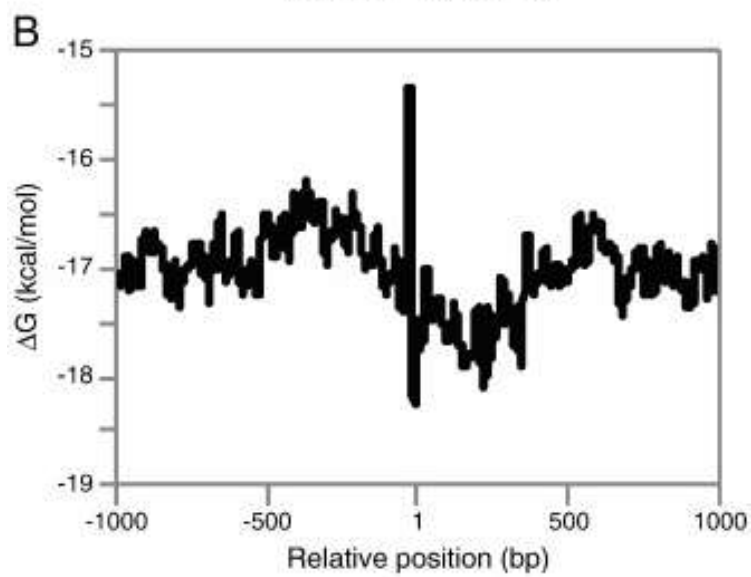
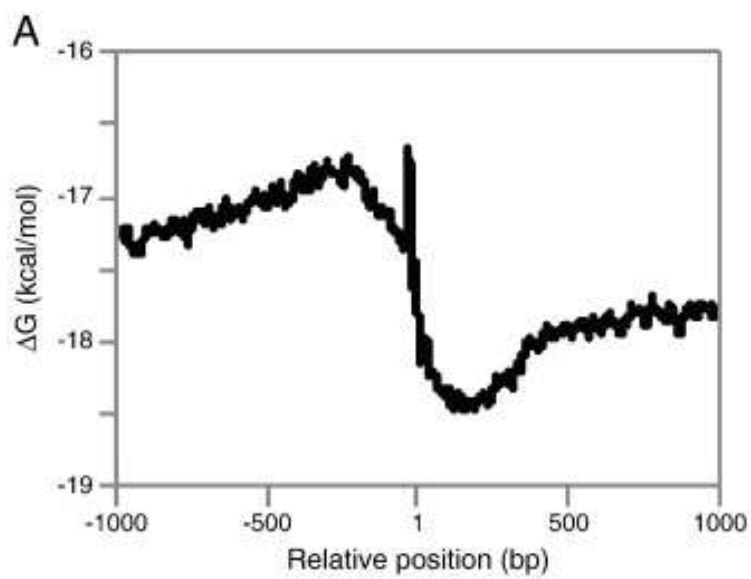
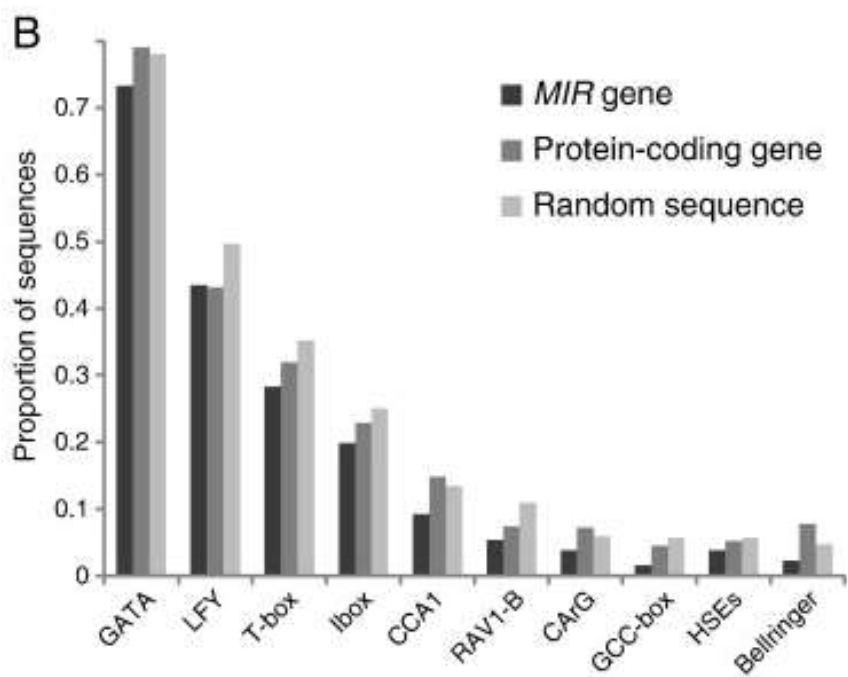
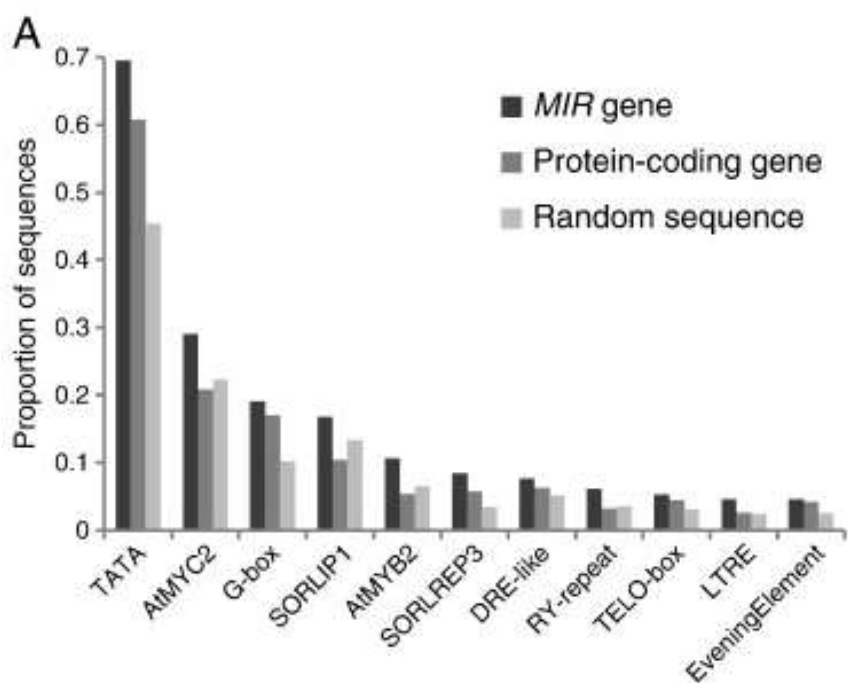


Figure 6. Over- and under-represented *cis*-elements found in *MIR* promoters.

- (A) Upstream 1 Kb regions from the predicted TSSs were scanned for *cis*-elements using 99 PWM. Compared to random *Arabidopsis* genome sequences, eleven motifs with posterior probability ($P_{miRNA} > P_{random}$) greater than 0.85 were considered to be over-represented in *MIR* promoters. Proportion of *MIR* promoters, protein-coding gene promoters as well as random genome loci containing these *cis*-elements is shown.
- (B) Under-represented *cis*-elements in *MIR* promoters. Ten motifs with posterior probability ($P_{miRNA} < P_{random}$) greater than 0.85 were considered to be under-represented. Proportion of *MIR* promoters, protein-coding gene promoters as well as random genome loci containing these *cis*-elements is shown. See Table S2 for details.

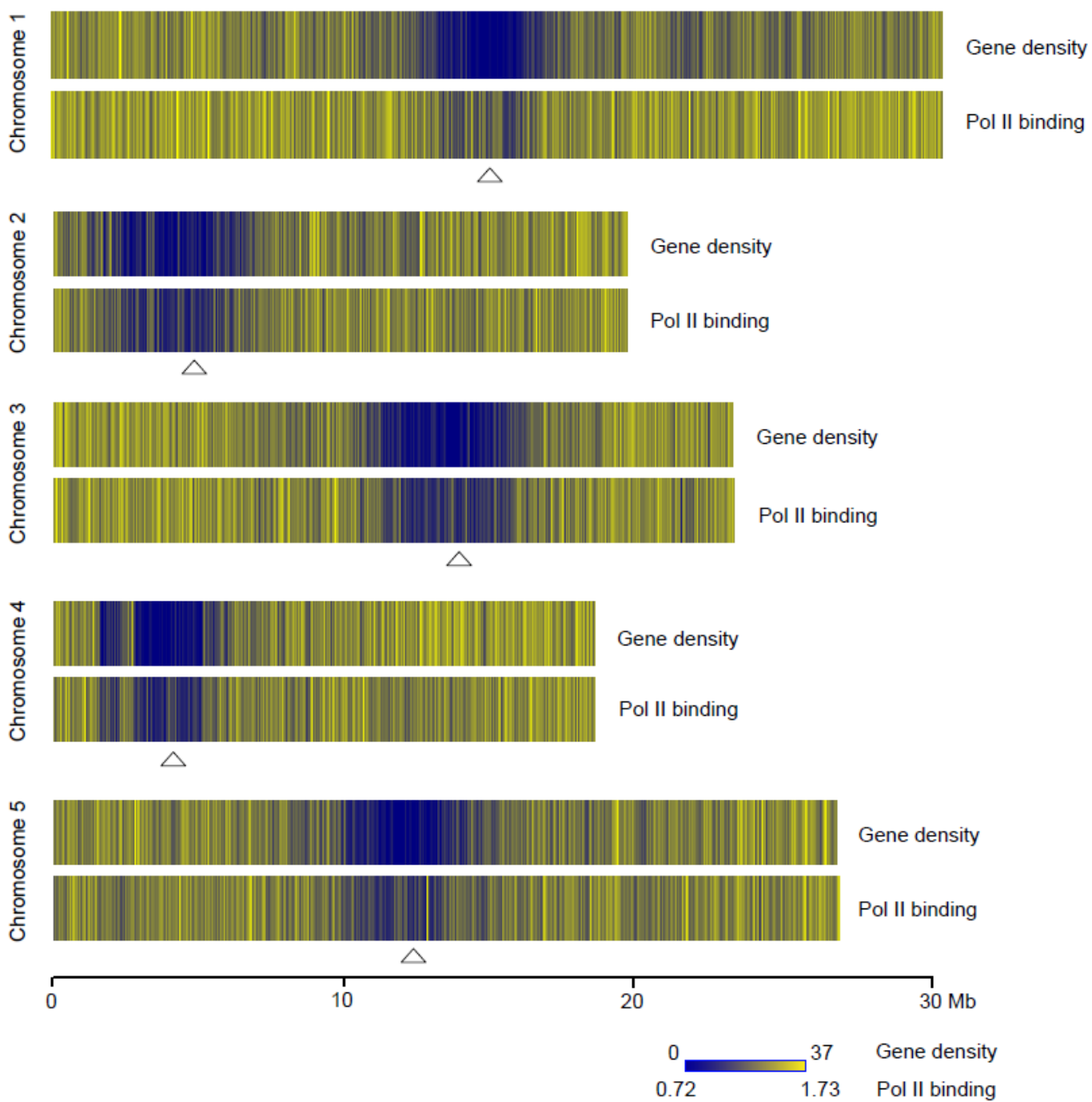


Supplementary materials

Supplementary Figure

Supplementary Figure 1. Chromosomal distribution of gene density and Pol II binding signal in *Arabidopsis*.

Gene density (top track) and Pol II signal (bottom track) for the five chromosomes were generated using a sliding window approach (window size of 50 kb, step size of 1 Kb). Within each window, number of annotated genes and average Pol II binding signal in the light sample were calculated and plotted against the chromosomal coordinates. Heatmap view was generated by Integrated Genome Browser. Position of the centromere of each chromosome is indicated with an open triangle.



Supplementary Tables

Supplementary Table 1. Information of 167 *MIR* genes with predicted TSS in *Arabidopsis*

Gene	Chr	Strand	Predicted TSS*	Gene	Chr	Strand	Predicted TSS*
<i>MIR156a</i>	2	-	10677118	<i>MIR398a</i>	2	+	1040076
<i>MIR156b</i>	4	+	15074944	<i>MIR398b</i>	5	+	4690954
<i>MIR156c</i>	4	-	1541619	<i>MIR399a</i>	1	+	10227048
<i>MIR156e</i>	5	+	3867055	<i>MIR399c</i>	5	+	24962423
<i>MIR156f</i>	5	+	9136029	<i>MIR399d</i>	2	-	14443200
<i>MIR156g</i>	2	-	8412862	<i>MIR399e</i>	2	+	14443389
<i>MIR156h</i>	5	-	22597179	<i>MIR399f</i>	2	+	14444966
<i>MIR157a</i>	1	-	24913755	<i>MIR400</i>	1	+	11785793
<i>MIR157b</i>	1	+	24920862	<i>MIR401</i>	4	-	5020495
<i>MIR157c</i>	3	-	6244825	<i>MIR403</i>	2	+	19414977
<i>MIR158a</i>	3	-	3366551	<i>MIR407</i>	2	+	13866073
<i>MIR158b</i>	1	+	20772177	<i>MIR408</i>	2	+	19319794
<i>MIR159a</i>	1	-	27713700	<i>MIR413</i>	1	-	23057582
<i>MIR159b</i>	1	+	6220360	<i>MIR414</i>	1	-	11231316
<i>MIR159c</i>	2	+	18994522	<i>MIR416</i>	2	+	7008303
<i>MIR160a</i>	2	+	16339853	<i>MIR418</i>	3	+	6516173
<i>MIR160c</i>	5	-	19009333	<i>MIR4221</i>	4	+	8460447
<i>MIR161</i>	1	+	17825624	<i>MIR4227</i>	1	+	28889054
<i>MIR162a</i>	5	-	2635437	<i>MIR4228</i>	1	+	28889117
<i>MIR162b</i>	5	-	7740857	<i>MIR4239</i>	4	+	11368616
<i>MIR163</i>	1	+	24883935	<i>MIR4240</i>	4	-	6547325
<i>MIR164a</i>	2	+	19520403	<i>MIR4243</i>	5	+	19567155
<i>MIR164c</i>	5	+	9852497	<i>MIR426</i>	1	+	22106937
<i>MIR165a</i>	1	-	79164	<i>MIR447a</i>	4	-	1529216
<i>MIR165b</i>	4	-	370020	<i>MIR447b</i>	4	-	1535811
<i>MIR166b</i>	3	+	22922002	<i>MIR447c</i>	4	-	1523686
<i>MIR166c</i>	5	+	2838502	<i>MIR472a</i>	1	-	4182941
<i>MIR166d</i>	5	+	2840552	<i>MIR5012</i>	2	+	8432658
<i>MIR166e</i>	5	-	16775687	<i>MIR5015a</i>	5	-	20555009
<i>MIR166f</i>	5	+	17516282	<i>MIR5016</i>	2	+	9919905
<i>MIR166g</i>	5	+	25504707	<i>MIR5017</i>	4	-	11963478
<i>MIR167a</i>	3	+	8108023	<i>MIR5018</i>	1	-	24544382
<i>MIR167b</i>	3	+	23406006	<i>MIR5019</i>	2	+	10403230

<i>MIR167c</i>	3	-	1306890	<i>MIR5020a</i>	4	-	6792158
<i>MIR167d</i>	1	+	11137352	<i>MIR5020b</i>	4	-	6787183
<i>MIR168a</i>	4	+	10578516	<i>MIR5021</i>	2	-	11975036
<i>MIR168b</i>	5	-	18358907	<i>MIR5023</i>	3	-	5497190
<i>MIR169a</i>	3	-	4359361	<i>MIR5026</i>	4	+	7844314
<i>MIR169b</i>	5	+	8527422	<i>MIR5027</i>	5	+	10943537
<i>MIR169c</i>	5	-	15871252	<i>MIR5028</i>	4	-	6547325
<i>MIR169d</i>	1	-	20039704	<i>MIR5029</i>	5	+	12267512
<i>MIR169e</i>	1	+	20041109	<i>MIR771a</i>	3	-	19659681
<i>MIR169g</i>	4	-	11483335	<i>MIR773b</i>	1	-	13050883
<i>MIR169h</i>	1	-	6695637	<i>MIR775</i>	1	+	29422095
<i>MIR169i</i>	3	-	9872649	<i>MIR776a</i>	1	+	22795617
<i>MIR169j</i>	3	-	9872649	<i>MIR777</i>	1	+	26637974
<i>MIR169k</i>	3	-	9876129	<i>MIR779</i>	2	+	9560430
<i>MIR169l</i>	3	-	9876129	<i>MIR781a</i>	1	+	7423517
<i>MIR169m</i>	3	-	9878786	<i>MIR783</i>	1	+	24720612
<i>MIR169n</i>	3	-	9878786	<i>MIR823</i>	3	-	4497179
<i>MIR170</i>	5	-	26411785	<i>MIR824</i>	4	+	12623528
<i>MIR171a</i>	3	+	19073097	<i>MIR825</i>	2	+	11159698
<i>MIR171b</i>	1	-	3961705	<i>MIR827</i>	3	-	22123160
<i>MIR171c</i>	1	-	22930426	<i>MIR828a</i>	4	+	13846836
<i>MIR172a</i>	2	-	11943613	<i>MIR829a</i>	1	-	11834180
<i>MIR172b</i>	5	-	1188916	<i>MIR831a</i>	2	+	10246788
<i>MIR172c</i>	3	-	3600094	<i>MIR832a</i>	4	+	6412982
<i>MIR172e</i>	5	+	23987955	<i>MIR833a</i>	1	+	29524798
<i>MIR2111a</i>	3	+	2854207	<i>MIR834a</i>	5	-	2641674
<i>MIR2111b</i>	5	+	400341	<i>MIR836</i>	2	-	10635509
<i>MIR2934</i>	3	-	5500236	<i>MIR839a</i>	1	-	25279199
<i>MIR2937</i>	5	+	13622234	<i>MIR841a</i>	4	-	7885479
<i>MIR2938</i>	5	-	22251921	<i>MIR841b</i>	4	+	2183727
<i>MIR319a</i>	4	+	12352486	<i>MIR842</i>	1	+	22577112
<i>MIR319b</i>	5	-	16660952	<i>MIR845a</i>	4	-	12217546
<i>MIR319c</i>	2	+	17029706	<i>MIR845b</i>	4	-	12214170
<i>MIR3440b</i>	3	+	6217457	<i>MIR846a</i>	1	+	22577112
<i>MIR390a</i>	2	+	16061878	<i>MIR847a</i>	1	+	2165246
<i>MIR390b</i>	5	+	23636802	<i>MIR849a</i>	3	-	16072937
<i>MIR391</i>	5	+	24293100	<i>MIR850a</i>	4	+	7845643
<i>MIR3932a</i>	4	+	2178855	<i>MIR851a</i>	3	-	19659681
<i>MIR3932b</i>	4	-	7891397	<i>MIR855</i>	2	+	4674389
<i>MIR393a</i>	2	+	16651968	<i>MIR856a</i>	1	+	11957459
<i>MIR393b</i>	3	+	20691178	<i>MIR857a</i>	4	-	7878922
<i>MIR394a</i>	1	+	7058012	<i>MIR858a</i>	1	-	26773856
<i>MIR394b</i>	1	+	28568545	<i>MIR860</i>	5	+	9098653

<i>MIR395a</i>	1	-	9363381	<i>MIR861a</i>	3	-	17838429
<i>MIR395b</i>	1	+	9364443	<i>MIR863a</i>	4	+	7845643
<i>MIR395d</i>	1	-	26270200	<i>MIR864</i>	1	+	6740499
<i>MIR395f</i>	1	+	26273857	<i>MIR865a</i>	5	+	5169809
<i>MIR396a</i>	2	-	4142564	<i>MIR867a</i>	4	+	11375026
<i>MIR396b</i>	5	+	13611460	<i>MIR868a</i>	3	+	6488010
<i>MIR397a</i>	4	+	2625915	<i>MIR869</i>	5	-	15891977

*According to TAIR10 genome coordinates.

Supplementary Table 2. Identified *cis*-regulatory elements in *Arabidopsis* *MIR* promoters

<i>Cis</i> -element	Count ¹	Proportion of sequences ² probability			Posterior	
		miRNA	PC ³	Random	PmiRNA>P	PmiRNA>Pran
TATA box ⁴	116	0.695	0.608	0.454	0.979	1.000
G-box promoter motif	32	0.191	0.170	0.102	0.715	0.998
SORLREP3	14	0.084	0.058	0.035	0.853	0.992
AtMYC2 BS in RD22	48	0.290	0.208	0.223	0.982	0.956
AtMYB2 BS in RD22	18	0.107	0.054	0.065	0.987	0.952
RY-repeat promoter motif	10	0.061	0.033	0.035	0.933	0.908
LTRE promoter motif	8	0.046	0.026	0.024	0.867	0.900
EveningElement promoter motif	8	0.046	0.043	0.025	0.531	0.887
TELO-box promoter motif	9	0.053	0.045	0.031	0.624	0.877
DRE-like promoter motif	13	0.076	0.063	0.052	0.697	0.861
SORLIP1	28	0.168	0.105	0.133	0.984	0.855
DPBF1&2 binding site motif	78	0.466	0.441	0.420	0.712	0.849
Bellringer/replumless/pennywise BS2	12	0.069	0.080	0.047	0.299	0.849
E2F binding site motif	3	0.015	0.008	0.006	0.735	0.808
ABFs binding site motif	3	0.015	0.019	0.007	0.297	0.793
ABRE-like binding site motif	18	0.107	0.134	0.085	0.160	0.785
octamer promoter motif	1	0.008	0.002	0.002	0.778	0.779
DRE promoter motif	1	0.008	0.003	0.002	0.685	0.775
CBF1 BS in cor15a	1	0.008	0.003	0.002	0.728	0.774
DREB1&2 BS in rd29a	1	0.008	0.003	0.002	0.676	0.774
MYB4 binding site motif	85	0.511	0.509	0.478	0.522	0.772
W-box promoter motif	75	0.450	0.455	0.419	0.467	0.764
Bellringer/replumless/pennywise BS1	27	0.160	0.206	0.140	0.091	0.723
MYB3 binding site motif	5	0.031	0.027	0.021	0.532	0.701
AP1 BS in AP3	1	0.008	0.002	0.003	0.832	0.690
AG BS in AP3	1	0.008	0.002	0.003	0.825	0.686
ATHB1 binding site motif	4	0.023	0.022	0.016	0.472	0.670
L1-box promoter motif	13	0.076	0.085	0.067	0.347	0.638
ABRE binding site motif	3	0.015	0.027	0.011	0.147	0.625
ATHB6 binding site motif	3	0.015	0.019	0.011	0.321	0.599
BoxII promoter motif	38	0.229	0.249	0.223	0.292	0.564
MYB1 binding site motif	8	0.046	0.036	0.042	0.684	0.537
SORLREP4	1	0.008	0.001	0.006	0.881	0.503
CBF2 binding site motif	3	0.015	0.021	0.014	0.248	0.480
GBF1/2/3 BS in ADH1	3	0.015	0.021	0.014	0.244	0.477
SORLIP5	9	0.053	0.043	0.056	0.675	0.419
ATHB5 binding site motif	4	0.023	0.033	0.024	0.212	0.404
ATB2/AtbZIP53/AtbZIP44/GBF5 BS	38	0.229	0.259	0.239	0.208	0.390

ATHB2 binding site motif	6	0.038	0.054	0.042	0.168	0.371
PI promoter motif	1	0.008	0.010	0.009	0.298	0.340
Z-box promoter motif	1	0.008	0.011	0.009	0.250	0.337
MYB binding site promoter	23	0.137	0.163	0.153	0.199	0.297
SBP-box promoter motif	1	0.008	0.013	0.012	0.207	0.231
CAAT box	105	0.626	0.673	0.658	0.141	0.230
Hexamer promoter motif	9	0.053	0.063	0.069	0.295	0.213
ARF binding site motif	36	0.214	0.229	0.244	0.347	0.213
ARF1 binding site motif	36	0.214	0.229	0.244	0.339	0.209
SORLIP2	22	0.130	0.197	0.159	0.020	0.169
RAV1-A binding site motif	121	0.725	0.767	0.766	0.149	0.162
HSEs binding site motif ⁵	6	0.038	0.052	0.056	0.199	0.149
GATA promoter motif [LRE]	122	0.733	0.790	0.780	0.068	0.115
CArG promoter motif	6	0.038	0.073	0.060	0.039	0.114
LFY consensus binding site motif	73	0.435	0.432	0.497	0.525	0.084
Ibox promoter motif	33	0.198	0.229	0.251	0.194	0.078
CCA1 binding site motif	15	0.092	0.148	0.134	0.025	0.060
Bellringer/replumless/pennywise BS3	4	0.023	0.078	0.048	0.004	0.059
T-box promoter motif	47	0.282	0.320	0.352	0.180	0.044
RAV1-B binding site motif	9	0.053	0.074	0.110	0.156	0.011
GCC-box promoter motif	3	0.015	0.045	0.057	0.019	0.005

¹Number of *MIR* promoters containing at least one copy of specific *cis*-element.

²Proportion of *MIR* promoters, promoters of protein-coding genes, and random genomic sequences that contain at least one copy of specific *cis*-element.

³Protein-coding genes.

⁴Yellow color indicates *cis*-elements specifically enriched in *MIR* promoters with posterior probability of $(P_{miRNA} > P_{random}) > 0.85$.

⁵Red color indicates *cis*-elements specifically under-represented in *MIR* promoters with posterior probability of $(P_{miRNA} > P_{random}) < 0.15$.

Chapter 3. Comparative Analysis of microRNA Promoters in *Arabidopsis* and Rice¹

¹Formatted as a first author manuscript published as:

Zhao X, Lei Li. 2013. *Genomics, Proteomics & Bioinformatics* 11 (2013) 56–60

Abstract

Endogenously-encoded microRNAs (miRNAs) are a class of small regulatory RNAs that modulate gene expression at the post-transcriptional level. In plants, miRNAs have increasingly been identified by experiments based on next-generation sequencing (NGS). However, promoter organization is currently unknown for most plant miRNAs, which are transcribed by RNA polymerase II. This deficiency prevents a comprehensive understanding of miRNA-mediated gene networks. In this study, by analyzing full-length cDNA sequences related to miRNAs, I mapped transcription start sites (TSSs) for 62 and 55 miRNAs in *Arabidopsis* and rice, respectively. The average free energy (AFE) profiles in the vicinity of TSSs were studied for both species. By employing position weight matrices (PWM) for 99 plant *cis*-elements, I discovered that three *cis*-elements were over-represented in the miRNA promoters of both species, while four and ten *cis*-elements were over-represented in *Arabidopsis* only and in rice only. Thus, comparison of miRNA promoters between *Arabidopsis* and rice provides a new perspective for studying miRNA regulation in plants.

Introduction

Following the initial discovery in the worm *Caenorhabditis elegans* (Lee et al., 1993; Wightman et al., 1993), microRNAs (miRNAs) are increasingly recognized as an important class of regulatory small RNA molecules in both animals and plants (Bartel, 2004; Voinnet, 2009). Endogenous miRNAs are encoded by *MIR* genes that are transcribed by RNA polymerase II (Bartel, 2004; Voinnet, 2009). The 20–24 nucleotide long mature miRNAs are processed from the primary transcripts called pri-miRNAs via stem-loop structured intermediates called pre-miRNAs (Bartel, 2004; Voinnet, 2009; Yang et al., 2012). In higher plants, both pri-miRNAs and pre-miRNAs are processed in the nucleus mainly by the endonuclease DICER-LIKE1 (Papp et al., 2003). Mature miRNAs are then transported to the cytoplasm and integrated into the RNA-induced silencing complex (RISC) (Khvorova et al., 2003; Schwarz et al., 2003). After integration into RISC, miRNAs interact with their cognate target transcripts through base pairing. In plants, such interactions typically lead to repression of gene expression through cleavage (Llave et al., 2002; Reinhart et al., 2002) or translational inhibition of the target mRNA (Brodersen et al., 2008). Down regulation of transcription by miRNA-directed DNA methylation at the target loci has also been reported in plants (Wu et al., 2010).

Given the critical role of miRNAs in gene regulation, temporal and spatial control of the expression of individual *MIR* genes needs to be elucidated before arriving at a complete understanding of the gene networks mediated by miRNAs. In plants, several studies have been carried out to computationally identify and analyze the miRNA promoters (Megraw et

al., 2006; Zhou et al., 2007). Results from these studies indicate that there are certain *cis*-regulatory elements enriched in the miRNA promoters (Megraw et al., 2006). In contrast to well-established programs for predicting the secondary structure of miRNA precursors or the miRNA-target interactions, computational methods to identify the promoter regions only have limited success.

Pinpointing the transcription start site (TSS) by locating the 5' end of primary transcript represents another approach to map the miRNA promoter. Using experimentally-obtained 5' transcript ends, Xie et al. successfully mapped the TSSs for 52 *MIR* genes in *Arabidopsis* (Xie et al., 2005). However, technical demands of this approach indicate that it is impractical for other plant species. On the other hand, full-length cDNA clones are regarded as critical resources for post-genomic research and have been extensively collected and sequenced in *Arabidopsis*, rice and tomato (Aoki et al., 2010; Kikuchi et al., 2003; Satoh et al., 2007; Seki et al., 2004). Utilization of these resources should generate knowledge on miRNA primary transcripts and facilitate further understanding of *cis*-regulatory elements governing miRNA transcription.

The goal of the current study is to identify and compare the promoter regions of miRNA genes between *Arabidopsis* and rice. Toward this goal, I mapped full-length cDNA sequences available to annotated miRNAs and collected TSS information for 62 and 55 miRNAs in *Arabidopsis* and rice, respectively. I then employed 99 position weight matrices (PWM) and discovered *cis*-elements that are statistically over-represented in *Arabidopsis* or rice miRNA promoters. This work thus represents a step forward in understanding regulation of miRNA genes in plants.

Results and discussion

Determination of TSSs for miRNAs by full-length cDNA mapping

The overall workflow of the current work is to employ full-length cDNA available in the model plants *Arabidopsis* and rice to precisely pinpoint the TSS for miRNA genes, and then use such information to analyze and compare the promoter features between the two species. To this end, I first mapped to the genome sequences the 299 and 591 annotated miRNA precursors in *Arabidopsis* and rice (Kozomara and Griffiths-Jones, 2011), respectively. I then mapped >155,000 full length-cDNA sequences in *Arabidopsis* (Seki et al., 2004) and >28,000 in rice (Kikuchi et al., 2003) to the corresponding genomes. Inspection of the mapping results indicates that there are 40 miRNA precursors located within full-length cDNA mapped loci in *Arabidopsis*. Further aligning the mapped regions to the annotated gene models revealed that out of the 40 miRNAs, one resides in the 5' UTR of protein-coding genes, one in exonic region, 13 in the intronic region, and four in the 3' UTR (Figure 1). Because these miRNAs are embedded within other genes and likely controlled by the host gene promoter (Baskerville and Bartel, 2005), they were excluded from further analysis. For the 21 miRNAs mapped to full-length cDNA and intergenic region, I consider the 5' end of the corresponding full length-cDNA sequence as the TSS for the miRNA. Combining these data with the previous dataset (Xie et al., 2005), I came up with a total of 62 miRNAs with experimentally-determined TSSs in *Arabidopsis*.

In rice, 157 miRNA precursors were found to be supported by full-length cDNA sequences.

Of these, 77 reside in the intron of protein-coding genes, 14 in the 5' UTR, four in the 3' UTR, and seven in the exon (Figure 1). After excluding miRNAs embedded in protein-coding genes, 55 miRNAs in rice were found in intergenic regions and their TSSs were assigned based on the full-length cDNA sequences. Even though the sample size is small for both species, it is interesting to notice that the proportions of miRNAs mapped to the introns and the intergenic regions are reversed in rice compared to those in *Arabidopsis*. It is currently unknown whether this phenomenon reflects different genome organization of miRNAs genes in the two species or is related to the quality of miRNA annotation.

DNA features at the miRNA TSSs

To systematically compare miRNA promoters between *Arabidopsis* and rice, I first calculated the distance from the TSS to the first nucleotide of the miRNA precursors. Consistent with the compact genome, more than 85% miRNAs in *Arabidopsis* have a distance between TSS and the stem-loop structured precursor of <1 kb. This distance is fewer than 200 bp for more than 70% miRNAs in *Arabidopsis* (Figure 2). By contrast, just over one third miRNAs in rice have their TSSs within 200 bp from the precursors and this distance can be as far as more than 3 kb (Figure 2). Therefore, this analysis indicates that, if our finding is applicable to all miRNAs, it is not suitable to use the first nucleotide of the miRNA precursor as the surrogate for TSS in functional studies of the promoters in rice.

I next compared the DNA structural features around miRNA TSSs between *Arabidopsis* and rice. To this end, average free energy (AFE) profiles in DNA melting were generated for genomic regions at the vicinity of TSSs in both plants. Using random sequences as a

control, I found that *Arabidopsis* miRNAs exhibit higher AFE (~ 1.5 kcal/mol) upstream of the TSS than the downstream region (Figure 3). Further, a sharp spike immediately upstream of the TSS is observed (Figure 3), similar to what was reported for protein-coding genes in *Arabidopsis* (Morey et al., 2011). As previously reported (Morey et al., 2011), DNA corresponding to the spike was found to be enriched with several AT-rich tetramers such as TATA-box. Overall, the AFE profile around the TSS of miRNAs is consistent with the regulatory landscape that the promoter region is thermodynamically less stable than the downstream transcribed region to favor transcription factor binding and transcription initiation.

In rice, AFE profiles for both the genomic control and the miRNAs have lower values than those in *Arabidopsis* (Figure 3). As GC-rich sequences tend to be more stable in DNA melting, this observation could be accounted for by the higher GC content in the rice genome. Compared to genomic control, significant AFE changes (~ 3 kcal/mol) between upstream and downstream regions of TSS were observed in rice as well as the spike immediately upstream of TSS (Figure 3). Taken together, the AFE profiles around TSS indicate high similarity of DNA structural features between miRNA genes in *Arabidopsis* and rice and between miRNAs and protein-coding genes.

Analyzing the *cis*-regulatory elements in miRNA promoters

The TSS information for miRNAs in *Arabidopsis* and rice enabled us to precisely pinpoint the promoter region and study the composition of *cis*-regulatory motifs. Previously, 99 PWM derived from known transcription factor binding sites were used to search 52 miRNA

promoters in *Arabidopsis* (Megraw et al., 2006). It was shown that 90% of predicted *cis*-elements were within the 800 bp upstream regions from TSS and that four *cis*-elements, TATA-box, AtMYC2, ARF, and SORLREP3 were most enriched in miRNA promoters based on posterior probability against random genomic sequences (Megraw et al., 2006). In this study, I used DNA fragment corresponding to the 1 kb upstream region from TSS to comprehensively identify putative *cis*-regulatory elements. Using the same PWM procedure, I analyzed miRNA promoters in both *Arabidopsis* and rice. I found from the expanded *Arabidopsis* miRNA dataset that three of the four motifs (except ARF) were indeed over-represented in miRNA promoters. Additionally, I found four more *cis*-elements (G-box, RY-repeat, LTRE and AtMYB2) that also show significant enrichment in *Arabidopsis* miRNA promoters based on high posterior probability ($P (P_{miRNA} > P_{random}) > 0.85$; Figure 4).

In rice, using the same cutoff of posterior probability ($P (P_{miRNA} > P_{random}) > 0.85$), I identified a total of 13 *cis*-elements enriched for miRNA promoters. By comparing rice and *Arabidopsis*, I found that three of these elements, TATA-box, RY-repeat and SORLREP3, are enriched for miRNA promoters in both species (Figure 4), suggesting that these *cis*-elements are fundamental to the expression of miRNAs. However, the other 10 enriched *cis*-elements (LFY, RAV1A, CAAT-box, MYB4, W-box, GCC-box, RAV1B, MYB, CCA1 and Bellringer BS2) in rice were not found to be over-represented in *Arabidopsis*. Conversely, the four *cis*-elements (AtMYC2, G-box, AtMYB2, and LTRE) enriched for *Arabidopsis* miRNAs were not over-represented in rice (Figure 4). As the PWM were

primarily derived from data in *Arabidopsis*, these findings will need to be validated with data from rice in the future.

Arabidopsis and rice are respective models for dicotyledonous and monocotyledonous plants. In the current work, genome-wide searches of full-length cDNA yielded a sizable number of TSSs for miRNAs in both species. Our analysis indicates that the general structural features of miRNA promoters are similar to those of the protein-coding genes, which is consistent with the observation that most *MIR* genes are transcribed by RNA polymerase II (Cai et al., 2004; Lee et al., 2004; Xie et al., 2005). Based on available PWM, cross-species comparison suggests that putative *cis*-regulatory elements in miRNA promoters display different degree of conservation. It is believed that new miRNAs have continuously appeared during evolution. These miRNAs, once incorporated into the gene regulatory networks, could generate genetic novelty in different plant lineages if their regulatory regions are different. Additional studies aimed at determining the precise function of the *cis*-elements should provide further insight into the complexity and evolution of miRNA-mediated gene networks in plants.

Methods

Data source

Annotated whole genome sequences, intron sequences and other gene model features of *Arabidopsis* and rice used in this study were downloaded from release 10 of The Arabidopsis Information Resource (TAIR) database (<http://www.arabidopsis.org/>) and release 6.1 of the Rice Genome Annotation Project (<http://www.rice.plantbiology.msu.edu/>). *Arabidopsis* and rice miRNA datasets were obtained from the miRBase (Kozomara and Griffiths-Jones, 2011) (release 19.0; <http://www.mirbase.org/>). Full-length cDNA sequences in *Arabidopsis* (Seki et al., 2004) and rice (Kikuchi et al., 2003) were downloaded from <http://www.rarge.psc.riken.jp/> and <http://www.cdna01.dna.affrc.go.jp/cDNA/>, respectively.

Mapping of full-length cDNA sequences

The >155,000 full-length cDNA sequences from *Arabidopsis* and >28,000 from rice were mapped to the corresponding genome assemblies using the BLAT program (Kent, 2002). Following previously-reported criteria (Satoh et al., 2007), only alignment has >90% coverage and >95% identity was accepted. Only the sequence hit with the highest score was considered as the correctly-mapped locus and the genomic position was recorded (Satoh et al., 2007). Full-length cDNA sequences that can be mapped to more than one locus with similarly high scores were excluded.

Generation of AFE profiles

The 62 and 55 miRNA loci in *Arabidopsis* and rice, respectively, were aligned within each group with the TSS set as the 0 position. The genomic sequences in the -1500 to 1500 region relative to the TSS were scanned by calculating the mean value of free energy in DNA melting at each position to generate the AFE profiles. The calculation was performed using dinucleotide parameters in DNA melting based on previous models (Allawi and SantaLucia, 1997; SantaLucia, 1998). A previously-described method was then employed to reduce noise (Morey et al., 2011), in which the dinucleotide parameters were averaged over a 15 bp sliding window with one nucleotide step. After that, the mean value assigned to the midpoint of each window was used to generate the AFE profile over the entire sequences. As a control in both *Arabidopsis* and rice, 1000 3-kb-long genomic sequences were randomly selected and subject to the same AFE profiling procedure.

Analysis of *cis*-elements

A previously-described method (Megraw et al., 2006) was followed to identify putative *cis*-regulatory elements in the miRNA promoters. In brief, PWM for 99 transcription factor binding sites (Megraw et al., 2006) were used to scan the 1 kb region upstream of the TSS for 62 and 55 miRNA genes in *Arabidopsis* and rice. Threshold employed for a specific matrix was set as the lowest score from using the matrix against all validated binding sites. As controls, 1000 genomic sequences (1 kb long each) from *Arabidopsis* and rice were randomly selected and subject to PWM analysis. For each *cis*-element identified in the miRNA promoters, proportion of sequences found to contain at least one copy of the *cis*-

element was calculated for miRNA genes (P_{miRNA}) and the random sequences (P_{random}) in both species. Posterior probability for two comparisons ($P_{miRNA} > P_{random}$ and $P_{miRNA} < P_{random}$) was calculated using 10,000 times Monte Carlo simulation in Matlab. For a given *cis*-element, if the posterior probability of ($P_{miRNA} > P_{random}$) is greater than 0.85, it was considered to be enriched in miRNA promoters.

References

- Allawi, H.T., and SantaLucia, J., Jr. (1997). Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* *36*, 10581-10594.
- Aoki, K., Yano, K., Suzuki, A., Kawamura, S., Sakurai, N., Suda, K., Kurabayashi, A., Suzuki, T., Tsugane, T., Watanabe, M., *et al.* (2010). Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics* *11*, 210.
- Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* *116*, 281-297.
- Baskerville, S., and Bartel, D.P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* *11*, 241-247.
- Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science* *320*, 1185-1190.
- Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* *10*, 1957-1966.
- Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* *12*, 656-664.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115, 209-216.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., *et al.* (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301, 376-379.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39, D152-157.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843-854.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23, 4051-4060.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297, 2053-2056.

Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA promoter element discovery in Arabidopsis. *RNA* 12, 1612-1619.

Morey, C., Mookherjee, S., Rajasekaran, G., and Bansal, M. (2011). DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes. *Plant Physiol* 156, 1300-1315.

Papp, I., Mette, M.F., Aufsatz, W., Daxinger, L., Schauer, S.E., Ray, A., van der Winden, J., Matzke, M., and Matzke, A.J. (2003). Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol* 132, 1382-1390.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. *Genes Dev* 16, 1616-1626.

SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 95, 1460-1465.

Satoh, K., Doi, K., Nagata, T., Kishimoto, N., Suzuki, K., Otomo, Y., Kawai, J., Nakamura, M., Hirozane-Kishikawa, T., Kanagawa, S., *et al.* (2007). Gene organization in rice revealed by full-length cDNA mapping and gene expression analysis through microarray. *PLoS One* 2, e1235.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z.S., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115, 199-208.

Seki, M., Satou, M., Sakurai, T., Akiyama, K., Iida, K., Ishida, J., Nakajima, M., Enju, A., Narusaka, M., Fujita, M., *et al.* (2004). RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. *J Exp Bot* 55, 213-223.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669-687.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75, 855-862.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. *Mol Cell* 38, 465-475.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. *Plant Physiol* 138, 2145-2154.

Yang, X., Zhang, H., and Li, L. (2012). Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. *Plant J* 70, 421-431.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol* 3, e37.

Figures

Figure 1. Distribution of full-length cDNA supported miRNAs in the annotated gene structures in *Arabidopsis* and rice

Pre-miRNAs and full-length cDNA sequences were aligned to the corresponding genome. Mapped pre-miRNAs in each species were grouped based on their physical relation with annotated gene models.

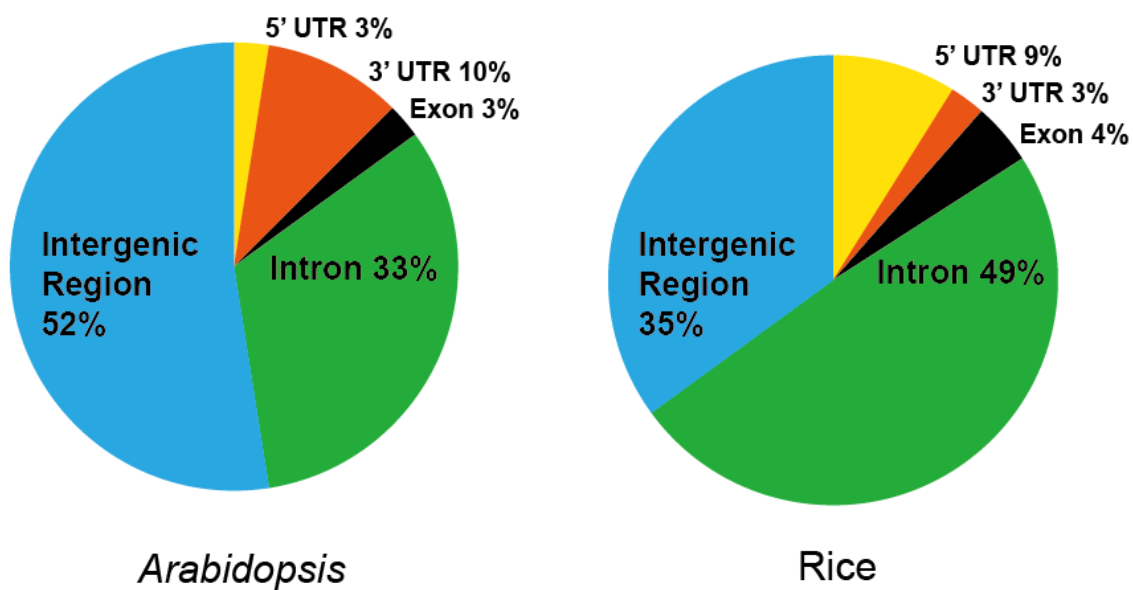


Figure 2. Distance between TSS and the start of miRNA precursor

Distance in nucleotides between the TSS and the first nucleotide of the pre-miRNA was calculated for 62 miRNAs in *Arabidopsis* and 55 in rice. The proportion of miRNAs having a given distance in 200 bp intervals in both species is respectively plotted.

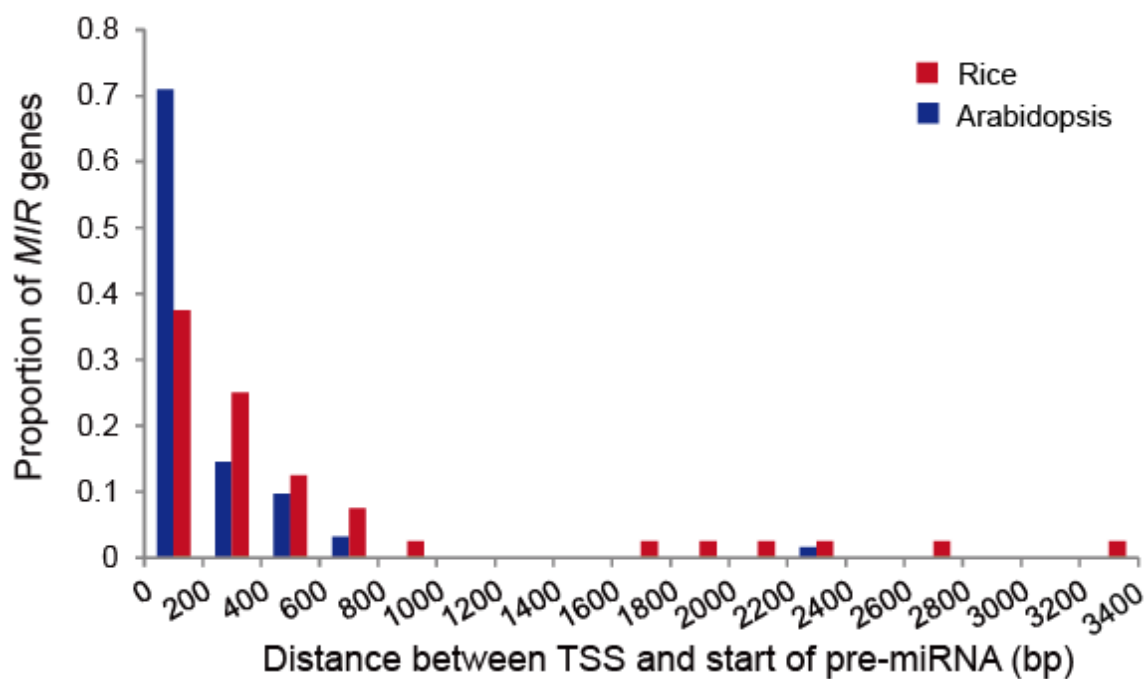


Figure 3. Comparison of AFE profiles in the vicinity of miRNA TSS in *Arabidopsis* and rice

AFE profiles were generated over the -1500 to +1500 bp region with respect to TSS of miRNAs using a sliding window approach. The AFE values for 1000 randomly-selected genomic sequences in *Arabidopsis* and rice was calculated as a control and shown as dashed lines. AFE=average free energy.

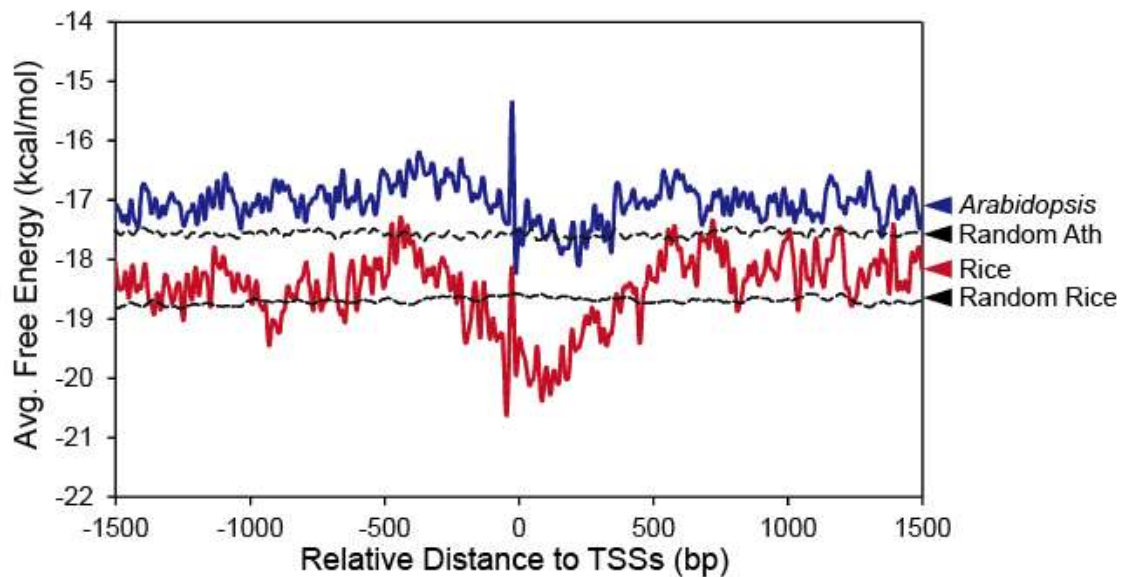
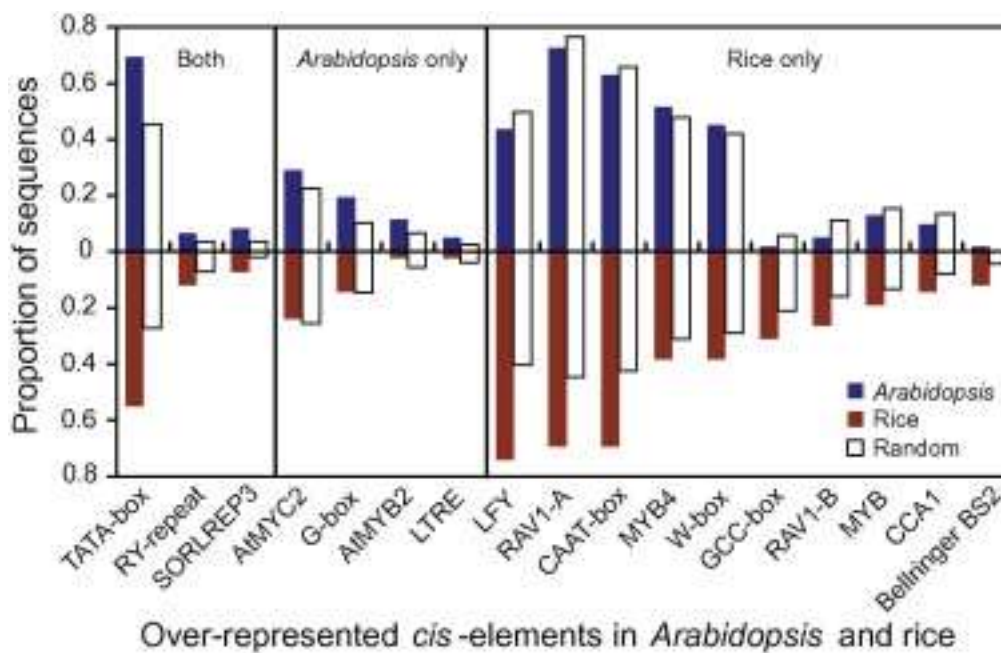


Figure 4. Over-represented *cis*-elements in miRNA promoters

Upstream 1 kb regions from the TSS of miRNAs in *Arabidopsis* and rice were scanned for *cis*-elements using 99 PWM. Compared to random genomic sequences in the two species, *cis*-elements with posterior probability ($P_{miRNA} > P_{random}$) greater than 0.85 were considered to be over-represented in miRNA promoters. Proportion of miRNA promoters containing a given *cis*-element is shown as blue (*Arabidopsis*) and red (rice) bars while proportion of the random genome loci containing these *cis*-elements is shown as blank bars. *Cis*-elements over-represented in both *Arabidopsis* and rice, in *Arabidopsis* or rice alone are shaded as indicated.



**Chapter 4. Construction and Analysis of a
microRNA-centered Regulatory Gene Network in
*Arabidopsis***

Abstract

The study of biological regulatory networks has successfully revealed the overall network topology and explained sophisticated regulatory behaviors. Despite the wealth of knowledge about transcriptional regulatory networks (TRNs), regulatory elements in addition to transcription factors (TFs), which can generate post-transcriptional level regulation, are still poorly understood. I used computational methods to incorporate microRNAs (miRNAs) into the current TRNs in *Arabidopsis* and constructed a miRNA-centered regulatory gene network. Through the integration of experimental data, computational prediction, and degradome sequencing data analysis, I collected 2189 highly reliable miRNA-target interactions (MTIs). By analyzing whole genome chromatin immunoprecipitation (ChIP) data on 35 TFs, I succeeded in building a miRNA-centered regulatory gene network of 1701 genes and 6424 interactions, including 900 TF-miRNA interactions (TMIs) and 3351 TF-miRNA target interactions (TTIs). The topological study indicated similar connectivity between miRNAs and TFs. Specifically, I found that through miRNAs, upstream regulation could reach a greater number indirect target genes, especially TFs. I showed that conserved miRNAs preferentially bound by more TFs, therefore could generate more potential crosstalk among TFs. Highly overlapped regulation by other TFs were identified on miRNAs by extracting the sub-network regulated by light-signaling TFs. The function of these TFs is related to the circadian clock, polarity identity, and flower development, providing evidence of the potential coordinated regulation on miRNAs in response to diverse input signals. Thus, the construction and analysis of the miRNA-

centered network provided insights and clues that could lead to a better understanding of miRNA-mediated gene regulation in plants.

Introduction

In the past century, biological research on individual cellular components and their functions in isolated forms, denoted as reductionism, has achieved unparalleled success (Barabasi and Oltvai, 2004). No longer satisfied by defining individual cellular constituents, researchers are now attempting to determine how they interact with each other and contribute to an organism's continuous adaptation to changing environmental conditions. The development of computational approaches and high-throughput technology has made it feasible to investigate these complex interactions by modeling them as networks (Jeong et al., 2000; Newman, 2003). Various types of networks (including signaling, metabolic, protein-protein interaction, and transcription regulatory networks) combine to form a global network (or a "network of networks") that is responsible for the overall behavior of the organism.

In the past decade, studies on transcription regulatory networks (TRNs) have elucidated the regulation of protein-coding genes in yeast (Harbison et al., 2004), *Caenorhabditis elegans* (Deplancke et al., 2006; Vermeirssen et al., 2007a), *Drosophila melanogaster* (Sandmann et al., 2007), and mammals (Boyer et al., 2005; Carro et al., 2010; Li et al., 2007). These TRNs have already revealed the overall network topology and the occurrence of particular network subgraphs (also referred to as network motifs), which appear more frequently in real networks than in randomized networks (Milo et al., 2002; Shen-Orr et al., 2002). These properties of the regulatory networks carry considerable information on the

regulation of gene expression in response to environmental change and the evolutionary origins of these networks.

miRNAs, initially discovered in *C. elegans* (Lee et al., 1993; Wightman et al., 1993), are recognized as a special class of small RNAs (sRNAs) that can regulate their target genes in a sequence-specific fashion by controlling either the availability of the mRNA for translation, mRNA stability, or its chromatin state (Chellappan et al., 2010; Voinnet, 2009; Wu et al., 2010). The function of miRNAs has been implicated in many biological processes. Unlike animal miRNAs, miRNAs in plants tend to have fewer targets (Jones-Rhoades and Bartel, 2004; Rhoades et al., 2002). However, the target genes of plant miRNAs often have regulatory function (e.g., TFs), which places miRNAs in a pivotal position in the gene regulation programs in plants. Recent research on plants revealed the involvement of miRNAs in governing plant behavior throughout the life cycle, as well as in the responses to stimulus and stress (Rubio-Somoza and Weigel, 2011). For instance, antagonistic activities of miR156 and miR172 facilitate the progression through different developmental phases (Wang et al., 2009; Wu et al., 2009); miR164 and miR319 are involved in the control of senescence (Kim et al., 2009; Schommer et al., 2008); and miR165/166 and miR390-TAS3 were found intimately associated with abaxial/adaxial polarity.

Similar to protein-coding genes, miRNAs transcription is achieved by RNA polymerase II (Lee et al., 2004; Xie et al., 2005), suggesting that miRNAs may be subject to a transcriptional regulatory mechanism similar to that of protein-coding genes. This indicates the possibility of incorporating miRNAs and TRNs to create a comprehensive regulatory

network that embodies both transcriptional and post-transcriptional regulation. For instance, by applying a method using the condition-independent yeast one-hybrid (Y1H) (Deplancke et al., 2004; Deplancke et al., 2006; Martinez et al., 2008; Vermeirssen et al., 2007a; Vermeirssen et al., 2007b), Martinez et al. mapped a genome-wide scale of TF–miRNA TRN in *C. elegans*. They discovered that feedback loops consisting of both TFs and miRNAs occur more frequently than in random networks, and they introduced the parameter of “flux capacity,” which captures the flow of information passing through the components of the feedback loops.

To study the functional importance of miRNAs and their regulatory relationship with current TRNs in *Arabidopsis*, I aim to construct a miRNA-centered regulatory gene network with increased inclusivity. This network will consist of three components—miRNAs, their upstream TFs, and downstream target genes—as well as three interactions—TF-miRNA interactions (TMIs), miRNA-target interactions (MTIs), and TF-miRNA target interactions (TTIs). To map this network, I employ both experimental data and computation predictions, which were optimized based on existing methods.

Regarding MTI prediction, several programs based on different algorithms have been developed (Allen et al., 2005; Dai and Zhao, 2011; Fahlgren et al., 2007; Stocks et al., 2012). However, a recent analysis revealed problems in these existing approaches and suggested potential improvement by combining different methods (Ding et al., 2012). In this study, I employ three computational methods to predict MTIs separately and integrate the results in conjugation with degradome sequencing data analysis and experimentally validate results to yield 2189 MTIs between 264 miRNAs and 1381 miRNA target genes.

For the systematic identification of TMIs, TSS predictions derived from distinct Pol II binding patterns (Chapter 2; Zhao et al., 2013) were collected to facilitate accurate TMI identification from whole genome ChIP data for 35 TFs. The ChIP data were also used to determine TTIs. 900 TMIs between 35 TFs and 218 miRNAs, and 3351 TTIs between 35 TFs and 951 miRNA target genes were gathered, respectively.

Based on these data, a hierarchical network centered on *Arabidopsis* miRNAs, consisting of 1701 genes and 6424 interactions, was built. The topology analysis of the network suggested that the connectivity profiles of miRNAs and TFs were similar. It also revealed the over-representation of the feed-forward loop (FFL), in which *MIR* genes may play a crucial role. In addition, combinatorial regulation at *MIR* loci was observed, indicating the central role of miRNAs in facilitating crosstalk among diverse TFs. Specifically, I discovered that the target miRNAs of light signaling pathway TFs are also heavily regulated by other pathways (e.g., polarity identity, circadian clock, and flower development). Finally, a *MIR408*-centered sub-network was extracted and used as a sample in the experimental investigation of the function of *MIR408* in Chapter 5.

Results

Construction of the miRNA-centered regulatory gene network

To construct an inclusive and accurate regulatory network centered on miRNAs in *Arabidopsis*, I first collected all available data to identify the TMIs, MTIs, and TTIs (Figure 1A). I began with a current miRNA annotation that includes 337 mature miRNAs from 298 pre-miRNAs (Kozomara and Griffiths-Jones, 2014). By combining the results of three computational prediction methods (psRNATarget, UAE_sRNA, TargetFinder) (Allen et al., 2005; Dai and Zhao, 2011; Fahlgren et al., 2007; Stocks et al., 2012) and MTIs that could be supported by degradome sequencing data (Addo-Quaye et al., 2008; German et al., 2008), I was able to identify 2153 MTIs. Experimentally validated MTIs collected from miRBase (Kozomara and Griffiths-Jones, 2014) and the literature (Addo-Quaye et al., 2008; Allen et al., 2005; Ding et al., 2012; German et al., 2008), which were not present in my prediction primarily due to low sequence complementarity, were retrieved and combined to generate 2189 high-confident MTIs between 264 miRNAs and 1381 target genes (Figure 1B).

To investigate transcriptional regulations on miRNAs and miRNA targets, I gathered and processed chromatin immunoprecipitation data of *Arabidopsis* TFs generated from ChIP-chip and ChIP-seq experiments. The resultant dataset included high quality genome-wide binding data for 35 TFs from 39 experiments obtained from the Gene Expression Omnibus and ArrayExpress (Edgar et al., 2002; Rustici et al., 2013). Based on the annotated functions, these TFs appeared to be involved in 12 different biological processes (Figure 1C;

Supplemental Table 1), particularly in flower development, light signaling pathway, and polarity specification. Utilizing the information on miRNA TSSs (Zhao et al., 2013), I obtained 900 high-confidence transcriptional regulatory interactions between 35 TFs and the promoters of 218 miRNAs, and 3351 interactions between 35 TFs and 951 miRNA target genes (Figure 1B; Supplemental Table 2). By incorporating all data, a hierarchical network centered on miRNAs was constructed for *Arabidopsis* (Figure 1D), which consisted of 1701 genes (289 miRNAs, 35 TFs, 1381 miRNA targets [including 4 of 35 TFs]), and 6424 interactions (2189 MTIs, 900 TMIs, and 3351 TTIs).

Topology analysis and motif discovery

I then studied the network topology to comprehend its overall properties and to quantify the interactions between miRNAs and the rest of the network. I first determined the distribution of the degrees of all the genes. I calculated the degree of each gene (i.e., the number of regulatory interactions that the gene is linked to) and plotted it against the quantity of genes with such degree. Although the majority of the genes had only a few links, the results showed that hubs with much higher connectivity clearly existed (Figure 2A). Further transformation revealed that this degree distribution follows a power law and therefore could be called scale-free (Barabasi and Albert, 1999). Next, I plotted the distribution of both the “in-degree” (i.e., the number of TFs that bind to a miRNA; see Figure 2B) and the “out-degree” (i.e., the number of targets a miRNA regulates; see Figure 2C) of the miRNAs. Similar to the TF target genes in *S. cerevisiae* (Shen-Orr et al., 2002), *E. coli* (Milo et al., 2002), and *C. elegans* (Deplancke et al., 2006), the “in-degree” of miRNAs exhibited an exponential distribution, which was also discovered for miRNAs in *C. elegans* (Li et al.,

2007). This finding indicates that miRNAs, which mostly are each regulated by fewer than three TFs, do not act differently from other TF target genes in terms of their transcriptional regulation.

A recent study on the miRNA-TF network in *C. elegans* found an exponential distribution of the out-degree of miRNAs (Li et al., 2007). However, by including non-TF miRNA targets in our network, the “out-degree” of miRNAs was actually best approximated by a power law distribution (Figure 2C), which has been observed on TFs in TRNs from various organisms. Given the proposed fine-tuning function of miRNAs in gene expression regulation (Bartel, 2004), this scale-free “out-degree” distribution unveiled that miRNAs could potentially globally regulate many genes rather than just a few specific targets. Interestingly, the aforementioned topological architecture has also been discovered in various regulatory networks without miRNAs in *S. cerevisiae* (Shen-Orr et al., 2002), *E. coli* (Milo et al., 2002), *C. elegans* (Deplancke et al., 2006), and mouse (Li et al., 2007). Thus, the incorporation of miRNAs into the conventional TRNs does not appear to change the basic network topology.

Next, I analyzed the network for the presence of recurring gene circuits (network motifs) by comparing their frequency to 1000-randomized networks in which the nodes’ characteristics remained the same. Several common motifs that have been previously found in the networks of other organisms (i.e., *E. coli*, *S. cerevisiae*, *C. elegans*, and mouse) also showed in the miRNA network (Supplemental Table 3). In brief, the single input and feed-forward loops (FFLs) are the most overrepresented 3-node motifs. Bi-fan and some motifs derived from FFL were overrepresented in 4-node sub-networks (Figure 2D). The repeated

appearance of FFLs and other motifs revealed that how miRNAs were connected to the rest of the network, and demonstrated their pivotal functions in fulfilling sophisticated regulatory tasks.

miRNAs play important roles in TF regulation crosstalk

Statistical analysis of the target genes of the 35 TFs reveals a linear correlation between the number of miRNAs and other genes targeted by the same TF, with a 1:60 ratio and Pearson correlation coefficient of 0.93 between the two groups (Figure 2E). Interestingly, comparison between the number of TF genes that can be regulated directly by TF and indirectly through miRNAs showed a similar correlation (correlation coefficient of 0.86) but with a much higher 1:4.7 ratio. These results suggest that TF could indirectly target more downstream genes through miRNAs. If this pathway is preferable, an implication is that the TF binding sites are expected to enrich within the regulatory region of miRNAs than in genes without regulatory function, which is supported by comparing the ratio between miRNAs and other genes that are bound by a particular TF (1:60) to the ratio between the *Arabidopsis* miRNAs and other genes count (1:100). Moreover, the proportions of miRNAs that are targeted by multiple TFs are anticipated to be higher. Hence, I calculated the frequency of *Arabidopsis* miRNAs, TFs, and other genes bound by different number of TFs and plotted this distribution (Figure 3A). I chose to plot genes with fewer than 15 TF binding sites, which included all miRNAs, up to 97.8% of TFs and 99.6% of other genes. The blue line (Figure 3A) indicates that nearly 74% of other genes were bound by only two or fewer TFs, compared to 51% in TFs and 55% in miRNAs. On the other hand, the frequency of TFs and miRNAs that possess TF binding sites from 3 to 15

exceeded other genes, respectively. This difference indicates that miRNAs and TFs on average have more TF binding sites than other genes. Together, these findings showed that TFs preferentially bind to miRNAs and other TFs perhaps as a mechanism to reach out to more downstream genes.

Previous phylogenetic studies have revealed that miRNAs include those that are deeply conserved and those that are species specific. The construction of a miRNA-centered regulatory network enabled comparisons of connectivity between conserved and species-specific miRNAs in *Arabidopsis*. Therefore, I grouped *Arabidopsis* miRNAs into four classes based on their conservation data gathered from miRBase (Kozomara and Griffiths-Jones, 2014): 1) miRNAs only discovered in *Arabidopsis thaliana* and/or *A. lyrata*; 2) miRNAs discovered in at least one other rosoid species excepting *Arabidopsis*; 3) miRNAs discovered in both rosoid and asterid species; and 4) miRNAs discovered not only in dicot but monocot or other species. Figure 3B demonstrates that most conserved miRNAs can target significantly more TFs than non-conserved miRNAs ($p < 1.6 \times 10^{-3}$), indicating its capability in bridging diverse TFs. Hence, such properties of miRNAs that could bring inter-pathway crosstalk by receiving and delivering regulatory signals to more than one pathway may be relevant to miRNAs' basic function and therefore deeply conserved.

In order to comprehend the extent to which conserved miRNAs bridge diverse TFs to fulfill sophisticated regulatory tasks, I quantified the frequency of genes that are simultaneously bound by any combination of two different TFs (two-TF module) for four groups of genes (e.g., conserved miRNAs, i.e. miRNAs that are found in both rosoids and asterids clades, non-conserved miRNAs, i.e. miRNAs that are only found in *A. thaliana* and/or *A. lyrata*,

TFs, and the other genes). The heat map of conserved miRNAs provided in Figure 3C shows fewer blanks (207 compared to 286 in non-conserved miRNAs) and deeper color (2.1% of conserved miRNAs on average were able to connect each two-TF module compared to 0.93% of non-conserved miRNAs), indicating that a higher proportion of conserved miRNAs created connections between a greater number of two-TF modules than non-conserved miRNAs did. When combined, miRNAs overall were able to link 432 (72%) of all 595 two-TF modules. A similar scenario was observed between TFs and other genes (Figure 3D): TFs connected 566 (95%) two-TF modules with on average 2.5% of genes connected each module compared to 0.82% of other genes. Both heat maps show higher connection ratios among TFs with close or correlated functions (data not shown), indicating that TFs involved in the same biological process tend to bind to similar sets of genes. Notably, heavy connections between the TFs belonging to three functional clusters (i.e., flower development, light signaling, polarity identity) were also revealed (data not shown), demonstrating the potentially massive crosstalk between these regulatory pathways.

Taken together, miRNAs in *Arabidopsis*, especially conserved ones, tend to be heavily regulated by TFs and in turn regulate many target TFs. Thus, miRNAs constitute an important layer in the regulatory network with great potential in integrating complex transcriptionally regulatory signal input.

miRNAs facilitate crosstalk between light signaling and other biological processes

To dissect the miRNA-centered regulatory gene network, I focused on the sub-network containing all miRNAs targeted by the light-signaling TFs (HY5, GBF1, FHY3, PIF3, PIF4, PIF5) and traced the entire set of TFs that regulate these miRNAs (Figure 4A). Primarily, this light-signaling TF-miRNA sub-network includes 20 miRNAs, 31 TFs, and 165 TMIs, thus possessing a higher density of interactions compared to the network average (82.6 TMIs per 20 miRNAs). Intriguingly, all 20 miRNAs belong to highly conserved miRNA families, which suggests that these miRNAs may play fundamental roles in mediating the transduction of light signals. The close examination of the TFs involved revealed three other functional clusters: circadian clock, polarity identity, and flower development. Given that light serves as the primary energy source and is one of the most dominating signal inputs for plants, the observation of the heavily overlapped regulation in miRNAs downstream of HY5 implies that miRNAs are the key nodes in facilitating crosstalk among diverse biological processes.

In the further investigation of the coordinated regulation on miRNAs, I focused on *MIR408*, which is among the most conserved miRNA families in land plants. A *MIR408*-centered regulatory gene network was generated to discover underlying crosstalk mediated by miR408 (Figure 4B). Eight TFs constituted the network, which fell into four functional clusters: light signaling (HY5, PIF3, PIF4, and PIF5); polarity identity (KAN1 and JAG); copper homeostasis (SPL7), and flower development (SEP3). Interestingly, the four light-signaling TFs potentially bind to the *MIR408* promoter at the same place (Figure 4C). Similarly, the collective placement of binding sites on the promoter region of *MIR408* by two TFs that determine polarity identity in *Arabidopsis* (KAN1 and JAG) was also

observed. Thus, binding to the cis-regulatory elements in the *MIR408* promoter by different TFs may serve as a platform for coordinating environmental input to regulate plant growth. The next chapter will describe in detail our attempt to demonstrate the role of miR408 in mediating light-copper crosstalk through the *SPL7-HY5-MIR408* circuit. I propose that this circuit represents a potentially conserved determinant of photosynthetic activity and hence plant growth and adaptation.

To sum up, I developed a computational approach to construct and visualize the miRNA-centered network in *Arabidopsis*. In addition to the structural and topological analysis of the network, I performed analysis of the light signaling-related miRNA sub-network. Although still preliminary, the findings described herein should provide an impetus and guidance for us to systematically analyze the cooperative character of miRNA regulation in mediating inter-pathway crosstalk.

Discussion

The study of the mechanisms of gene regulation is important because it leads to understanding the adaptation and behavior of all organisms. The development of methods that map, visualize, and analyze the sophisticated TRNs in plants has provided ways to inspect gene expression and extended our understanding of the complexity of real biological networks. However, the absence of miRNAs indicates that the current regulatory network models are far from holistic. In plants, miRNAs have rarely been depicted in the context of networks. Hence, the integration of miRNAs into current regulatory networks provides much-needed insight into the dynamic regulation of gene activity through post-transcriptional mechanisms.

Based on several lines of evidence, most topological properties of miRNAs in the network are clearly similar to those of TFs (Figures 2, 3). For example, the out-degree, which has been illustrated to follow an unsymmetrically distributed power law for TFs, showed similar features in miRNAs. Moreover, the exponential distribution of in-degree was also discovered on miRNAs (Figure 2B). On the other hand, the quantity of tested TFs that potentially bind to miRNA and TF loci demonstrated their similar preferentiality in recruiting more TFs than other genes do. The fact that the same set of overrepresented network motifs was discovered in networks with or without miRNAs suggests that the incorporation of miRNAs into existing regulatory networks has no effect on the overall topology. This finding implies that the role that non-coding RNAs play in the network is not to modify network structure but instead to build parallel routes to TF-mediated

transcriptional regulation for the transduction of regulatory signals from upstream regulators to downstream targets through a post-transcriptional mechanism.

The comparison of the capacity of miRNAs and TFs to interact with diverse two-TF modules revealed that both conserved miRNAs and TFs preferentially receive regulatory signals from multiple upstream TFs. Therefore, although not surprising, it is interesting to assume that miRNAs may facilitate crosstalk between different signal transduction pathways. For example, molecular genetic analysis showed that *MIR408* mediates light-copper crosstalk (Figure 4B). Thus, regulatory relationships denoted in the miRNA-centered network should provide guidelines for the functional analysis of other miRNAs that integrate the signals from multiple TFs.

Finally, the extraction of the sub-network regulated by light-signaling TFs revealed heavily overlapping transcriptional regulation of miRNAs. For example, most light-regulated miRNAs are also targeted by TFs involved in circadian clock, polarity identity, and flower development. Light is an essential abiotic stimulus of plants and serves as the primary energy source. The involvement of light in plant development varies from essential processes, such as photosynthesis, to photoperiodism and developmental stage transition. The light-signaling pathway is therefore fundamental to the fitness of plants and regulates the activities of diverse related pathways. The observation of heavily overlapping regulations of miRNA by different signaling pathways is consistent with the undeniable requirement for light signaling to coordinate with other pathways to control plant growth and development. Hence, mapping and characterization of the light-signaling sub-network uncovered potentially instrumental clues to understand how multiple biological processes

are orchestrated by miRNAs at the post-transcriptional level.

It should be noted that in this study, only limited information on TFs was used to construct the miRNA-centered gene network. In the absence of quantitative and functional data, only global TF ChIP results were considered in mapping the TF-DNA association. With mainly focused on protein-protein/protein-DNA interactions and genome sequencing, the global validation and characterization of the regulatory power of miRNA remains extremely inadequate for current high-throughput technologies. Moreover, because of the obvious incompleteness of the mapped network, a systematic validation method is also lacking. Nonetheless, the construction of the miRNA-centered regulatory network undoubtedly extended our knowledge of miRNAs by characterizing their network topology, unveiling their role as potential crosstalk mediators, and providing clues for functional validation. Finally, although much of the work described in this chapter centered on the depiction of the overall role of miRNAs in the context of a regulatory network, the sub-networks should provide sufficient information to carry out functional analysis of individual miRNAs. These efforts, combined with further improvement of the overall network, should advance our understanding of plant miRNAs.

Materials and methods

MicroRNA target genes identification

Sequences of 298 pre-miRNAs and 337 mature miRNAs of *Arabidopsis* were obtained from miRBase Release 20 (Kozomara and Griffiths-Jones, 2014). Experimentally validated MTIs were collected from miRTarBase 4.5 (101 MTIs) (Hsu et al., 2014), and several publications (Addo-Quaye et al., 2008; Allen et al., 2005; Ding et al., 2012; German et al., 2008) (102 MTIs). Computational prediction methods, including psRNATarget (Dai and Zhao, 2011); UAE_sRNA (Stocks et al., 2012); TargetFinder (Allen et al., 2005; Fahlgren et al., 2007), and online target dataset PMRD (Zhang et al., 2010) and PMTED (Plant MicroRNA Target Expression Database), (Sun et al., 2013) were combined for miRNA target prediction. As suggested in (Ding et al., 2012), multi-method with a loose cutoff increases the true positives and improves the filter power of MTI prediction. Therefore, I collected MTIs from all three computational prediction methods and employed the multi-method cutoff of 2 (MTIs predicted in at least two different methods) to collect highly confident MTI predictions. Default settings for plant miRNAs were applied when running psRNATarget, UAE_sRNA, and TargetFinder.

Additional, degradome library (Addo-Quaye et al., 2008; German et al., 2008) analyzed by CLEAVELAND pipeline (Addo-Quaye et al., 2009) were collected from (Ding et al., 2012). Low confident MTIs that were only predicted by a single computational method but were supported by degradome sequencing data (enough 5' end of uncapped, polyadenylated RNAs found at predicted miRNA cleavage site) were retained for further analysis as well.

ChIP-chip analysis

As described in (Zhang et al., 2011), raw CEL data files obtained online were normalized and analyzed by two separate software packages: TAS (Version 1.1.02, Affymetrix) and Cisgenome (Ji et al., 2008; Ji and Wong, 2005). All probes were mapped to the *Arabidopsis* genome build TAIR 10 (<ftp://ftp.arabidopsis.org/>). With the TAS software, probe intensity was computed based on both PerfectMatch and MisMatch (PM/MM) with a bandwidth 250 bp. The enriched intervals were defined by p -value <0.005 , with maximum gap set to 300 bp and minimum run 100 bp.

For Cisgenome, probe intensity was computed by perfect match-mismatch (PM-MM), log₂-transformed and quantile normalized. The Moving Average (MA) method was performed for peak detection with cutoff of 2.5 fold change and window size 250 bp. Max gap and minimum run was set to 300 bp, and 100 bp.

ChIP-seq analysis

ChIP-seq reads have no unrecognized nucleotides (Ns) and passed quality filtering were aligned to the TAIR10 genome assembly using Bowtie (Langmead et al., 2009) version 0.12.7 with no more than 2 mismatches allowed. Only reads that are uniquely mapped to the nuclear genome were retained for further analysis.

In order to generate reproducible peaks, strategy described in (Wuest et al., 2012) was used. Briefly, for each dataset, two independent peak calling methods were performed before merge together. Numerous peak detection algorithms have been tested (Laajala et al., 2009;

Wilbanks and Facciotti, 2010), and demonstrated with similar performance in terms of sensitivity and specificity. Thereby, I chose to use MACS (Model-based Analysis of ChIP-Seq) (Zhang et al., 2008), and QuEST (Quantitative Enrichment of Sequence Tags) (Valouev et al., 2008), which have relatively higher true positive rate and lower false positive rate and false discovery rate (FDR), for peak detection. With the MACS, peak detection was performed with default setting, except the effective genome size, which was set to 110 Mb, and the p -value cutoff, which was increased to 10^{-3} . For QuEST, parameters were set to (bandwidth 30bp, ChIP seedling fold enrichment 10, ChIP extension fold enrichment 3, and ChIP-to-background fold enrichment 2.5).

As suggested in (Zhang et al., 2013), for each TF, reads from technical replicates were aligned and pooled together, and processed as single library. If reads are from biological replicates, independent Bowtie, MACS/QuEST analyses was performed. Only replicate-specific peaks that were identified at the same genomic location in at least two biological replicates were defined as reproducible peaks and retained.

Peak mergence and bound gene identification

For each TF binding data (either ChIP-chip or ChIP-seq), two independent peak calling algorithms, as previously described could improve both precision and coverage (Schweikert et al., 2012), were applied, overlapped peaks were retained as real binding sites. Peaks identified by two independent peak calling runs (TAS & Cisgenome or MACS & QuEST) were merged based on intersection according to (Schweikert et al., 2012), by which both precision and coverage improved slightly. Briefly, overlapping peaks among two methods

were combined together to form new merged regions, start position of the merged peaks were defined as the minimum of all start position of its overlapping peaks, end position set to the maximum of all end positions. Peaks identified for same TF from different conditions and experiments were combined.

The peak regions determined by the above method were compared to TAIR10 gene annotation to associate TFs binding sites to putative target genes using Perl script. For miRNAs, our previously computational predicted TSSs were used to refine the miRBase miRNA annotation. Genes that has a binding site locate within 3kb upstream of its transcription start position to 1kb downstream of its transcriptional end position were then collected as the TF bound ones (Kaufmann et al., 2009; Wuest et al., 2012). If one peak can be assigned to more than two genes, the one has smaller distance between its TSS and the binding peak is reserved.

Network construction and topology study

Network was generated by Cytoscape 3 (Cline et al., 2007). Degree distributions of the whole network and miRNAs were generated by Cytoscape. Significantly overrepresented motifs were discovered by comparing real network to 1000-randomized networks, Mfinder v1.2 (Kashtan et al., 2004) was used. Only motifs with Z-score greater than 5 and appear more than 500 times in the real network were considered significant.

MiRNA conservation study and two-TF module study

Lists of *Arabidopsis* TFs were retained from AtTFDB and PlantTFDB (Davuluri et al., 2003; Jin et al., 2014). Combined miRNA conservation information was gathered from miRBase. Pri-miRNA expression data in seedling and rosette leaf from eight developmental stages were derived from mirEX2 (Bielewicz et al., 2012). For each two-TF module, frequency of genes bound by both TFs was calculated using Perl script and plotted by R for four gene types: conserved miRNAs, non-conserved miRNAs, TFs and other genes.

References

Addo-Quaye, C., Eshoo, T.W., Bartel, D.P., and Axtell, M.J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol* 18, 758-762.

Addo-Quaye, C., Miller, W., and Axtell, M.J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25, 130-131.

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* 121, 207-221.

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509-512.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101-113.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-297.

Bielewicz, D., Dolata, J., Zielezinski, A., Alaba, S., Szarzynska, B., Szczesniak, M.W., Jarmolowski, A., Szweykowska-Kulinska, Z., and Karlowski, W.M. (2012). mirEX: a platform for comparative exploration of plant pri-miRNA expression data. *Nucleic Acids Res* 40, D191-197.

Blais, A., and Dynlacht, B.D. (2005). Constructing transcriptional regulatory networks. *Gene Dev* 19, 1499-1511.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.R., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947-956.

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X.D., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., *et al.* (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463, 318-U368.

Chellappan, P., Xia, J., Zhou, X., Gao, S., Zhang, X., Coutino, G., Vazquez, F., Zhang, W., and Jin, H. (2010). siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Res* 38, 6883-6894.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2, 2366-2382.

Dai, X., and Zhao, P.X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 39, W155-159.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4, 25.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J. (2004). A gateway-compatible yeast one-hybrid system. *Genome Res* *14*, 2093-2101.

Deplancke, B., Mukhopadhyay, A., Ao, W.Y., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S., Hope, I.A., *et al.* (2006). A gene-centered *C. elegans* protein-DNA interaction network. *Cell* *125*, 1193-1205.

Ding, J., Li, D., Ohler, U., Guan, J., and Zhou, S. (2012). Genome-wide search for miRNA-target interactions in *Arabidopsis thaliana* with an integrated approach. *BMC Genomics* *13 Suppl 3*, S3.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* *30*, 207-210.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., *et al.* (2007). High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS One* *2*, e219.

German, M.A., Pillay, M., Jeong, D.H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta, K., German, R., *et al.* (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* *26*, 941-946.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* *431*, 99-104.

Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., *et al.* (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* *42*, D78-85.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature* *407*, 651-654.

Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., and Wong, W.H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* *26*, 1293-1300.

Ji, H.K., and Wong, W.H. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* *21*, 3629-3636.

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* *42*, D1182-1187.

Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant MicroRNAs and their targets, including a stress-induced miRNA. *Molecular Cell* *14*, 787-799.

Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* *20*, 1746-1758.

Kaufmann, K., Muino, J.M., Jauregui, R., Airoidi, C.A., Smaczniak, C., Krajewski, P., and Angenent, G.C. (2009). Target Genes of the MADS Transcription Factor SEPALLATA3: Integration of Developmental and Hormonal Pathways in the Arabidopsis Flower. *PLoS Biology* 7, 854-875.

Kim, J.H., Woo, H.R., Kim, J., Lim, P.O., Lee, I.C., Choi, S.H., Hwang, D., and Nam, H.G. (2009). Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in Arabidopsis. *Science* 323, 1053-1057.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 42, D68-73.

Laajala, T.D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., and Elo, L.L. (2009). A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* 10.

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

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843-854.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23, 4051-4060.

Li, J., Liu, Z.J.J., Pan, Y.C.C., Liu, Q., Fu, X., Cooper, N.G.F., Li, Y.X., Qiu, M.S., and Shi, T.L. (2007). Regulatory module network of basic/helix-loop-helix transcription factors in mouse brain. *Genome Biol* 8, R244.

Martinez, N.J., Ow, M.C., Barrasa, M.I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F.P., Ambros, V.R., and Walhout, A.J. (2008). A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev* 22, 2535-2549.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* 298, 824-827.

Newman, M.E.J. (2003). The structure and function of complex networks. *Siam Rev* 45, 167-256.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513-520.

Rubio-Somoza, I., and Weigel, D. (2011). MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci* 16, 258-264.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., *et al.* (2013). ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res* 41, D987-990.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E.M. (2007). A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Gene Dev* 21, 436-449.

Schommer, C., Palatnik, J.F., Aggarwal, P., Chetelat, A., Cubas, P., Farmer, E.E., Nath, U., and Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol* 6, e230.

Schweikert, C., Brown, S., Tang, Z., Smith, P.R., and Hsu, D.F. (2012). Combining multiple ChIP-seq peak detection systems using combinatorial fusion. *BMC Genomics* 13 *Suppl* 8, S12.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31, 64-68.

Stocks, M.B., Moxon, S., Mapleson, D., Woolfenden, H.C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T., and Moulton, V. (2012). The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 28, 2059-2061.

Sun, X., Dong, B., Yin, L., Zhang, R., Du, W., Liu, D., Shi, N., Li, A., Liang, Y., and Mao, L. (2013). PMTED: a plant microRNA target expression database. *BMC Bioinformatics* 14, 174.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* 5, 829-834.

Vermeirssen, V., Barrasa, M.I., Hidalgo, C.A., Babon, J.A.B., Sequerra, R., Doucette-Stamm, L., Barabasi, A.L., and Walhout, A.J.M. (2007a). Transcription factor modularity in a gene-centered *C. elegans* core neuronal protein-DNA interaction network. *Genome Research* 17, 1061-1071.

Vermeirssen, V., Deplancke, B., Barrasa, M.I., Reece-Hoyes, J.S., Arda, H.E., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Brent, M.R., *et al.* (2007b). Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat Methods* 4, 659-664.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669-687.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. *Cell* 138, 738-749.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75, 855-862.

Wilbanks, E.G., and Facciotti, M.T. (2010). Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS One* 5, e11471.

Wu, G., Park, M.Y., Conway, S.R., Wang, J.W., Weigel, D., and Poethig, R.S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* 138, 750-759.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. *Mol Cell* 38, 465-475.

Wuest, S.E., O'Maoileidigh, D.S., Rae, L., Kwasniewska, K., Raganelli, A., Hanczaryk, K., Lohan, A.J., Loftus, B., Graciet, E., and Wellmer, F. (2012). Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proc Natl Acad Sci USA* 109, 13452-13457.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. *Plant Physiol* 138, 2145-2154.

Zhang, H., He, H., Wang, X., Yang, X., Li, L., and Deng, X.W. (2011). Genome-wide mapping of the HY5-mediated gene networks in Arabidopsis that involve both transcriptional and post-transcriptional regulation. *Plant J* 65, 346-358.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9.

Zhang, Y., Mayba, O., Pfeiffer, A., Shi, H., Tepperman, J.M., Speed, T.P., and 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 Genet* 9.

Zhang, Z., Yu, J., Li, D., Liu, F., Zhou, X., Wang, T., Ling, Y., and Su, Z. (2010). PMRD: plant microRNA database. *Nucleic Acids Res* 38, D806-813.

Zhao, X., Zhang, H., and Li, L. (2013). Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis. *Genomics* 101, 187-194.

Figures

Figure 1. Construction of the miRNA-centered gene regulatory network in *Arabidopsis*

- (A) Schematic diagram of the miRNA-centered regulatory gene network. Three classes of genes (TFs, miRNAs, and miRNA target genes) and three interactions (MTIs, TMIs, TTIs) were included to build the network.
- (B) Flow chart of the procedures to collect 2189 miRNA-target interactions (MTIs). High/low-confident MTIs were first collected from computational prediction methods. Low-confident predictions with degradome data support were retained and combined with highly confident predictions. Finally, experimentally validated MTIs were incorporated, producing 2189 reliable MTIs.
- (C) Functional distribution of 35 TFs; twelve functional clusters were shown.
- (D) miRNA-centered regulatory gene network. Nodes indicate genes; lines indicate regulations. Specifically, the color of nodes represents different gene classes; all interactions between these genes are depicted as gray directional lines.

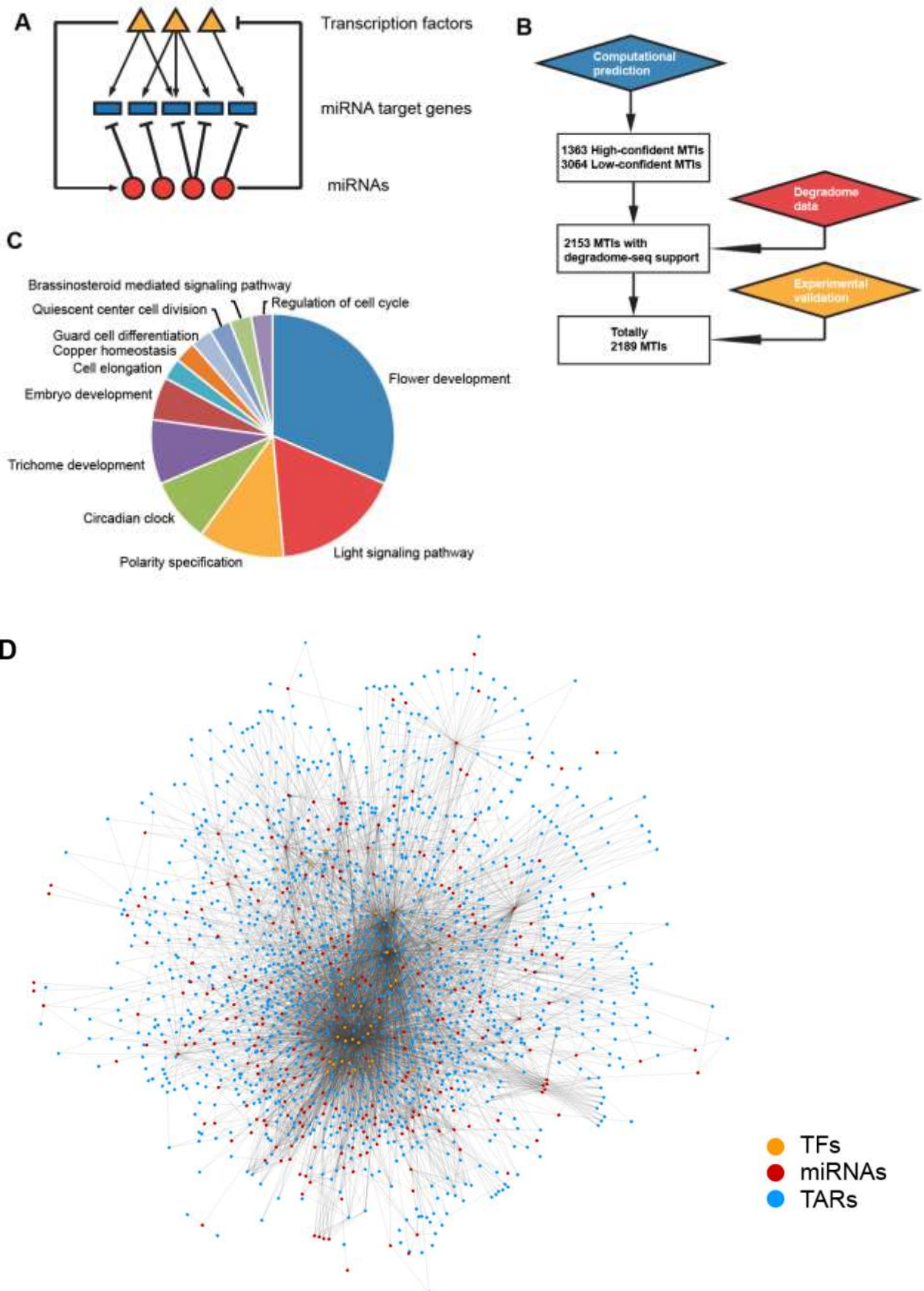


Figure 2. Topological analysis of the network and target gene distribution of 35 TFs

- (A) Degree (both in and out) distribution of the whole network. Degree k was plotted against frequency of nodes $P(k)$ for all genes. Transformed data were able to fit power law.
- (B) In-degree distribution of miRNAs is shown in both original and transformed layouts, with in-degree k -in and frequency $P(k$ -in) indicated. Transformed data were fitted by an exponential distribution.
- (C) Out-degree distribution of miRNAs is shown in both original and transformed layouts, with out-degree k -out and frequency $P(k$ -out) indicated. Transformed data were fitted by power law.
- (D) Eleven significantly over-represented 3- and 4-node motifs with count in real network and 1000-time randomized networks. Structures of motifs are shown along the X-axis.
- (E) Quantitative distribution of target gene for 35 TFs. Upper part indicates number of miRNAs targeted by 35 TFs and the number of TFs that are predicted targets of these miRNAs, which are denoted as the indirect TF targets. Lower part shows the number of genes (except miRNAs) that are directly bound to 35 TFs, and the number of TFs within them. Black line represents the ratio of indirect TF targets versus the number of miRNAs bound by each TF. Gray line shows the ratio between the number of direct TF targets and the number of all direct targets except miRNAs. Scales of the black and gray lines are depicted on the right. Functional clusters for 35 TFs are indicated at the bottom.

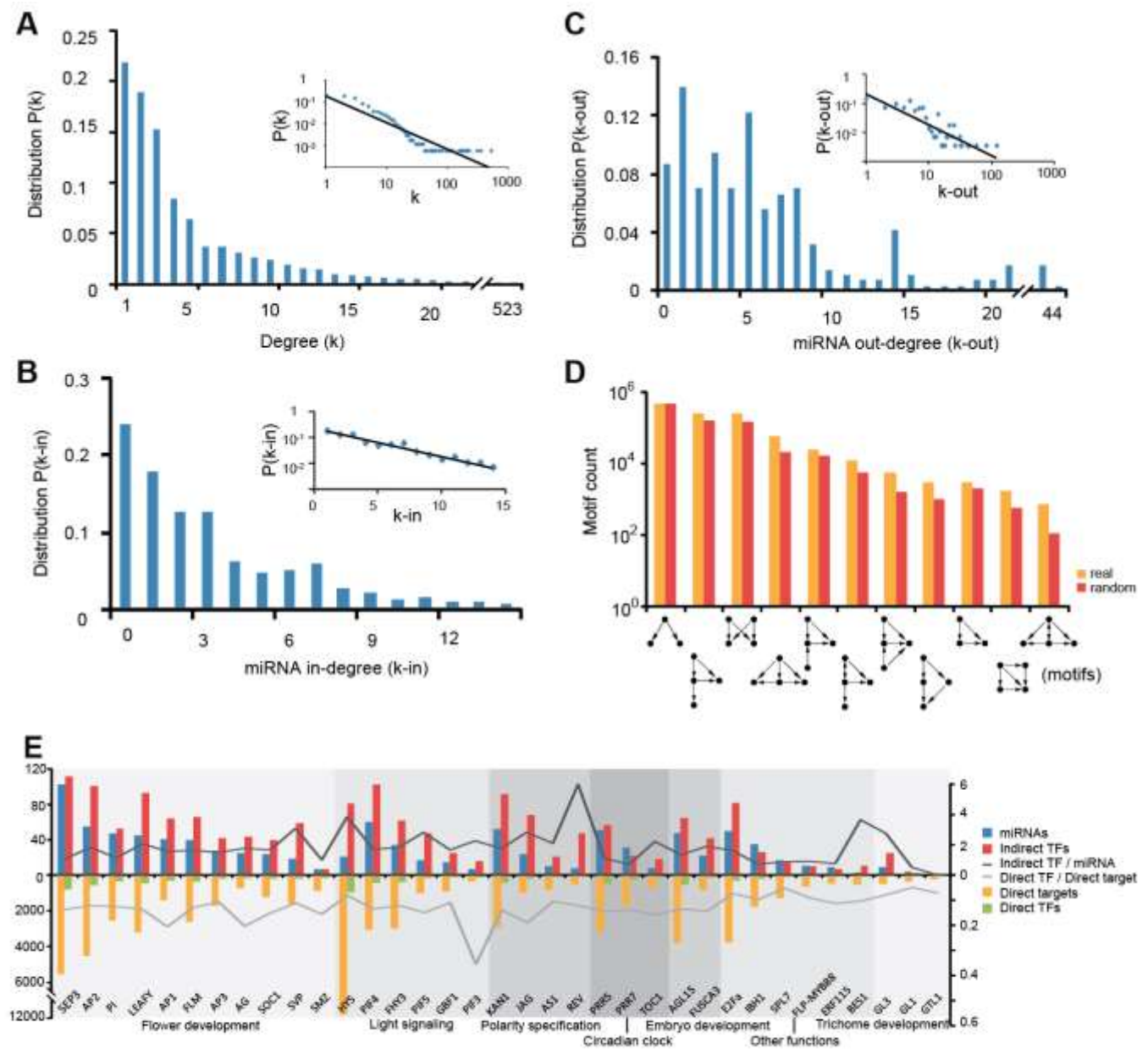


Figure 3. miRNAs facilitate crosstalk between different TFs

- (A) Quantitative distribution of TF ChIP binding sites on miRNAs, all TFs, and other genes. Only genes with fewer than 15 TF binding sites were included.
- (B) Quantitative distribution of TF binding sites at miRNA loci (TF binding) and TFs among target genes for miRNAs (TF target) with different conservation depicted at the bottom. miRNAs were grouped into four classes based on conservation: 1) miRNAs only discovered in *Arabidopsis thaliana* and/or *A. lyrata*; 2) miRNAs discovered in at least one other rosoid specie excepting *Arabidopsis*; 3) miRNAs discovered in both rosoid and asterid species; and 4) miRNAs discovered not only in dicot but monocot or other species). Error bar indicates 95% confident interval.
- (C) Heat map shows the proportion of miRNAs that are simultaneously bound by different two-TF modules. Each cell of the heat map indicates one two-TF module, with a total 595 combinations in each half (lower-left or upper-right.). Conserved miRNAs are depicted on the lower-left, while non-conserved miRNAs are shown on the upper-right. Intensity of the color for each cell means the proportion of miRNAs (conserved or non-conserved) simultaneously bound by each two-TF module. Intensity varies from 0 to 20%.
- (D) Heat map shows the proportion of all TFs and other genes that are simultaneously bound by each two-TF module chosen from the 35 TFs. TFs are depicted on the lower left, and the remaining genes on the upper-right. Intensity varies from 0 to 20%.

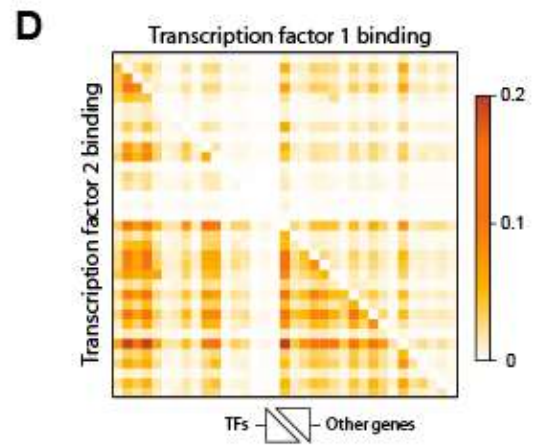
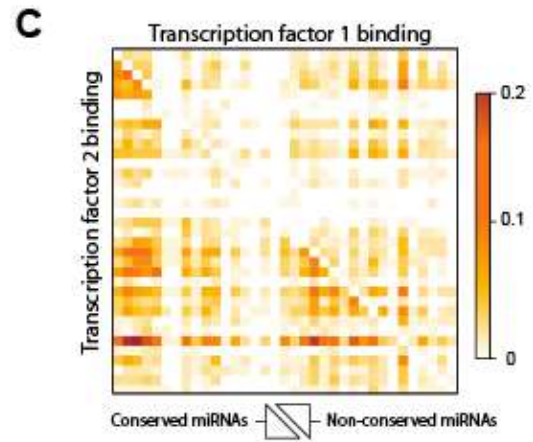
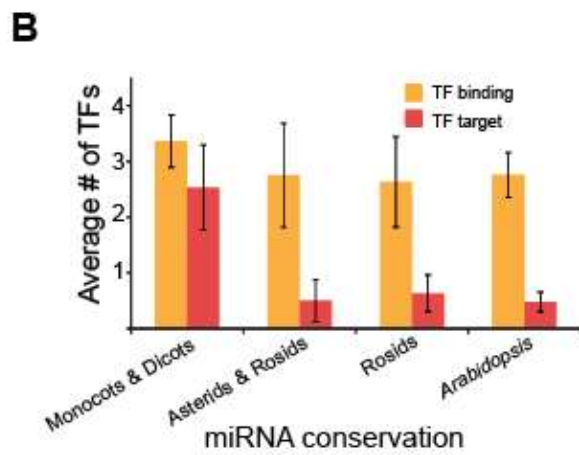
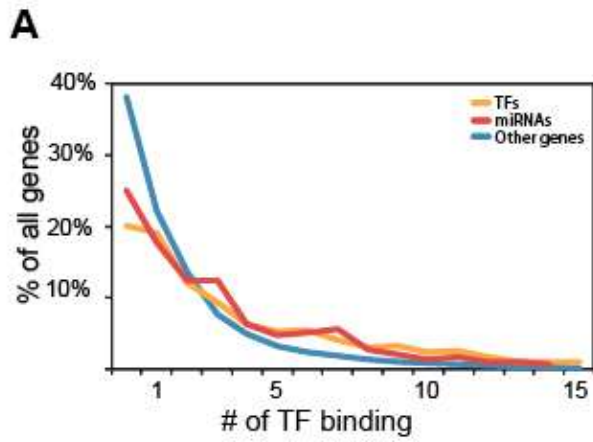
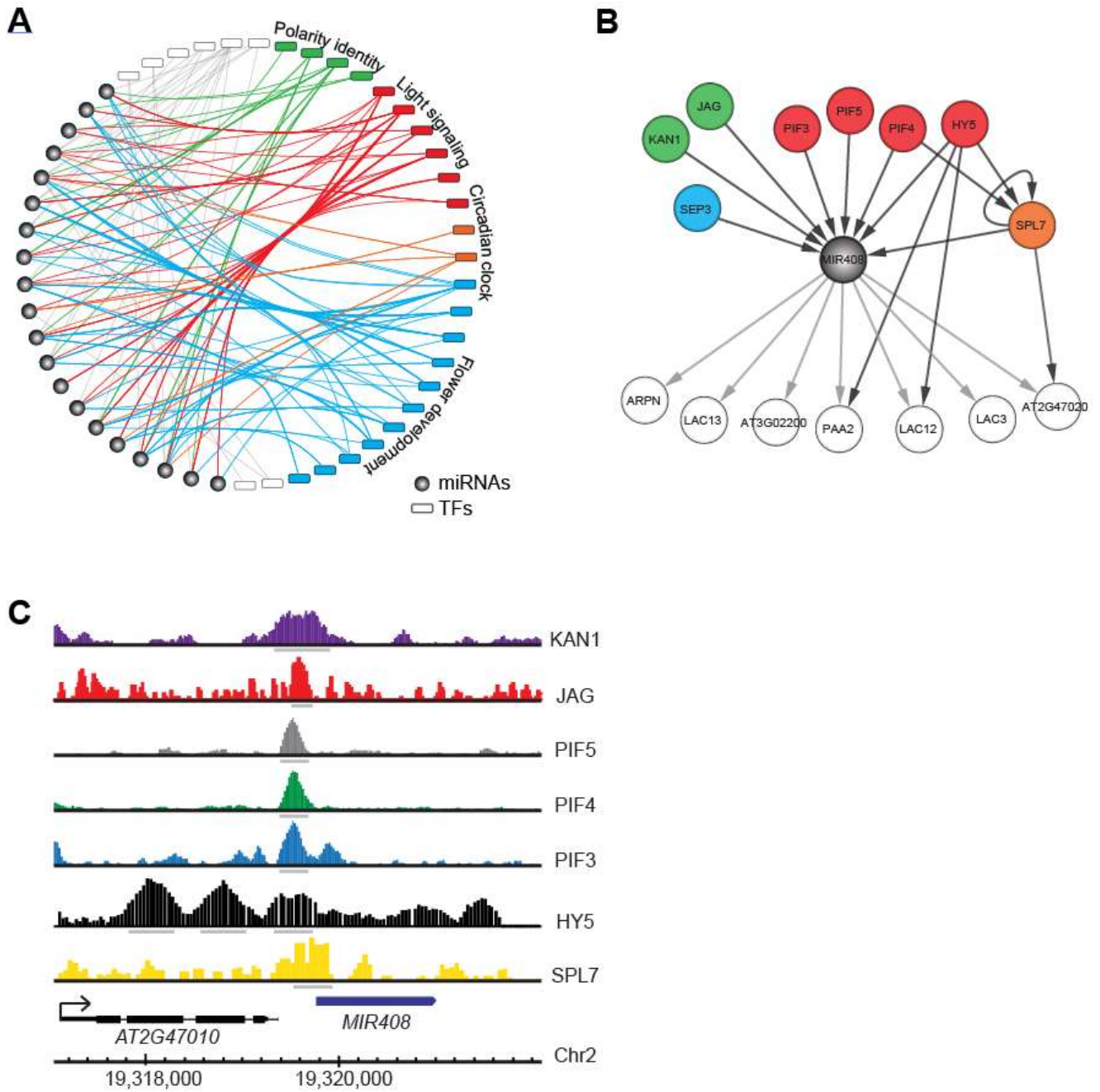


Figure 4. Light-signaling TFs regulated sub-network and *MIR408*-centered gene network.

- (A) Sub-network includes all 20 miRNAs that are targeted by light-signaling TFs, and all other TFs it recruits. TFs within four functional clusters that are significantly highly connected to these miRNAs are highlighted; the bended edges indicate their transcriptional regulations.
- (B) A *MIR408*-centered regulatory gene network. Eight TFs that bind to *MIR408* gene loci are shown at the top and highlighted as corresponding to different functional clusters. Seven target genes of *MIR408* are shown at the bottom. The black directed edges represent TMIs, while the gray edges show post-transcriptional regulation by *MIR408* (MTIs).
- (C) ChIP binding profile of eight TFs at *MIR408* gene loci. Gene structure of *MIR408* and its neighbor gene together with their genomic coordinate are shown at the bottom. The gray bar under each TF binding pattern indicates a corresponding binding peak.



Supplemental materials












Supplemental Table 1. List of 35 TFs included in the miRNA-centered regulatory gene network in *Arabidopsis*

TF	Data type	Accession	Developmental stage	Sampling Time	Year	TF domain	TF function	Binding sites
AG	ChIP-seq	GSE45938	inflorescences	4 WEEKS	2013	MADS	floral development	765
AGL15	ChIP-chip	GSE17717	embryonic culture tissue		2009	MADS	embryogenesis	4362
AP1	ChIP-seq	GSE20176	apical meristematic tissue		2010	MADS	floral development	1295
	ChIP-seq	GSE46986	inflorescences	2,4,8 DAYS after induction	2013			587
AP2	ChIP-chip	E-MEXP-2653	rosette leaves	5 WEEKS	2010	AP2/ERF	floral development	2203
	ChIP-seq	GSE21301	inflorescences		2010			3951
AP3	ChIP-seq	GSE38358	inflorescences	4 WEEKS	2012	MADS	floral development	1446
AS1	ChIP-chip	GSE44872	seedling	14 DAYS	2013	MYB	polarity identity	923
BES1	ChIP-chip	GSE24684	seedling	2 WEEKS	2010	BZR	brassinosteroid (BR) signalling	544
E2Fa	ChIP-seq	GSE53422			2013	E2F/DP	regulation of cell cycle	3913
ERF115	ChIP-seq	GSE48793		7 DAYS	2013	AP2/ERF	Quiescent center cell division	475
FHY3	ChIP-seq	GSE30711	seedling	4 DAYS	2011	ZINC-FINGER	light signaling	3166
FLM	ChIP-seq	GSE48082	seedling		2013	MADS	floral development	2850
FLP-MYB88	ChIP-chip	GSE19763	shoots	10 DAYS	2010	MYB	guard mother cells differentiation	771
FUSCA3	ChIP-chip	GSE43291	embryonic culture tissue		2013	B3	embryogenesis	982
GBF1	ChIP-chip	GSE36965	seedling	4 DAYS	2012	bZIP	light signaling	1030
GL1	ChIP-chip	GSE13090	green tissue	3 WEEKS	2008	MYB	trichome development	376
GL3	ChIP-chip	GSE13090	green tissue	3 WEEKS	2008	bHLH	trichome development	511
GTL1	ChIP-chip	GSE40519	whole aerial parts	12 DAYS	2012	Trihelix	trichome development	236
HY5	ChIP-chip	GSE24974	seedling	4 DAYS	2010	bZIP	light signaling	11797
IBH1	ChIP-seq	GSE51120	seedling	10 DAYS	2013	bHLH	cell and organ elongation	1861
JAGGED	ChIP-seq	GSE51537	inflorescence tips		2013	C2H2	polarity identity	1047
KANADI1	ChIP-seq	GSE48081		10 DAYS in liquid culture	2013	G2-like	polarity identity	3193
LEAFY	ChIP-chip	GSE28063	seedling	9 DAYS & 19 DAYS	2011	LFY	floral development	2912
	ChIP-seq	GSE24568	seedling	15 DAYS	2010			1691
PI	ChIP-seq	GSE38358	inflorescences	5 WEEKS	2012	MADS	floral development	2593
PIF3	ChIP-seq	GSE39215	seedling	2 DAYS	2012	bHLH	light signaling	703
PIF4	ChIP-seq	GSE35315	seedling		2012	bHLH	light signaling	3529
PIF5	ChIP-seq	GSE35059	seedling	14 DAYS	2012	bHLH	light signaling	1083
PRR5	ChIP-seq	GSE36361	whole plant	18 DAYS	2012	CCT	circadian clock	3582
PRR7	ChIP-seq	GSE49282	seedling	15 DAYS	2013	CCT	circadian clock	1890
REV	ChIP-seq	GSE26722	seedling		2011	homeobox	polarity identity	520
	ChIP-seq	GSE46986	inflorescences	2,4,8 DAYS after induction	2013			3617
SMZ	ChIP-chip	E-MEXP-2068	seedling	9 DAYS	2009	AP2/ERF	floral development	1016
SOC1	ChIP-chip	GSE33297	seedling	9 DAYS	2011	MADS	floral development	796
	ChIP-seq	GSE45846			2013			782
SPL7	ChIP-seq	GSE45213	seedling	7 DAYS	2013	SBP-box	copper homeostasis	1535
SVP	ChIP-chip	GSE33297	seedling	9 DAYS	2011	MADS	floral development	1162
	ChIP-seq	GSE33120	inflorescences	2 WEEKS	2011			729
TOC1	ChIP-seq	GSE35952	seedling	2 WEEKS	2012	CCT	circadian clock	803

Supplemental Table 2. List of MTIs, TMIs, and TTIs

This supplemental table includes 3 sheets and over 6,000 lines, is available from the author upon request.

Supplemental Table 3. Over-represented network motifs

Size	Structure	# in real network	# in randomized network	Z-score	p-value
3		475281	474414.23	9.3757	0
4		269179	159349.89	6.4727	0
4		267145	145015.71	25.997	0
4		56562	21378.83	9.8978	0
4		24555	16982.15	5.7114	0
4		11841	5542.98	6.2765	0
4		5692	1636.57	16.599	0
4		3010	982.05	20.24	0
3		2880	1943.89	8.8513	0
4		1737	591.68	13.658	0
4		750	115.05	12.589	0

Chapter5. microRNA408 Is Critical for the HY5-SPL7 Gene Network That Mediates Coordinated Response to Light and Copper¹

(This section of work was done in collaboration with Dr. Huiyong Zhang)

¹Formatted as a co-author manuscript accepted and in press as:

Zhang H, Zhao X, Li J, Cai H, Deng XW, Li L. *Plant Cell* 12 (2014)

doi:10.1105/tpc.114.127340

Abstract

Light and copper are important environmental determinants for plant growth and development. Despite the wealth of knowledge on both light and copper signaling, molecular mechanisms for integrating the two pathways are poorly understood. Here, in collaboration with Dr. Zhang, we demonstrate in *Arabidopsis* an interaction between *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 7 (SPL7)* and *ELONGATED HYPOCOTYL 5 (HY5)*, which mediates copper and light signaling, respectively. Through whole genome chromatin immunoprecipitation (ChIP) and RNA-sequencing analyses, I elucidated the *SPL7* regulon and compared it with that of *HY5*. I found that the two transcription factors co-regulate many genes such as those involved in anthocyanin accumulation and photosynthesis. Specifically, I show that *SPL7* and *HY5* act coordinately to transcriptionally regulate *MIR408*, which results in differential expression of miR408 as well as its target genes in response to changing light and copper conditions. We demonstrate that this regulation is tied to copper allocation to the chloroplast and plastocyanin level. Finally, we found that constitutively activated miR408 rescues distinct developmental defects of the *hy5*, *spl7*, and *hy5 spl7* mutants. These findings revealed a previously uncharacterized light-copper crosstalk mediated by a *HY5-SPL7* network. Further, integration of transcriptional and post-transcriptional regulations is critical for governing proper metabolism and development in response to combined copper and light signaling.

Introduction

Essentially sessile in nature, plants have evolved sophisticated mechanisms to maintain metabolic homeostasis and remarkable capacity to reprogram development when resource levels vary. Light and copper are among the most important environmental factors for plant growth. In addition to providing the source of energy, light regulates many plant processes and induces massive reprogramming of the transcriptome (Chen et al., 2004; Jiao et al., 2007). Genetic and molecular approaches undertaken in the model plant *Arabidopsis* have identified many regulatory factors and pathways required for proper light signaling. *HY5* encodes a bZIP type transcription factor that functions downstream of multiple photoreceptors to promote photomorphogenesis (Oyama et al., 1997; Ang et al., 1998). Detailed characterization of the *hy5* mutants revealed a myriad of phenotypic defects, including elongated hypocotyl, reduced pigment content, aberrant chloroplast development, altered root morphology, and compromised hormonal responses (Oyama et al., 1997; Cluis et al., 2004; Vandenbussche et al., 2007). HY5 is known to bind to G-box-like motifs in the light responsive promoters (Chattopadhyay et al., 1998; Yadav et al., 2002; Shin et al., 2007; Song et al., 2008). Using ChIP coupled with tiling microarray analysis, it has been shown that HY5 binds to approximately 40% of the coding loci in *Arabidopsis* and detectably impacts the expression level of approximately 3,000 genes (Lee et al., 2007; Zhang et al., 2011).

As a transition metal, copper is an essential cofactor for numerous proteins. The most abundant copper protein in plants is plastocyanin (PC), which transfers electrons from the

cytochrome *b₆f* complex to photosystem I (Burkhead et al., 2009). Copper is also used as a cofactor by plant proteins involved in neutralizing reactive oxygen species, lignification of the cell wall, ethylene perception, and formation of phenolics in response to pathogens (Burkhead et al., 2009). Regulating the abundance of these proteins is important to maintain copper homeostasis and prioritize the use of cellular copper in plants. Studies in the green alga *Chlamydomonas* revealed CRR1 as a zinc finger transcription factor that is specifically activated under copper deficiency (Kropat et al., 2005). The GTAC motif found in CRR1 targets is recognized as the core of copper-response elements in diverse plants (Quinn and Merchant, 1995; Quinn et al., 1999; Kropat et al., 2005; Nagae et al., 2008; Yamasaki et al., 2009). In *Arabidopsis*, *SPL7* is orthologous to *CRR1* (Yamasaki et al., 2009). In a *spl7* mutant, many of the genes related to copper homeostasis, including the transporters COPT1 and COPT2, the copper chaperones CCH and CCS, as well as the copper/zinc superoxide dismutases CSD1 and CSD2, are dysregulated (Yamasaki et al., 2009; Bernal et al., 2012). It was demonstrated that copper specifically inhibits DNA binding activity of both CRR1 and SPL7, and prevents transcription activation in vitro (Sommer et al., 2010). SPL7 is thus likely a copper sensor in *Arabidopsis* that regulates gene expression in response to changing cellular copper levels (Sommer et al., 2010).

In addition to transcriptional regulators, microRNAs (miRNAs) are emerging as a class of sequence-specific, trans-acting regulatory small RNA molecules that modulate gene expression at the post-transcription level. After processing from stem-loop-structured precursors and integration into the RNA-induced silencing complex, miRNAs function in general as gene repressors by directing cleavage or translational repression of the target

transcripts (Voinnet, 2009). The large numbers of miRNAs in eukaryotes entail increased complexity of gene regulatory mechanisms. In plants, miRNAs have been implicated in light and copper signaling. HY5 occupancy at the promoter region of eight miRNAs was noticed in a global ChIP analysis (Zhang et al., 2011). Further, several miRNAs were predicted to target transcripts encoding copper proteins (Jones-Rhoades and Bartel, 2004; Sunkar et al., 2006; Yamasaki et al., 2007; Abdel-Ghany and Pilon, 2008). Their expression was found to be induced by copper deficiency in a *SPL7*-dependent manner (Yamasaki et al., 2009; Bernal et al., 2012). It was thus hypothesized that these miRNAs are used to repress nonessential copper proteins (such as CSD1 and CSD2) and their chaperone to save copper for essential proteins such as PC under impending deficiency (Yamasaki et al., 2007).

However, how light and copper signals work together to regulate gene expression for optimal plant growth and development is poorly understood. In the current work, we demonstrate the interaction between *SPL7* and *HY5*, providing evidence for the integration of copper and light signaling. By elucidating the *SPL7* regulon through ChIP- and RNA-sequencing and comparing with previously identified *HY5* regulon (Zhang et al., 2011), I found that these two transcription factors have significant overlap in terms of the genes that they directly bind and the genes whose expression levels they regulate. Further, we show that miR408 is a critical component of the *SPL7-HY5-MIR408* network and that constitutively activated miR408 rescues the developmental defects of the *hy5*, *spl7*, and *hy5 spl7* mutants. These findings thus revealed a molecular basis incorporating transcriptional and post-transcriptional regulation for coordinated plant responses to changing copper and light regimes.

Results

SPL7* Interacts with *HY5

We observed that *Arabidopsis* seedlings respond decisively to combined light and copper regimes when grown under relatively high (HL; $170 \mu\text{molm}^{-2}\text{s}^{-1}$), or low light intensity (LL; $40 \mu\text{molm}^{-2}\text{s}^{-1}$), and relatively high (HC; $5 \mu\text{M}$), or low copper concentration (LC; $0.1 \mu\text{M}$). Although morphology was primarily influenced by light with seedling grown in HL having shortened hypocotyl and root but increased fresh weight (Supplemental Figures 1A and 1B), contents of essential metabolites (chlorophyll, anthocyanin, and glucose) were influenced by both light and copper (Supplemental Figures 1C to 1E). Overall, seedlings displayed distinctive morphological and metabolic profiles under the four combinations of light and copper conditions (Supplemental Figure 1F). These results indicate that there is a crosstalk between the light and copper signaling pathways for orchestrating plant growth and metabolism.

To begin elucidating the light-copper crosstalk, I examined the developmental expression profiles of *HY5* and *SPL7* in *Arabidopsis* and found that both are expressed in the six examined organ types (Supplemental Figure 2A). This result prompted us to investigate whether the two transcription factors interact with each other. Dr. Zhang performed a yeast two-hybrid assay in which *HY5* and *SPL7* were fused in-frame to the binding domain of LexA and the activation domain of B42, respectively. Specifically increased reporter activity was observed when both the *HY5* and *SPL7* fusion proteins were expressed in the same yeast cells (Figure 1A). Dr. Zhang also performed an *in vitro* pull down assay using

recombinant *SPL7* and *HY5*, which showed that GST-tagged *HY5*, but not GST alone, was able to pull down 6×His-tagged *SPL7* (Figure 1B). Further, Dr. Zhang generated transgenic *Arabidopsis* plants expressing N-terminal FLAG-tagged *SPL7* driven by the Cauliflower Mosaic Virus 35S promoter (*35S:FLAG-SPL7*). Using these plants, we performed a coimmunoprecipitation (co-IP) assay and found that the anti-FLAG and the anti-*HY5* antibodies could specifically pull down *HY5* and FLAG-*SPL7* from plant extracts, respectively (Figure 1C). Together these results demonstrate the *SPL7*-*HY5* physical interaction.

Next, I tested whether *SPL7* and *HY5* influence each other's expression. I extracted the 2 Kb sequences upstream of the transcription start site (TSS) of *SPL7* and *HY5* as approximation of their promoters, respectively. Scanning the *SPL7* promoter identified a cluster of five *HY5*-binding motifs, which coincides with a strong *HY5* binding peak at this region according to the whole genome occupancy data (Zhang et al., 2011; Supplemental Figure 2B). This observation was confirmed by electrophoretic mobility shift assay (EMSA) using recombinant *HY5* and a DNA fragment from the *SPL7* promoter region containing all five motifs as the probe (Supplemental Figure 2C) as well as quantitative PCR analysis of the *SPL7* promoter region following ChIP (ChIP-qPCR) by the anti-*HY5* antibody (Figure 1D). Monitoring transcript levels by quantitative reverse transcription coupled PCR (qRT-PCR) revealed that *SPL7* expression increases over two-fold in *hy5* (Figure 1E). These results indicate that *HY5* directly binds to the *SPL7* promoter and negatively regulates its transcription.

Scanning the proximal promoter region of *HY5* revealed no GTAC motifs. Consistently,

ChIP-qPCR analysis revealed that *SPL7* does not bind to the *HY5* promoter. However, we found that *HY5* abundance showed an over two-fold increase in the *spl7* mutant compared to wild type at both the mRNA and protein levels (Figures 1F and 1G), indicating that *HY5* is negatively regulated by *SPL7*. Together, these results demonstrate that *SPL7* and *HY5* interact with each other both physically and genetically, which suggests a feedback mechanism for linking a light and copper responsive gene network.

Genome-wide Analyses of the *SPL7* Regulon

To facilitate global identification of *SPL7* binding sites in *Arabidopsis*, Dr. Zhang generated transgenic lines expressing FLAG-tagged *SPL7* in the *spl7* mutant background (*35S:FLAG-SPL7/spl7*). Characterization of the transgenic lines indicated that *FLAG-SPL7* is properly expressed, levels of *SPL7*-regulated miR398 (Yamasaki et al., 2009) and miR408 (Zhang and Li, 2013) restored, and growth defects of the *spl7* mutant, which include reduced fresh weight and shorter root when grown under LC, rescued. Thus, the *FLAG*-tagged *SPL7* transgene is functional *in vivo*.

Using the *35S:FLAG-SPL7/spl7*, we performed anti-FLAG ChIP-sequencing and generated 17.8 million reads. As a control, the ChIP-sequencing procedure was applied to the *spl7* mutant, which yielded 14.8 million reads (Supplemental Table 1). A total of 1,535 specific *SPL7*-binding peaks were identified and found to locate predominantly near the TSS (Figure 2A; Supplemental Table 2, sheet 1). For verification, we performed ChIP-qPCR analysis on randomly selected *SPL7*-occupied regions (Supplemental Figure 3A) and confirmed *SPL7* binding for fifteen of the sixteen tested loci (94%; Figure 2B). The *SPL7*-

binding peaks overlap with 1,266 genes including ten miRNA genes (Supplemental Figure 3B; Supplemental Table 2, sheet 2), whereas 11% of the peaks reside in intergenic regions (Figure 2C). Of the binding sites assigned to genes, most are localized to the exons (48%) and the proximal promoter regions (22%; Figure 2C). Gene Ontology (GO) analysis revealed that these genes preferentially associate with GO terms such as “response to stimulus and stress”, “photosynthesis”, “regulation of biological quality”, and “post-embryonic development” (Supplemental Figure 3C). Regarding annotated pathways, photosynthesis, carbon fixation, nitrogen metabolism, and glyoxylate and dicarboxylate metabolism are among the most significantly enriched (Supplemental Figure 3D). These results indicate that *SPL7* target genes are involved in primary metabolism and responsive to environmental stimuli.

SPL7 is known to recognize DNA motifs containing the GTAC sequence (Yamasaki et al., 2009; Sommer et al., 2010). In the identified *SPL7*-binding sites, this tetranucleotide is significantly over-represented compared to random genome sequences (Figure 2D). Closer inspection revealed that the hexanucleotide in which GTAC is symmetrically flanked by A/T, but not other GTAC-encompassing hexanucleotides, is enriched in *SPL7*-binding sites (Figure 2D). By EMSA analysis, we confirmed that the A/TGTACT/A motif indeed has stronger affinity for *SPL7* than other similar sequences (Supplemental Figure 4A). Moreover, I discovered a novel motif overrepresented in the *SPL7* occupied regions that resembles DNA elements recognized by the zinc finger family of transcription factors to which *SPL7* belongs (Badis et al., 2008; Supplemental Figure 4B). Binding of *SPL7* to this motif was confirmed by EMSA (Supplemental Figure 4C). Further, systematic scanning of

the *SPL7* binding sites revealed that several other known *cis*-elements appear at a higher frequency than the genome average (Supplemental Figure 4D). Together, these results indicate that *SPL7* recognizes different classes of DNA motifs and acts with other factors for targeting a broad spectrum of genes.

Next, I conducted whole-transcriptome RNA-sequencing using wild type and *spl7* seedlings grown under either LC or HC conditions. I obtained a total of 158 million reads for the four tested samples (Supplemental Table 1) and made four pairwise comparisons, revealing that expression of 4,090 genes is influenced by copper (Supplemental Figure 5A). Clustering analysis revealed that these genes form four major groups (Figure 2E; Supplemental Table 2, sheets 3-6). There are roughly the same numbers of genes in each group (1,018 in group I, 1,147 in group II, 964 in group III, and 961 in group IV; Supplemental Figure 5B). To validate the RNA-sequencing data, I performed qRT-PCR on a handful of copper responsive as well as randomly selected genes. As shown in Supplemental Figure 6, the two methods generated highly agreeable results.

The four groups of genes exhibit distinctive and contrasting transcriptional profiles. Group I genes have lower levels in the WT/LC vs WT/HC comparison and higher levels in the *spl7*/HC vs *spl7*/LC comparison, indicating that these genes are induced by the HC condition. Further, most group I genes show higher levels in the *spl7*/HC vs WT/HC comparison but lower levels in the WT/LC vs *spl7*/LC comparison (Figure 2E), indicating that *SPL7* acts as a negative regulator for these genes. Interestingly, group III genes exhibit generally the opposite behavior as group I genes (Figure 2E), indicating these genes are induced by LC with *SPL7* acting as a positive regulator though exceptions exist. By

contrast, group II genes show lower levels in the *spl7*/HC vs WT/HC comparison, higher levels in the WT/LC vs *spl7*/LC comparison, but not much change in the WT/LC vs WT/HC and the *spl7*/HC vs *spl7*/LC comparisons (Figure 2E), indicating that these genes do not respond to change in copper regime but require *SPL7* to maintain proper levels. Thus, *SPL7* acts primarily as a positive regulator for these genes regardless of the copper condition. Again, the behavior of group IV are generally opposite to that of group II (Figure 2E), indicating expression of the majority of group IV genes is independent of copper with *SPL7* acting as a negative regulator. Thus, the general trend of Figure 2E revealed that *SPL7* could function either as a positive or negative regulator, both in the presence or absence of changing copper conditions.

Global analysis showed that the four groups of genes are preferentially associated with different GO terms and annotated pathways, suggesting that *SPL7* can broadly modulate primary metabolism and participates in stress responses (Supplemental Figure 5B). Interestingly, group III and IV are about twice enriched with genes bound by *SPL7* than group I and II, respectively (Figure 2E; Supplemental Figure 5B). Collectively, analyses of the *SPL7* regulon indicate that *SPL7* is a global regulator for proper molecular responses to copper regimes and participates in the regulation of other genes through distinct modes of action.

***SPL7* and *HY5* Co-regulate a Large Cohort of Genes**

Availability of the *SPL7* regulon allowed comprehensive elucidation of the light-copper crosstalk mediated by the *SPL7*-*HY5* feedback loop. I first sought to identify protein-coding

and miRNA genes directly targeted by both *SPL7* and *HY5* based on the global ChIP data reported here and previously (Zhang et al., 2011; Supplemental Figure 7A). Consistent with their interaction (Figure 1), the G-box, which is recognized by *HY5*, was found over-represented in the *SPL7* binding sites (Figure 3A). This observation prompted us to examine whether *SPL7* and *HY5* binding sites are clustered. To this end, I performed computational simulation and found that *SPL7*- and *HY5*-occupied regions are more likely to locate in close proximity than randomly selected genomic sequences (Figure 3B). Together, these two pieces of evidence support the notion that *SPL7* and *HY5* bind to a specific set of targets. Using a distance of 750 bp as the threshold, 586 genes, including *MIR159a*, *MIR398b*, and *MIR408*, were identified as the common targets of *SPL7* and *HY5* (Figure 3C; Supplemental Table 2, sheet 7). Global analysis revealed that the GO terms “photosynthesis” and “response to stimulus” are most significantly associated with these genes (Supplemental Figures 7B and 7C).

At the transcript level, *SPL7* and *HY5* each influence the expression of hundreds of genes based on RNA-sequencing data (Supplemental Table 2; Zhang et al., 2011). Comparison of these data revealed that *SPL7* and *HY5* commonly impact the transcript level of a set of 1,090 genes. Interestingly, *SPL7* and *HY5* modulate 582 of these genes in the opposite direction (up-regulated in *spl7* but down-regulated in *hy5* in comparison to wild type and vice versa) and 508 genes in the same direction (Figure 3D; Supplemental Table 2, sheet 8). Although the exact numbers in this comparison may not be taken literally as the RNA-sequencing experiments were not performed side by side, this result does reveal that *SPL7* and *HY5* act together to differentially regulate a large cohort of genes. To understand these

genes in more detail, I analyzed their associated biochemical pathways.

Production of anthocyanin as part of the flavonoid pathway requires 58 genes that encode the multiple enzymes involved (Solfanelli et al., 2006). According to the RNA-sequencing data, 29 of these genes are significantly influenced by either *SPL7* or *HY5*, encompassing essentially every enzymatic step of anthocyanin biosynthesis (Figure 3E). For example, the gene encoding chalcone synthase (CHS), the first committed enzyme in flavonoid biosynthesis, showed decreased expression in *hy5* and *spl7* mutants. Interestingly, because the pathway involves many isozymes that are encoded by members of small paralogous gene families, *SPL7* and *HY5* appear to work on different family members such that an overall up-regulation is achieved. Examples in this regard include genes encoding phenylalanine ammonia-lyase (PAL), the first and committed step in the phenyl propanoid pathway leading to flavonoid biosynthesis, and leucoanthocyanidin dioxygenase (LDOX), which is involved in proanthocyanin biosynthesis. Thus, a clear trend was observed that *SPL7* and *HY5* coordinately promote gene expression leading to increased anthocyanin synthesis.

I also analyzed genes related to photosynthesis, which revealed that *SPL7* and *HY5* both promote expression of genes involved in the light reactions (Supplemental Figure 8A) but repress genes involved in photorespiration (Supplemental Figure 8B). Interestingly, genes involved in the Calvin cycle are regulated by *SPL7* and *HY5* in different ways. It appears that genes responsible for the first stage of the cycle, in which a CO₂ molecule is incorporated into one of the two three-carbon molecules, are coordinately promoted by *SPL7* and *HY5*. By contrast, genes involved in the stage to regenerate ribulose-1,5-

bisphosphate are not positively regulated by *SPL7* and *HY5* (Supplemental Figure 8C). Taken together, our results indicate that *SPL7* and *HY5* are able to exert sophisticated regulations over their target genes involved in various pathways and processes, which presumably constitutes the molecular basis for the light-copper crosstalk.

As an example, I tested *MIR408* in detail, which is encoded at a single locus (At2g47015) in *Arabidopsis*. Searching the proximal promoter region upstream of the TSS revealed a G-box (CACGTG) and an array of GTAC motifs that clearly coincide with the strong *HY5* and *SPL7* binding peaks, respectively (Figure 4A). Previously, we have demonstrated through EMSA that *SPL7* binds to the GTAC motifs in the *MIR408* promoter *in vitro* (Zhang and Li, 2013). Here we show through ChIP-qPCR analysis that the anti-FLAG antibody could pull down the *MIR408* promoter from *35S:FLAG-SPL7/spl7* but not *spl7* plants (Figure 4B). Likewise, the binding of *HY5* to the *MIR408* promoter was confirmed both *in vivo* by ChIP-qPCR and *in vitro* by EMSA (Figures 4B and 4C).

Next, Dr. Zhang and I investigated whether *SPL7* and *HY5* could bind simultaneously to the *MIR408* promoter *in vitro*. A DNA fragment (-292 to -250) that contains two GTAC motifs and the G-box (Figure 4A) was tested in EMSA for binding with *HY5* and *SPL7* by Dr. Zhang. As shown in Figure 4D, addition of both *SPL7* and *HY5* in one reaction produced a super-shift band with further reduced mobility than that by adding either protein alone. When two mutant versions of the probe (one in which the G-box is mutated from CACGTG to CTGCAG and the other the two GTAC motifs both mutated to CATG) were used as competitors, production of the super-shifted band was effectively abolished (Figure 4D). Together these results indicate that *SPL7* and *HY5* are able to interact together with

the same DNA molecule *in vitro*. Further, transformation of yeast cells containing the *LacZ* reporter gene placed under control of the *MIR408* promoter with either *HY5* or *SPL7* resulted in *LacZ* expression, whereas additive activation was achieved with co-transformation of both *SPL7* and *HY5* (Figure 4E). Together these data demonstrate that *SPL7* and *HY5* bind simultaneously to the *MIR408* promoter via the GTAC and G-box motifs, respectively.

***MIR408* Participates in Coordinated Copper and Light Response**

To functionally dissect the *SPL7-HY5* network, we focused on the expression of *MIR408* in different light conditions. Dr. Zhang examined miR408 abundance in response to changing levels of copper and light. Both RNA gel blot and qRT-PCR analyses revealed that miR408 is present at the highest level under the LC/HL condition, intermediate under LC/LL and HC/HL, and lowest under HC/LL (Figures 5A and 5B). Additionally, we found that induction of *MIR408* by LC/HL requires both *SPL7* and *HY5* as miR408 accumulation is impaired in the *spl7* mutant and partially so in *hy5* (Figures 5A and 5B). Previously the *LAC13* gene, which encodes a copper-containing laccase, was identified as a cleavable miR408 target (Abdel-Ghany and Pilon, 2008). We therefore investigated whether expression of *LAC13* is affected by growth conditions that impact miR408 abundance. Based on qRT-PCR analysis, we found that *LAC13* is influenced by copper level and light intensity in a pattern contrary to that of miR408 (Figure 5C). Interestingly, in both *spl7* and *hy5* mutants, *LAC13* becomes less sensitive to light and copper changes and remains at relatively high levels (Figure 5C). These results indicate that *MIR408* and its target gene are regulated by coordinated light and copper signaling.

Dr. Zhang generated two reporter constructs in which the β -glucuronidase (*GUS*) gene is fused with either the native *MIR408* promoter (*pMIR408:GUS*) or a mutated *MIR408* promoter where the G-box sequence CACGTG was mutated to CTGCAG (*pMIR408m:GUS*). Dr. Zhang then transformed these two reporters into *Arabidopsis* plants of different genotypes. In wild type seedlings expressing *pMIR408:GUS*, GUS activity is the strongest under the LC/HL condition, intermediate under LC/LL and HC/HL, and weakest under HC/LL (Figure 5D), which are consistent with our results of transcript analyses (Figures 5A and 5B). Further, *pMIR408m:GUS* generates drastically reduced GUS activity with abolished light responsiveness, which could be phenocopied by expressing the intact *pMIR408:GUS* in the *hy5* background (Figure 5D). Previously, we showed that induction of *pMIR408:GUS* by LC requires *SPL7* (Zhang and Li, 2013). When expressing *pMIR408:GUS* in the *spl7* mutant, the GUS activity exhibits an overall significant decrease with a weak light responsiveness still remaining in the cotyledons (Figure 5D). These results confirmed that *MIR408* activation in response to light and copper is controlled at the transcription level by *HY5* and *SPL7*, respectively.

The GUS staining pattern revealed differences between cotyledons and hypocotyls (Figure 5D). We thus sampled these two organs separately for qRT-PCR analysis, which shows that *miR408* level is prominent in wild type hypocotyls under LC/HL but under both LC/LL and LC/HL in cotyledons. In hypocotyls, *MIR408* and *LAC13* levels show a much stronger anti-correlation than in cotyledons. *SPL7* is clearly required for *MIR408* induction in both cotyledons and hypocotyls under all tested conditions. While this is also the case for *HY5* in the cotyledons, it only acts positively for *MIR408* expression in the hypocotyls under the

HL conditions. These observations are generally consistent with GUS activities that indicate HY5 as well as SPL7 is required for *MIR408* expression in the cotyledons while in the hypocotyls HY5 is only responsible for light induction of *MIR408* (Figure 5D). These results further indicate that environment-induced *MIR408* expression is specifically regulated in different organ types, which likely involves additional regulators.

To further analyze the joint effect of *HY5* and *SPL7* on *MIR408*, we generated *hy5 spl7* double mutant. Northern blot analysis revealed that accumulation of miR408 is essentially abolished in *hy5 spl7* seedlings as in *spl7* (Figure 5E). qRT-PCR analysis shows that accumulation of miR408 decreases further in the *hy5 spl7* double mutant compared to either single mutant (Figure 5E). As expected, *LAC13* expression level is higher in *hy5* and *spl7* seedlings and further increases in the *hy5 spl7* double mutant (Figure 5F), which strongly correlates with the decreased level of miR408 in the same mutants. Finally, we checked the activity of *pMIR408:GUS* in the *hy5 spl7* double mutant and found further diminished GUS staining compared to that in both single mutants (Figure 5G). Together our results demonstrate that *SPL7* and *HY5* act additively to control *MIR408* transcription in response to varying growth conditions with *SPL7* playing a more dominant role in determining miR408 level.

The *SPL7-HY5-MIR408* Circuit Controls Gene Expression Dynamics

To investigate the molecular role of *MIR408* in the *SPL7-HY5* network, I began by examining two paralogous miR408 target genes, *LAC12* and *LAC13*, which share essentially identical binding sites for miR408 (Figure 6A). Inspection of the global

occupancy data revealed that *LAC12* but not *LAC13* is targeted by *HY5* while neither is directly regulated by *SPL7*, a pattern also confirmed by ChIP-qPCR analysis (Figure 6B). A diagram summarizing these and aforementioned results is shown as Figure 6C. According to this diagram, one of the molecular functions of the *SPL7-HY5-MIR408* circuit is to differentially regulate *LAC12* and *LAC13*. We therefore monitored expression dynamics of these genes by qRT-PCR analysis in wild type and *hy5* seedlings transitioning from LL to HL.

After switching from LL to HL, *LAC13* transcript level fluctuates modestly in wild type seedlings (Figure 6D). However, in *hy5*, as miR408 level continues to increase, *LAC13* level declines linearly throughout the time course (Figure 6D), indicating that miR408-mediated post-transcriptional regulation becomes predominant in *hy5*. By contrast, *HY5*-based regulation of *LAC12* results in a strong transient induction in wild type and low expression in *hy5* (Figure 6D), suggesting that miR408-mediated repression of *LAC12* is secondary though it has identical miR408 binding site as *LAC13*. Distinct expression patterns of *LAC12* and *LAC13* thus indicate that the *SPL7-HY5-MIR408* loop is capable of differentially regulating paralogous target genes, providing functional support to its role in executing precise gene expression programs.

It should be noted that the above time-course experiment was done with plants grown under continuous light. Yet *HY5*, *SPL7*, and *MIR408* all exhibit pulse-like or rhythmic expression dynamics that peak in a time-dependent manner (Figure 6D), suggesting a possible circadian regulation. Meanwhile, analyzing previous global ChIP studies revealed that *HY5* targets several key genes in the circadian clock, including *CCA1*, *LHY*, *LHCB1*, and *COL1*

(Lee et al., 2007; Zhang et al., 2011). In this work, I also found that these genes (except *LHY*) are targeted by *SPL7* (Supplemental Table 2, sheet 7). In addition, copper homeostasis has been shown to affect circadian rhythm-related plant growth (Andres-Colas et al., 2010). Thus, it awaits future investigations to elucidate the interplay between circadian rhythms and the *HY5-SPL7* circuit.

The *SPL7-HY5-MIR408* Loop Controls Plant Development

The *spl7* and *hy5* mutants display distinct phenotypes during early development. When grown in light, the *hy5* seedling has characteristically long hypocotyl while *spl7* has reduced growth under LC (Figure 7A; Oyama et al., 1997; Yamasaki et al., 2009). Analyzing the *hy5 spl7* double mutant revealed that *SPL7* and *HY5* interact in different ways to regulate different aspects of development. For example, the *hy5 spl7* double mutant displays intermediate phenotypes in terms of hypocotyl length (Figures 7A and 7B) and fresh weight under LC (Figures 7A and 7C); phenotypes similar to the single mutants (e.g. chlorophyll content; Figure 7D); and more severe phenotypes such as lower anthocyanin content than the single mutants (Figure 7E). Consistent with previous report (Yamasaki et al., 2009), we found that the *spl7* defects could be rescued by HC. Interestingly, under this condition, the double mutant still has intermediate hypocotyl length, but no longer displays more severe phenotype regarding anthocyanin accumulation but instead does so for fresh weight. Thus, *SPL7* and *HY5* follow distinct genetic modes for controlling diverse aspects of plant development under different growth conditions, which is consistent with the large set of genes regulated by both transcription factors (Figure 3).

To further examine the role of miR408 in development, we employed a previously developed *amiR408* line in which *MIR408* is silenced by a constitutively expressed artificial miRNA (*amiR408*) that targets the endogenous *MIR408* gene (Zhang and Li, 2013). We found that *amiR408* seedlings display elongated hypocotyls and reduced fresh weight and anthocyanin content (Figure 7). Thus, loss-of-function in *MIR408* impacts all the examined phenotypes, which are unlike either *spl7* or *hy5* but similar to the double mutant, indicating that *MIR408* is a critical component of the *HY5- SPL7* network. Finally, we sought to test whether over-expression of *MIR408* could rescue the developmental defects of *spl7*, *hy5*, or *hy5 spl7*. Consistent with previous finding (Zhang and Li, 2013), we found that the *35S:pre-miR408* transgene partially rescues the reduced growth vigor of *spl7* under LC (Figure 7). Strikingly, introduction of the *35S:pre-miR408* transgene into the *hy5* and *hy5 spl7* backgrounds, which results in accumulation of miR408 and down-regulation of miR408 target genes (Supplemental Figure 9), was found to completely or partially rescue all the examined phenotypes (Figure 7).

Sucrose is known to influence growth and development of *Arabidopsis* seedlings (Gibson, 2005). To test whether growth defects of the *hy5*, *spl7* and *hy5 spl7* mutants and hence the ability of miR408 to rescue such defects is dependent on sucrose, Dr. Zhang supplemented the growth medium with two concentrations of sucrose. In addition to the experiment reported above (Figure 7) in which 1% exogenous sucrose was used, we performed new experiments with only 0.1% sucrose (data not shown). Comparing to Figure 7, our results revealed that low sucrose reduced all growth parameters of seedlings except hypocotyl length. However, even with low sucrose supplementation, we found that silencing *MIR408*

mimics the *hy5 spl7* double mutant and that constitutively activated miR408 completely or partially rescues all the examined phenotypes of *hy5*, *spl7*, as well as *hy5 spl7*. These results thus suggest that action of miR408 in the *HY5-SPL7* network is likely independent of sucrose regimes.

In contrast to drastic phenotypes in seedling development, *hy5* plants have moderate defects in later stages (Ang et al., 1998). In our experiments, the *hy5* mutant displays normal fresh weight but reduced pigmentation in juvenile plants (Supplemental Figure 10). Consequently, the *hy5 spl7* double mutant displays a phenotype on fresh weight similar to *spl7* and pigmentation phenotypes intermediate between *hy5* and *spl7*. Regarding miR408, it is interesting to observe that the *amiR408* line still displays phenotypes most similar to the *hy5 spl7* double mutant. Further, over-expression of miR408 could rescue all examined developmental defects in *hy5*, *spl7*, and *hy5 spl7* (Supplemental Figure 10). These results indicate that the *SPL7-HY5-MIR408* loop is functional beyond the seedling stage.

Cellular Function of miR408

The ability of miR408 to regulate plant growth, chlorophyll level, and repress copper proteins with non-photosynthetic usage prompted us to further investigate its cellular function. Given that PC is a copper-binding protein located in the thylakoid lumen and serves as a photosynthetic electron carrier, we first examined the involvement of miR408 in copper allocation and PC abundance. In *Arabidopsis*, PC is encoded by two paralogous genes *PETE1* and *PETE2* (Weigel et al., 2003). By means of immunoblotting, we analyzed the levels of PETE1 and PETE2 in various lines in which *MIR408* expression is altered. As

shown in Figure 8A, the PC antibody detects both PETE1 and PETE2 in seedlings with PETE2 being the more abundant isoform as previously reported (Abdel-Ghany et al. 2009; Pesaresi et al., 2009). Quantification of total PC (PETE1 and PETE2 combined) revealed that its levels are reduced in mutants with impaired *MIR408* expression (*hy5*, *spl7*, *hy5 spl7*, and *amiR408*) and increased when *MIR408* is constitutively produced (*MIR408-OX/hy5*, *MIR408-OX/spl7*, and *MIR408-OX/hy5 spl7*), compared to wild type (Figure 8B).

As controls, we examined protein levels of other photosynthetic electron carriers (Figures 8A, 8B). We found that the levels of chloroplast cytochrome *b₆* and ferredoxin do not change as *MIR408* expression alters. Thus, miR408 appears to specifically regulate PC in the photosynthetic electron transport chain. Previous studies have shown that mutations of *PETE1* and *PETE2* lead to impaired vegetative growth (Weigel et al. 2003; Joliot and Joliot 2006; Abdel-Ghany, 2009; Pesaresi et al., 2009). In addition, mutations in the transporters that deliver copper to the chloroplast and thylakoid lumen result in drastic reduction of PC and growth defects (Shikanai et al. 2003; Abdel-Ghany et al. 2005). These results suggest that the effects of miR408 levels on plant growth might be exerted through PC abundance.

Because PC is a major copper sink in the chloroplast (Ramshaw et al., 1973), Dr. Zhang examined whether miR408-regulated changes in PC abundance are accompanied with corresponding changes in chloroplastic copper levels. To this end, we measured copper content in whole seedlings of various genotypes as well as chloroplasts isolated from these lines. Compared to wild type, copper in both whole plant and chloroplasts decreases slightly when *MIR408* expression is compromised (*hy5*, *spl7*, *hy5 spl7*, and *amiR408*) but increases when miR408 accumulates to higher levels (Supplemental Figure 11).

Interestingly, we found that copper content in chloroplasts as a percentage of total cellular copper tracks the PC levels and correlates with miR408 abundance in various genetic backgrounds with varied *MIR408* expression (Figure 8C). Therefore, copper allocation to the chloroplast and hence PC abundance relates to miR408 level. Together with other data aforementioned, our results delineate the *SPL7-HY5-MIR408* loop as a cellular mechanism for modulating plant growth based on integration of light and copper signaling (Figure 8D).

Discussion

Regulation of gene expression is fundamental to the integrity and function of all organisms. We now appreciate that complex and sophisticated regulatory networks have evolved in plants for proper gene expression and have the means to map and study these networks. In addition to transcription factors, miRNAs play crucial roles in the regulatory networks by modulating gene expression at the post-transcriptional level (Voinnet, 2009). In the context of regulatory networks, one of the major challenges to understand gene expression is to identify and analyze regulatory gene circuits incorporating both transcriptional and post-transcriptional control mechanisms. Such inquiries should provide much-needed insights into dynamical understanding of gene activity that is critical for plant development and responses to environmental changes.

The SPL7 Regulon in *Arabidopsis*

Copper is an essential mineral micronutrient for plants and participates as a redox catalytic cofactor in a variety of physiological processes (Pilon et al., 2006; Burkhead et al., 2009). While diminished cellular copper impedes photosynthesis, excessive copper leads to the generation of harmful reactive oxygen species and replacement of other metal cofactors (Burkhead et al., 2009). Copper homeostasis is therefore fundamental to the fitness of plants and concerns expression of genes involved in diverse pathways. As a member of the SBP family of zinc finger transcription factors, *SPL7* in *Arabidopsis* (Yamasaki et al., 2009) and its ortholog *CRR1* in *Chlamydomonas* (Kropat et al., 2005) are regulator for copper

homeostasis. It is demonstrated that copper inhibits DNA binding activity of CRR1 and SPL7 and prevents transcription of specific target genes *in vitro* (Sommer et al., 2010). These and other observations have led to the proposal that SPL7 is the copper sensor and copper deficiency promotes its binding to the GTAC motif, which in turn transcriptionally activates the target genes (Yamasaki et al., 2009; Beauclair et al., 2010; Sommer et al., 2010; Bernal et al., 2012).

In this study, I fully elucidated the SPL7 regulon in *Arabidopsis*. By whole genome ChIP-sequencing, I identified 1,535 high confidence SPL7-bound genomic regions aligned with 1,266 gene loci (Figure 2; Supplemental Figure 3; Supplemental Table 2, sheets 1 and 2). A strong enrichment of SPL7 binding in the proximity of the TSS of target genes was observed (Figure 2A), suggesting that distribution of the recognition sites for SPL7 in the genome is highly selective. As a known recognition motif for SPL7 (Kropat et al., 2005; Birkenbihl et al., 2005; Yamashaki et al., 2009; Sommer et al., 2010), the GTAC tetranucleotide, particularly in the context of A/TGTACT/A, is actually over-represented in the SPL7-occupied regions (Figure 2D; Supplemental Figure 4A). Additionally, a novel motif in the SPL7 binding sites that resembles DNA elements recognized by zinc finger transcription factors was found (Supplemental Figures 6B and 6C; Badis et al., 2008). Moreover, the SPL7 binding sites are also enriched with recognition motifs for other transcription factors such as the G-box (Figure 3A; Supplemental Figure 4D). These results clearly demonstrate that SPL7 is well connected in the regulatory network through recognizing different classes of DNA motifs.

By means of RNA-sequencing, transcription of approximately 4,000 genes is found to be

influenced by *SPL7* under both copper deficient and sufficient conditions (Figure 2E; Supplemental Figure 5; Supplemental Table 2, sheets 3-6), including many involved in primary metabolism as have noted (Bernal et al., 2012). Through clustering analysis, the *SPL7*-dependent genes are recognized to form four groups with distinct transcriptional behavior (Figure 2E; Supplemental Figure 5B). Previously, Yamashaki et al. (2009) reported that *SPL7* positively regulates several *MIR* genes while Bernal et al. (2012) noted that *SPL7* represses many metabolism-related genes. Our results indicate that *SPL7* could function either as a positive or negative regulator, which is consistent with analysis of *CRR1* in *Chlamydomonas* (Moseley et al., 2002). Combining the ChIP- and RNA-sequencing data revealed that half of the *SPL7*-bound genes (634 out of 1,266) exhibit *SPL7*-dependent expression (Supplemental Figure 5B). Although this proportion is much higher than the genome average, *SPL7* binding to many target genes may not be sufficient to cause changes in their expression levels. Conversely, a majority of the genes for which proper expression is dependent on *SPL7* are not directly bound by the transcription factor (Supplemental Figure 5B). Together, our results suggest that *SPL7* controls its regulon through interwoven sub-programs coordinated with other transcriptional regulators.

The *HY5-SPL7* Interaction Defines a Light-Copper Crosstalk

By virtue of having a specific set of protein-binding sites in their promoters, most genes are regulated by multiple transcription factors. Combinatorial control is thus a major mechanism underlying transcriptional regulation in eukaryotes (Carey, 1998). An increasing number of studies in plants have documented combinatorial control involving transcription factors that are implicated in multiple biological processes, such as pigment

biosynthesis, carbon metabolism, hormonal responses, organ development, and plant immunity (Mol et al., 1996; Shultz et al., 1998; Hobo et al., 1999; Yanagisawa, 2000; Lara et al., 2003; Shin et al., 2007; Liu and Howell, 2010; Moore et al., 2011; Park et al., 2011). Enrichment of various DNA elements in *SPL7*-occupied regions (Figure 3A; Supplemental Figure 4D) indicates that much of the regulatory function of *SPL7* is likely fulfilled together with other transcription factors.

Given that copper is essential for photosynthesis (Pilon et al., 2006; Burkhead et al., 2009), it is not surprising but nevertheless interesting to find that *SPL7* interacts with *HY5* in *Arabidopsis* (Figures 1A to 1C). Because they each impact thousands of genes, the *SPL7*-*HY5* interaction thus defines a previously unknown transcriptional level light-copper crosstalk. Indeed, through global comparison of the *SPL7* and *HY5* regulons, I show that the *SPL7*-*HY5* feedback loop regulates a large cohort of genes (Figure 3; Supplemental Figures 7 and 8), indicating that the light-copper crosstalk is extensive. Our results further demonstrate that interplay of the *SPL7*-*HY5* network is carried out through different molecular mechanisms. On one hand, *HY5* binds directly to the G-box like motifs in the *SPL7* promoter and represses its expression (Figures 1D and 1E; Supplemental Figure 1). Thus, one aspect of the crosstalk is indirect attenuation of the *SPL7* regulon achieved through *HY5* mediated *SPL7* repression, presumably to put a counterweight on copper responsive genes based on input from light signaling.

On the other hand, interaction between the two transcription factors entails that they exert combinatorial control over their commonly regulated genes. Because *HY5* and *SPL7* both can serve as positive as well as negative regulators, the net impact of these two factors on

the target genes could be cooperative or antagonistic. In supporting of this notion, through RNA-sequencing, I found that *SPL7* and *HY5* commonly regulate 1,090 genes, with roughly half of which being regulated in the same direction by the two transcription factors and half in the opposite direction (Figure 3D; Supplemental Table 2, sheet 8). Further, genes co-regulated by *SPL7* and *HY5* are involved in such processes as photosynthesis (Supplemental Figure 8) and biosynthesis of anthocyanins (Figure 3E), which help to absorb blue-green light and thereby protect photosynthetic tissues from photoinhibition (Feild et al., 2001; Neill and Gould, 2003; Hughes et al., 2005; Solfanelli et al., 2005; Merzlyak et al., 2008). These findings are consistent with observations that chlorophylls and anthocyanins accumulate to high levels under the HC/HL condition (Supplemental Figure 1). Therefore, photosynthesis appears to be a major point of convergence of light signaling through *HY5*, which perceives solar energy available for harvesting, and copper sensing through *SPL7*, which is tied to copper allocation to the electron transport chain and cycling of reactive oxygen species (Burkhead et al., 2009). The sophisticated combinatorial regulations exerted by the *HY5-SPL7* loop thus likely constitute one of the molecular mechanisms for mediating the light-copper crosstalk.

As a specific example for functional study, I show that coordinated *SPL7-HY5* regulation represents a mechanism for calculated miR408 accumulation with *SPL7* playing a more dominant role in determining miR408 levels (Figures 4 and 5). This mechanism ensures that miR408 level is low when sufficient copper is present, intermediate in copper deficient and low light conditions, and high when copper is low but light is strong (Figure 5). Thus, transcriptional regulation of *MIR408* by both *HY5* and *SPL7* allows distinct temporal and

spatial expression dynamics to be established by combining different input signals. Significantly, we demonstrate that *MIR408* is sufficient to activate the *HY5-SPL7* network as constitutively expressed miR408 could rescue or partially rescue all the examined developmental defects of the *hy5*, *spl7*, and *hy5 spl7* mutants (Figures 7 and 8; Supplemental Figure 10).

Complementation of the upstream regulators by *MIR408* is an intriguing finding because all validated miR408 target genes encode copper proteins designated to the extracellular space (Yamasaki et al., 2007; Abdel-Ghany and Pilon, 2008), which suggests a primary role of *MIR408* in copper homeostasis. Our finding indicates that *MIR408* is an integral and critical component of the *HY5-SPL7* network and provides a specific signaling route that connects the transcriptional and post-transcriptional gene regulatory branches. One implication of this finding is that there are multiple parallel routes for regulatory information flow in the *HY5-SPL7* network and that activation of individual routes is sufficient to excite the network. In addition to *MIR408*, *SPL7* and *HY5* co-regulate two more miRNAs and 28 transcription factors. It is plausible that the *HY5-SPL7* feedback outputs calculated gene expression programs through clustered regulatory loops involving other gene regulators. Further elucidating the interplay among genes in the *HY5-SPL7* network should provide much needed insights into gene batteries that underpin plant development in response to changing environments.

A Model for *HY5-SPL7* Regulated *MIR408* Activation in Plant Growth

Based on several lines of evidence, *MIR408* is clearly a critical component of the *HY5-*

SPL7 network. Physiologically, relative copper content in the chloroplast is decreased when *MIR408* expression is undermined as in *hy5*, *spl7*, *hy5 spl7*, and *amiR408* plants while constitutive activation of *MIR408* is sufficient to reverse such decreases (Figure 8C). Molecularly, abundance of PC, which is a key component of the photosynthetic electron transport chain and the major copper sink in chloroplast (Marschner, 2002; Burkhead et al., 2009), is reduced as a consequence of compromised *MIR408* activation but restored by miR408 over production (Figures 8A and 8B). Consequently, cellular contents of chlorophyll (Figure 7D), which is an approximation of the amount of photons harvested (Maxwell and Johnson, 2000), and anthocyanin (Figure 7E; Supplemental Figure 10D), which is modulated by the redox status of the plastoquinone pool of the electron transport chain (Das et al., 2011), are reduced in mutants and *amiR408* plants but elevated by constitutively expressing miR408.

Reconciling all observations, we propose that transcriptional regulation of *MIR408* by HY5 and *SPL7* constitutes a mechanism for integrating light and copper signals to control photosynthesis (Figure 8D). In this model, elevated miR408 level coordinately promoted by HY5 and *SPL7* in the HL/LC condition would reduce copper usage in the extracellular space by repressing transcripts encoding copper proteins such as laccases and plantacyanin (Figure 6A; Yamashaki et al., 2007, Abdel-Ghany and Pilon, 2008). This would result in preferential allocation of cellular copper to the chloroplast and thus PC (Figures 7D, 7E; 8A to 8C; Supplemental Figures 10 and 11). We hypothesize that increased copper delivery to PC drives up its expression, which in turn would increase flux of photosynthetic electron transport as evident by the chlorophyll content. Conversely, compromised *MIR408*

activation would deprive a plant of the ability to preferentially deliver copper to PC in all light and copper regimes. This would result in diminished photosynthetic electron flux due to reduced PC and pigmentation.

Given the critical function of PC in vegetative growth (Weigel et al. 2003; Joliot and Joliot 2006; Abdel-Ghany, 2009; Pesaresi et al., 2009), an inference is that miR408 levels should correlate with vigor of plant growth. This is indeed the case as we have previously shown that constitutive activation of *MIR408* results in enhanced vegetative growth while silencing *MIR408* causes impaired growth (Zhang and Li, 2013). Further, we found that the *hy5*, *spl7*, and *hy5 spl7* mutants display defects in vegetative development that could be rescued by constitutively activated *MIR408* (Figure 7). Given the inherent complexity of copper homeostasis (Figure 8D), it is not yet clear how miR408-based regulation works coordinately with the web of copper chaperones and transporters for economic distribution of this critical transition metal under varying growth conditions. For example, *MIR408-OX* apparently could increase total copper content in *spl7* plants (Supplemental Figure 11A). This implies that miR408 may enhance expression or activity of the copper transporters such as COPT1 or COPT2 in the absence of a functional SPL7, though the signaling mechanism is elusive. Nevertheless, our results collectively indicate that cellular function of miR408 is to promote copper allocation to the chloroplast and abundance of PC, thereby constitutes one specific regulatory route downstream of *SPL7* and *HY5* to modulate vegetative growth based on light and copper inputs (Figure 8D).

It should be noted that *MIR408* is among the most conserved miRNA families in land plants (Axtell and Bowman, 2008), suggesting that its role in mediating light-copper

crosstalk is fundamental to plants. Thus, the *SPL7-HY5-MIR408* circuit represents a potentially conserved determinant for photosynthetic activity and hence plant growth and adaptation in changing environment. Photosynthesis in plants is a relatively inefficient process, with far below 10% of the received solar energy being converted to chemical energy (Zhu et al., 2010). This relatively low efficiency has provided an impetus for researchers to genetically modify plants to achieve greater efficiencies and enhanced growth. So far, much of this effort has centered on carbon fixation (Kirschbaum, 2011). Our finding suggests that manipulating capacity of the electron transport chain mediated by the *SPL7-HY5-MIR408* circuit is an alternative strategy that warrants further investigation.

Methods

Plant Materials and Growth Conditions

Wild-type plant used was *Arabidopsis thaliana* ecotype Col-0. Mutant defective in *HY5* and *SPL7* was respectively *hy5-215* (Oyama et al, 1997) and the T-DNA insertion line SALK_093849 (Yamasaki et al., 2009). To obtain the *35S:FLAG-SPL7/spl7* line, N-terminal FLAG-tagged *SPL7* was generated by PCR using forward primer containing the *FLAG* sequence, inserted into the pJim19 binary vector under control of the 35S promoter, and transformed into the *spl7* background. Transgenic plants were selected with 20mg/L Basta. T₃ generation homozygous lines were used for all experiments. Transgenic plants expressing the GUS reporter were generated as previously described (Zhang and Li, 2013). Briefly, a mutated miR408 promoter was obtained from the native promoter sequence by bridge PCR to change the CACGTG sequence of the G-box to CTGCAG, and cloned into the pCAMBIA-1381Xa vector (CAMBIA) and transformed into various genetic backgrounds as indicated. For each transgene, at least three independent lines were selected with 25 mg/L Hygromycin. T₂ generation plants of representative lines were used for histochemical staining of GUS activity as previously described (Zhang and Li, 2013). The *hy5 spl7* double mutant was generated by crossing *hy5-215* and *spl7*. F₂ progenies homozygous for both alleles were identified by PCR analysis of genomic DNA for the presence of T-DNA and the *hy5* allele. This procedure was repeated for the F₃ generation in which a single *hy5 spl7* double mutant line was selected and used for all subsequent analyses. The *35S:pre-miR408* transgene, which contains the sequence encompassing the

pre-miR408 stem-loop structure (Zhang and Li, 2013), was used to overexpress miR408 in various genetic backgrounds. Transgenic plants were selected with Basta and T₂ generation homozygous lines were used for subsequent experiments.

To grow *Arabidopsis* seedlings, seeds were surface sterilized, plated on agar-solidified MS media including 0.1% or 1% sucrose and the indicated concentrations of CuSO₄. The plates were incubated at 4 °C for four days in the dark, and then transferred to continuous white light with intensity of either 40 or 170 $\mu\text{molm}^{-2}\text{s}^{-1}$ and allowed to grow at 22 °C for seven days or other indicated length of time. To obtain adult plants, seedlings were transferred to soil and maintained in a growth chamber with the following setting: standard long-day (16h light/8h darkness) condition, light intensity of approximately 120 $\mu\text{molm}^{-2}\text{s}^{-1}$, 50% relative humidity, and temperature at 22 °C. All molecular and physiological experiments were conducted with at least three independently harvested biological samples.

ChIP-Sequencing

Chromatin isolation was performed with whole seedlings of *35S:FLAG-SPL7/spl7* and *spl7* grown under the LC/HL condition according to the procedure described by Bowler et al. (2004). The resuspended chromatin pellet was sonicated at 4 °C with a Diagenode Bioruptor set at high intensity for 10 min (30s on, 30s off intervals). Chromatin was immunoprecipitated with monoclonal an anti-FLAG antibody (Sigma-Aldrich) according to the Affymetrix Chromatin Immunoprecipitation Assay Protocol Rev.3. The precipitated DNA (one biological replicate for each sample) was sequenced using the HiSeq2000 system (Illumina) according to the manufacturer's instructions. Sequencing reads of 100-bp

were mapped to the TAIR10 genome release of *Arabidopsis* using Bowtie (Langmead et al., 2009), allowing two mismatches and no gaps. Only uniquely mapped reads were retained for further analysis. MACS (Zhang et al., 2008) with customized parameters (bandwidth = 300 bp; P value = $1e^{-05}$; mfold = 10-50; nolambda) was used to call peaks representing enriched SPL7 binding specifically in *35S:FLAG-SPL7/spl7* but not *spl7*.

We used Multiple Em for Motif Elicitation (Bailey et al., 2006) to identify sequence motifs over-represented in SPL7 binding sites. To search for the presence other transcription factor binding sites, a position weight matrix method based on experimentally validated data derived from the *Arabidopsis thaliana* Promoter Binding Element Database (<http://exon.cshl.org/cgi-bin/atprobe/atprobe.pl>) and the Arabidopsis Gene Regulatory Information Server dataset (Davuluri et al., 2003) was used as previously described (Zhao et al., 2013). Posterior probability was calculated using 10,000 times Monte Carlo simulation in Matlab.

RNA-Sequencing

Total RNA from wild type and *spl7* seedlings grown under LC and HC conditions was isolated using the RNeasy Plus Mini Kit (Qiagen). Library construction and sequencing on the HiSeq2000 platform were performed according to the manufacturer's instructions (Illumina). One biological replicate for each sample was analyzed. The resultant 100 bp reads were aligned to the TAIR10 genome using TopHat (Trapnell et al., 2009), allowing two mismatches and maximal intron size of 2 Kb. Only uniquely mapped reads were used for subsequent analysis. Differentially expressed genes were identified using cufflinks

(Trapnell et al., 2010) with the following parameters: minimal number of alignment = 50, quartile normalization, as False Discovery Rate < 0.01, and P < 0.05. Results from four pairwise comparisons (*spl7*/HC vs WT/HC; WT/LC vs WT/HC; *spl7*/HC vs *spl7*/LC; and WT/LC vs *spl7*/LC) were included for hierarchical clustering analysis based on Pearson's correlation. For a given gene, value for each comparison was set to the logarithm of fold change of normalized read counts (per Kb transcript per million mapped reads). Heatmap of the differentially expressed genes (rows) across the four comparisons (columns) was generated using the row Z-score. For each row in a column, Z-score was calculated by subtracting that gene's mean relative expression level across the four comparisons from its value in that particular comparison and then dividing by the standard deviation across all the comparisons. GO analysis was performed using BiNGO (Maere et al., 2005). ChIP- and RNA-sequencing data are available in the Gene Expression Omnibus database under accession number GSE45213.

Yeast Assays

For yeast two-hybrid assays, full-length *SPL7* open reading frame was amplified by RT-PCR from wild type plants and cloned into the B42AD vector (Clontech) to generate the B42AD-SPL7 construct. The LexA-HY5 and B42AD-COP1 constructs were obtained as described previously (Ang et al., 1998). The respective combinations of B42AD and LexA fusion constructs were cotransformed into the yeast strain EGY48 containing the reporter plasmid *p8op:LacZ* (Clontech). For yeast one-hybrid assay, plasmids for AD fusions (AD-SPL7 and AD-HY5) were co-transformed with the *LacZ* reporter driven by the *MIR408* promoter as described (Zhang and Li, 2013) into the yeast strain EGY48. Transformants

were grown on proper dropout plates containing X-gal for blue color development. Yeast transformation and liquid assay were conducted as described in the Yeast Protocols Handbook (Clontech).

Protein Analyses

Protein extraction and Western blotting were carried out as previously described (Feng et al., 2004). The blots were probed with different primary antibodies as follows: anti-FLAG (GenScript), anti-HY5 (Osterlund et al., 2000), anti-RPT5 (Kwok et al., 1999), anti-His (Qiagen), anti-plastocyanin, anti-Cytochrome *b₆*, and anti-ferredoxin (Acris Antibodies). For co-IP experiment on the in vivo binding between SPL7 and HY5, total protein extracts were prepared from *35S:FLAG-SPL7/spl7* transgenic seedlings. The anti-FLAG or anti-HY5 antibody was added to the protein extracts and precipitated with Protein A agarose beads (Sigma) following a method previously described (Feng et al., 2004). The precipitates and total extracts were then subjected to immunoblot analysis with antibodies against HY5, FLAG, or RPT5. For in vitro binding, two μg of purified recombinant bait proteins (GST-HY5 and GST) and two μg of prey protein (6xHis-SPL7) were added to 1 mL of binding buffer containing 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, and 0.6% Triton X-100. After incubation at 4 °C for 2h, Glutathione Sepharose 4B beads (Amersham Biosciences) were then added and incubated for one further hour. After washing three times with the binding buffer, pulled-down proteins were eluted in 2 \times SDS loading buffer at 95 °C for 10 min, separated on SDS-PAGE gels, and detected by immunoblotting using the anti-His antibody.

RNA Analyses

Total RNA was extracted using the TRIzol reagent (Invitrogen) as suggested by the manufacturer to include the low-molecular-weight fraction of RNA. For qRT-PCR quantification of protein-coding genes and pri-miR408, DNaseI treated RNA was reverse-transcribed using the SuperScript II reverse transcriptase (Invitrogen). The resultant cDNA was analyzed using the SYBR Green master mix with the ABI 7500 Fast Real-Time PCR System (Applied Biosystems) in triplicate. The *Actin7* amplicon was used for normalization. For qRT-PCR quantification of mature miR408, poly (A) tailing and first strand cDNA synthesis were carried out using the NCode miRNA First-Strand cDNA Kit (Invitrogen). A miR408-specific forward primer (complementary to mature miR408) and a universal reverse primer supplied by the manufacturer were used. 5S ribosome RNA was used for normalization. Determination of relative gene expression level was carried out using the standard $2^{-\Delta\Delta C(T)}$ method. All experiments were performed on three independent biological samples with each including three technical replicates. Northern blot analysis of miRNA was performed as previously described (Zhang and Li, 2013). Primer and probe sequences are listed in Supplemental Table 3.

EMSA

Full length *HY5* was amplified by RT-PCR and cloned into the vector pET-28a (+) (Novagen). The resulting plasmid was introduced into *E. coli* strain BL-21 and His-tagged HY5 purified with the Ni-NAT Agarose system (Qiagen). Recombinant SPL7 was prepared as described previously (Zhang and Li, 2013). EMSA was performed using digoxigenin-

labeled probes and the second generation DIG Gel Shift Kit (Roche) according to the manufacture's instruction. Sequences of probes and primers used in this study are shown in Supplemental Table 3.

Chlorophyll, Anthocyanin, and Glucose Measurements

Measurement of chlorophyll and anthocyanin was performed as described previously (Chory et al., 1989). Briefly, seven-day-old seedlings were harvested, weighed, and homogenized in liquid nitrogen. Chlorophyll *a/b* was extracted into 80% acetone and quantified as microgram per gram fresh weight using MacKinney's specific absorption coefficients for which chlorophyll *a* = $12.7(A_{663}) - 2.69(A_{645})$ and chlorophyll *b* = $22.9(A_{645}) - 4.48(A_{663})$. For anthocyanin, homogenized samples were incubated overnight in 0.3 ml of 1% HCl in methanol at 4 °C and extracted using an equal volume of chloroform after addition of 0.2 ml of water. The quantity of anthocyanin was determined by spectrophotometric measurement of the aqueous phase ($A_{530} - 0.25 \times A_{657}$) and normalized to fresh weight of each sample. Measurement of glucose content was performed using the Glucose and Sucrose Assay Kit (Biovision) according to the manufacturer's instructions. Seedlings were homogenized in a glucose assay buffer and the supernatant was collected after centrifuging at 12,000 rpm for 10 min. The reaction system was set up in 100 µl total volume including 50 µl of sample in glucose assay buffer and 50 µl of glucose assay mix (46 µl glucose assay buffer, 2 µl glucose probe, and 2 µl glucose enzyme mix), and incubated at 37 °C for 30 min. The absorbance at 570 nm was collected and glucose concentrations of the test samples calculated based on the standard curve.

Measurement of Cellular Copper

Harvested ten-day-old seedlings of various genotypes were weighted, and then washed twice with 1 mM EDTA and once with double-distilled water. The materials were then desiccated, digested in 1% nitric acid, and used directly for copper analysis with inductively coupled plasma-atomic emission spectroscopy as described previously (Cohu and Pilon, 2007). Chloroplasts were isolated from shoots of seedlings as described previously (Kubis et al., 2008). Briefly, plant tissues were homogenized in the chloroplast isolation buffer (0.3 M sorbitol, 5 mM MgCl₂, 5 mM EGTA, 5 mM EDTA, 20 mM HEPES/KOH pH 8.0, 10 mM NaHCO₃). The homogenate was filtered through two layers of Miracloth and centrifuged at 3,000 rpm in the Sorvall RC6 centrifuge with an SLA-1500 rotor for 5 min to pellet the crude chloroplast. To separate the intact chloroplasts from broken chloroplasts and other debris, the pellet was further purified on a continuous 50% (v/v) Percoll gradient through centrifugation at 7,000 rpm in the Sorvall RC6 centrifuge with an HB-6 rotor for 10 min. The lower green band in the gradient, which contains intact chloroplasts, was collected. Intactness of the chloroplasts was assessed by phase-contrast light microscopy and only samples with more than 80% intact chloroplasts were retained. Isolated chloroplasts were then dried, digested in 1% nitric acid, and assayed for copper content.

Accession Number

Sequence data from this article can be found in the *Arabidopsis* Genome Initiative or GenBank/EMBL databases under the following accession numbers: *MIR408* (At2g47015),

SPL7 (At5g18830), *HY5* (At5g11260), *LAC12* (At5g05390), *LAC13* (At5g07130), *PETE1* (At1g76100), *PETE2* (At1g20340), *FD1* (At1g10960), *PETB* (ATCG00720), *ACTIN7* (At5g09810). T-DNA insertion mutant used is *spl7* (SALK_093849).

References

Abdel-Ghany, S.E. (2009). Contribution of plastocyanin isoforms to photosynthesis and copper homeostasis in *Arabidopsis thaliana* grown at different copper regimes. *Planta* 229, 767-779.

Abdel-Ghany, S.E., and Pilon, M. (2008). MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in *Arabidopsis*. *J. Biol. Chem.* 283, 15932-15945.

Abdel-Ghany, S.E., Muller-Moule, P., Niyogi, K.K., Pilon, M., and Shikanai, T. (2005). Two P-type ATPases are required for copper delivery in *Arabidopsis thaliana* chloroplasts. *Plant Cell* 17, 1233-1251.

Andres-Colas, N., Perea-Garcia, A., Puig, S., and Penarrubia, L. (2010). Deregulated copper transport affects *Arabidopsis* development especially in the absence of environmental cycles. *Plant Physiol.* 153, 170-184.

Ang, L.H., Chattopadhyay, S., Wei, N., Oyama, T., Okada, K., Batschauer, A., and Deng, X.W. (1998). Molecular interaction between COP1 and HY5 defines a regulatory switch for light control of *Arabidopsis* development. *Mol. Cell* 1, 213-222.

Axtell, M.J., and Bowman, J.L. (2008). Evolution of plant microRNAs and their targets. *Trends Plant Sci.* 13, 343-349.

Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D.,

Gossett, A.J., Hasinoff, M.J., Warren, C.L., Gebbia, M., Talukder, S., Yang, A., Mnaimneh, S., Terterov, D., Coburn, D., Li Yeo, A., Yeo, Z.X., Clarke, N.D., Lieb, J.D., Ansari, A.Z., Nislow, C., and Hughes, T.R. (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* 32, 878-887.

Bailey, T.L., Williams, N., Mischak, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369-373.

Beauchair, L., Yu, A., and Bouche, N. (2010). microRNA-directed cleavage and translational repression of the copper chaperone for superoxide dismutase mRNA in *Arabidopsis*. *Plant J.* 62, 454-462.

Bernal, M., Casero, D., Singh, V., Wilson, G.T., Grande, A., Yang, H., Dodani, S.C., Pellegrini, M., Huijser, P., Connolly, E.L., Merchant, S.S., and Kramer, U. (2012). Transcriptome sequencing identifies SPL7-regulated copper acquisition genes FRO4/FRO5 and the copper dependence of iron homeostasis in *Arabidopsis*. *Plant Cell* 24, 738-761.

Birkenbihl, R.P., Jach, G., Saedler, H., and Huijser, P. (2005). Functional dissection of the plant-specific SBP-domain: overlap of the DNA-binding and nuclear localization domains. *J. Mol. Biol.* 352, 585-596.

Bowler, C., Benvenuto, G., Laflamme, P., Molino, D., Probst, A.V., Tariq, M., and Paszkowski, J. (2004). Chromatin techniques for plant cells. *Plant J.* 39, 776-789.

Burkhead, J.L., Reynolds, K.A., Abdel-Ghany, S.E., Cohu, C.M., and Pilon, M. (2009).

Copper homeostasis. *New Phytol.* 182, 799-816.

Carey, M. (1998). The enhanceosome and transcriptional synergy. *Cell* 92, 5-8.

Castruita, M., Casero, D., Karpowicz, S.J., Kropat, J., Vieler, A., Hsieh, S.I., Yan, W., Cokus, S., Loo, J.A., Benning, C., Pellegrini, M., and Merchant, S.S. (2011). Systems biology approach in *Chlamydomonas* reveals connections between copper nutrition and multiple metabolic steps. *Plant Cell* 23, 1273-1292.

Chattopadhyay, S., Ang, L.H., Puente, P., Deng, X.W., and Wei, N. (1998). Arabidopsis bZIP protein HY5 directly interacts with light-responsive promoters in mediating light control of gene expression. *Plant Cell* 10, 673-683.

Chen, L. (1999). Combinatorial gene regulation by eukaryotic transcription factors. *Curr. Opin. Struct. Biol.* 9, 48-55.

Chen, M., Chory, J., and Fankhauser, C. (2004). Light signal transduction in higher plants. *Annu. Rev. Genet.* 38, 87-117.

Chory, J., Peto, C., Feinbaum, R., Pratt, L., and Ausubel, F. (1989). Arabidopsis thaliana mutant that develops as a light-grown plant in the absence of light. *Cell* 58, 991-999.

Cluis, C.P., Mouchel, C.F., and Hardtke, C.S. (2004). The Arabidopsis transcription factor HY5 integrates light and hormone signaling pathways. *Plant J.* 38, 332-347.

Cohu, C.M., Pilon, M. (2007). Regulation of superoxide dismutase expression by copper availability. *Physiologia Plantarum* 129, 747-755.

Das, P.K., Geul, B., Choi, S.B., Yoo, S.D., and Park, Y.I. (2011). Photosynthesis-dependent anthocyanin pigmentation in Arabidopsis. *Plant Signal Behav.* 6, 23-25.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4, 25.

Feild, T.S., Lee, D.W., and Holbrook, N.M. (2001). Why leaves turn red in autumn. The role of anthocyanins in senescing leaves of red-osier dogwood. *Plant Physiol.* 127, 566-574.

Feng, S., Shen, Y., Sullivan, J.A., Rubio, V., Xiong, Y., Sun, T.P., and Deng, X.W. (2004). Arabidopsis CAND1, an unmodified CUL1-interacting protein, is involved in multiple developmental pathways controlled by ubiquitin/proteasome-mediated protein Degradation. *Plant Cell* 16, 1870-1882.

Gao, Z., Zhao, R., and Ruan, J. (2013). A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. *BMC Genomics* 14 Suppl 1, S4.

Garcia-Molina, A., Andres-Colas, N., Perea-Garcia, A., Neumann, U., Dodani, S.C., Huijser, P., Penarrubia, L., and Puig, S. (2013). The Arabidopsis COPT6 transport protein functions in copper distribution under copper-deficient conditions. *Plant Cell Physiol.* 54, 1378-1390.

Gibson, S.I. (2005). Control of plant development and gene expression by sugar signaling.

Curr. Opin. Plant Biol. 8:93–102.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36, D154-158.

Grotjohann, I., and Fromme, P. (2005). Structure of cyanobacterial photosystem I. *Photosynth. Res.* 85, 51-72.

Hobo, T., Kowyama, Y., and Hattori, T. (1999). A bZIP factor, TRAB1, interacts with VP1 and mediates abscisic acid-induced transcription. *Proc. Natl. Acad. Sci. USA* 96, 15348-15353.

Hughes, N.M., Neufeld, H.S., and Burkey, K.O. (2005). Functional role of anthocyanins in high-light winter leaves of the evergreen herb *Galax urceolata*. *New Phytol.* 168, 575-587.

Jiao, Y., Lau, O.S., and Deng, X.W. (2007). Light-regulated transcriptional networks in higher plants. *Nat. Rev. Genet.* 8, 217-230.

Joliot, P., and Joliot, A. (2006). Cyclic electron flow in C3 plants. *Biochim. Biophys. Acta.* 1757, 362-368.

Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* 14, 787-799.

Kirschbaum, M.U. (2011). Does enhanced photosynthesis enhance growth? Lessons learned from CO₂ enrichment studies. *Plant Physiol.* 155, 117-124.

Kropat, J., Tottey, S., Birkenbihl, R.P., Depege, N., Huijser, P., and Merchant, S. (2005). A regulator of nutritional copper signaling in *Chlamydomonas* is an SBP domain protein that recognizes the GTAC core of copper response element. *Proc. Natl. Acad. Sci. USA* 102, 18730-18735.

Kubis, S.E., Lilley, K.S., and Jarvis, P. (2008). Isolation and preparation of chloroplasts from *Arabidopsis thaliana* plants. *Methods Mol. Biol.* 425, 171-186.

Kwok, S.F., Staub, J.M., and Deng, X.W. (1999). Characterization of two subunits of *Arabidopsis* 19S proteasome regulatory complex and its possible interaction with the COP9 complex. *J. Mol. Biol.* 285, 85-95.

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

Lara, P., Onate-Sanchez, L., Abraham, Z., Ferrandiz, C., Diaz, I., Carbonero, P., and Vicente-Carbajosa, J. (2003). Synergistic activation of seed storage protein gene expression in *Arabidopsis* by ABI3 and two bZIPs related to OPAQUE2. *J. Biol. Chem.* 278, 21003-21011.

Lee, J., He, K., Stolc, V., Lee, H., Figueroa, P., Gao, Y., Tongprasit, W., Zhao, H., Lee, I., and Deng, X.W. (2007). Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. *Plant Cell* 19, 731-749.

Li, F., Han, Y., Feng, Y., Xing, S., Zhao, M., Chen, Y., and Wang, W. (2013). Expression of wheat expansin driven by the RD29 promoter in tobacco confers water-stress tolerance

without impacting growth and development. *J. Biotechnol.* 163, 281-291.

Liu, J.X., and Howell, S.H. (2010). bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in *Arabidopsis*. *Plant Cell* 22, 782-796.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448-3449.

Marschner, H. (2002). *Mineral Nutrition in Higher Plants*. (London: Academic Press).

Maxwell, K., and Johnson, G.N. (2000). Chlorophyll fluorescence--a practical guide. *J. Exp. Bot.* 51, 659-668.

Merzlyak, M.N., Chivkunova, O.B., Solovchenko, A.E., and Naqvi, K.R. (2008). Light absorption by anthocyanins in juvenile, stressed, and senescing leaves. *J. Exp. Bot.* 59, 3903-3911.

Mol J, J.G., Schafer E, Weiss D (1996). Signal perception, transduction, and gene expression involved in anthocyanin biosynthesis. *Crit. Rev. Plant Sci.* 15, 525-557.

Moore, D.L., and Goldberg, J.L. (2011). Multiple transcription factor families regulate axon growth and regeneration. *Dev. Neurobiol.* 71, 1186-1211.

Moseley, J.L., Page, M.D., Alder, N.P., Eriksson, M., Quinn, J., Soto, F., Theg, S.M., Hippler, M., and Merchant, S. (2002). Reciprocal expression of two candidate di-iron

enzymes affecting photosystem I and light-harvesting complex accumulation. *Plant Cell* 14, 673-688.

Nagae, M., Nakata, M., and Takahashi, Y. (2008). Identification of negative cis-acting elements in response to copper in the chloroplastic iron superoxide dismutase gene of the moss *Barbula unguiculata*. *Plant Physiol.* 146, 1687-1696.

Neill, S.O., Gould, K.S. (2003). Anthocyanins in leaves: light attenuators or antioxidants? *Functional Plant Biology* 30, 865-873.

Osterlund, M.T., Hardtke, C.S., Wei, N., and Deng, X.W. (2000). Targeted destabilization of HY5 during light-regulated development of *Arabidopsis*. *Nature* 405, 462-466.

Oyama, T., Shimura, Y., and Okada, K. (1997). The *Arabidopsis* HY5 gene encodes a bZIP protein that regulates stimulus-induced development of root and hypocotyl. *Genes Dev.* 11, 2983-2995.

Park, J., Lee, N., Kim, W., Lim, S., and Choi, G. (2011). ABI3 and PIL5 collaboratively activate the expression of SOMNUS by directly binding to its promoter in imbibed *Arabidopsis* seeds. *Plant Cell* 23, 1404-1415.

Pesaresi, P., Scharfenberg, M., Weigel, M., Granlund, I., Schroder, W.P., Finazzi, G., Rappaport, F., Masiero, S., Furini, A., Jahns, P., and Leister, D. (2009). Mutants, overexpressors, and interactors of *Arabidopsis* plastocyanin isoforms: revised roles of plastocyanin in photosynthetic electron flow and thylakoid redox state. *Mol. Plant* 2, 236-248.

Pilon, M., Abdel-Ghany, S.E., CoHu, C.M., Gogolin, K.A., and Ye, H. (2006). Copper cofactor delivery in plant cells. *Curr. Opin. Plant Biol.* 9, 256-263.

Quinn, J.M., and Merchant, S. (1995). Two copper-responsive elements associated with the *Chlamydomonas* C_{yc6} gene function as targets for transcriptional activators. *Plant Cell* 7, 623-628.

Quinn, J.M., Nakamoto, S.S., and Merchant, S. (1999). Induction of coproporphyrinogen oxidase in *Chlamydomonas* chloroplasts occurs via transcriptional regulation of Cpx1 mediated by copper response elements and increased translation from a copper deficiency-specific form of the transcript. *J. Biol. Chem.* 274, 14444-14454.

Ramshaw, J.A., Brown, R.H., Scawen, M.D., and Boulter, D. (1973). Higher plant plastocyanin. *Biochim. Biophys. Acta* 303, 269-273.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C.O., Forrest, A.R., Gough, J., Grimmond, S., Han, J.H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C.R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R.D., Tegner, J., Lenhard, B., Teichmann, S.A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D.A., Ideker, T., and Hayashizaki, Y. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744-752.

Remenyi, A., Scholer, H.R., and Wilmanns, M. (2004). Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* 11, 812-815.

Schultz, T.F., Medina, J., Hill, A., and Quatrano, R.S. (1998). 14-3-3 proteins are part of an abscisic acid-VIVIPAROUS1 (VP1) response complex in the Em promoter and interact with VP1 and EmBP1. *Plant Cell* 10, 837-847.

Shi, L.X., and Schroder, W.P. (2004). The low molecular mass subunits of the photosynthetic supracomplex, photosystem II. *Biochim. Biophys. Acta* 1608, 75-96.

Shikanai, T., Muller-Moule, P., Munekage, Y., Niyogi, K.K., and Pilon, M. (2003). PAA1, a P-type ATPase of Arabidopsis, functions in copper transport in chloroplasts. *Plant Cell* 15, 1333-1346.

Shin, J., Park, E., and Choi, G. (2007). PIF3 regulates anthocyanin biosynthesis in an HY5-dependent manner with both factors directly binding anthocyanin biosynthetic gene promoters in Arabidopsis. *Plant J.* 49, 981-994.

Singh, K.B. (1998). Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol.* 118, 1111-1120.

Solfanelli, C., Poggi, A., Loreti, E., Alpi, A., and Perata, P. (2006). Sucrose-specific induction of the anthocyanin biosynthetic pathway in Arabidopsis. *Plant Physiol.* 140, 637-646.

Sommer, F., Kropat, J., Malasarn, D., Grosseohme, N.E., Chen, X., Giedroc, D.P., and

Merchant, S.S. (2010). The CRR1 nutritional copper sensor in *Chlamydomonas* contains two distinct metal-responsive domains. *Plant Cell* 22, 4098-4113.

Song, Y.H., Yoo, C.M., Hong, A.P., Kim, S.H., Jeong, H.J., Shin, S.Y., Kim, H.J., Yun, D.J., Lim, C.O., Bahk, J.D., Lee, S.Y., Nagao, R.T., Key, J.L., and Hong, J.C. (2008). DNA-binding study identifies C-box and hybrid C/G-box or C/A-box motifs as high-affinity binding sites for STF1 and LONG HYPOCOTYL5 proteins. *Plant Physiol.* 146, 1862-1877.

Sunkar, R., Kapoor, A., and Zhu, J.K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in *Arabidopsis* is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18, 2051-2065.

Tomancak, P., and Ohler, U. (2010). Mapping the complexity of transcription control in higher eukaryotes. *Genome Biol.* 11, 115.

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

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

Vandenbussche, F., Habricot, Y., Condiff, A.S., Maldiney, R., Van der Straeten, D., and Ahmad, M. (2007). HY5 is a point of convergence between cryptochrome and cytokinin

signalling pathways in *Arabidopsis thaliana*. *Plant J.* 49, 428-441.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669-687.

Weigel, M., Varotto, C., Pesaresi, P., Finazzi, G., Rappaport, F., Salamini, F., and Leister, D. (2003). Plastocyanin is indispensable for photosynthetic electron flow in *Arabidopsis thaliana*. *J. Biol. Chem.* 278, 31286-31289.

Yadav, V., Kundu, S., Chattopadhyay, D., Negi, P., Wei, N., Deng, X.W., and Chattopadhyay, S. (2002). Light regulated modulation of Z-box containing promoters by photoreceptors and downstream regulatory components, COP1 and HY5, in *Arabidopsis*. *Plant J.* 31, 741-753.

Yamasaki, H., Hayashi, M., Fukazawa, M., Kobayashi, Y., and Shikanai, T. (2009). SQUAMOSA Promoter Binding Protein-Like7 Is a Central Regulator for Copper Homeostasis in *Arabidopsis*. *Plant Cell* 21, 347-361.

Yamasaki, H., Abdel-Ghany, S.E., Cohu, C.M., Kobayashi, Y., Shikanai, T., and Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in *Arabidopsis*. *J. Biol. Chem.* 282, 16369-16378.

Yanagisawa, S. (2000). Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. *Plant J.* 21, 281-288.

Zhang, H., and Li, L. (2013). SQUAMOSA promoter binding protein-like7 regulated microRNA408 is required for vegetative development in *Arabidopsis*. *Plant J.* 74, 98-109.

Zhang, H., He, H., Wang, X., Wang, X., Yang, X., Li, L., and Deng, X.W. (2011). Genome-wide mapping of the HY5-mediated gene networks in Arabidopsis that involve both transcriptional and post-transcriptional regulation. *Plant J.* 65, 346-358.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.

Zhao, X., Zhang, H., and Li, L. (2013). Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis. *Genomics* 101, 187-194.

Zhu, X.G., Long, S.P., and Ort, D.R. (2010). Improving photosynthetic efficiency for greater yield. *Annu. Rev. Plant. Biol.* 61, 235-261.

Figures

Figure 1. Interaction of SPL7 and HY5.

- (A) HY5 interacts with SPL7 in a yeast two-hybrid assay. The β -galactosidase activities resulted from the HY5-SPL7 interaction and various controls are shown. The HY5-COP1 interaction (Ang et al., 1998) was used as a positive control. Error bars indicate SD (n = 4).
- (B) HY5 can pull down SPL7 in vitro. Purified 6 \times His-tagged SPL7 was incubated with recombinant GSTHY5 or GST and immunoprecipitated with agarose beads conjugated with a GST antibody. The precipitates were subject to Western blotting using the anti-His antibody. Input, 5% of the purified SPL7 used in the pull-down assays.
- (C) HY5 associates with SPL7 *in vivo*. Total protein extracts from 35S:FLAG-SPL7/*spl7* seedlings were incubated with anti-FLAG or HY5 antibody-conjugated agarose beads. The precipitates and total extracts were subject to immunoblotting with antibodies against HY5 and FLAG, respectively. RPT5 was used as a control.
- (D) Confirmation of HY5 binding to the *SPL7* promoter by ChIP-qPCR. ChIP was performed in wild type and *hy5* seedlings with or without the anti-HY5 antibody. The resultant DNA was analyzed by qPCR with the values normalized to their respective DNA inputs.
- (E) and (F) qRT-PCR analyses of *SPL7* and *HY5* transcripts levels. Reverse-transcribed cDNA from wild type, *hy5*, or *spl7* seedlings was examined by qPCR with the values normalized to those of the wild type.

(G) Immunoblot analysis of HY5 protein levels in wild type and *spl7* seedlings. Values below the blots represent HY5 levels normalized against the loading control RPT5 using Image J software and set to one for wild type. Data for ChIP-qPCR or qRT-PCR are means \pm SD from three biological replicates.

(This figure was generated by Dr. Huiyong Zhang)

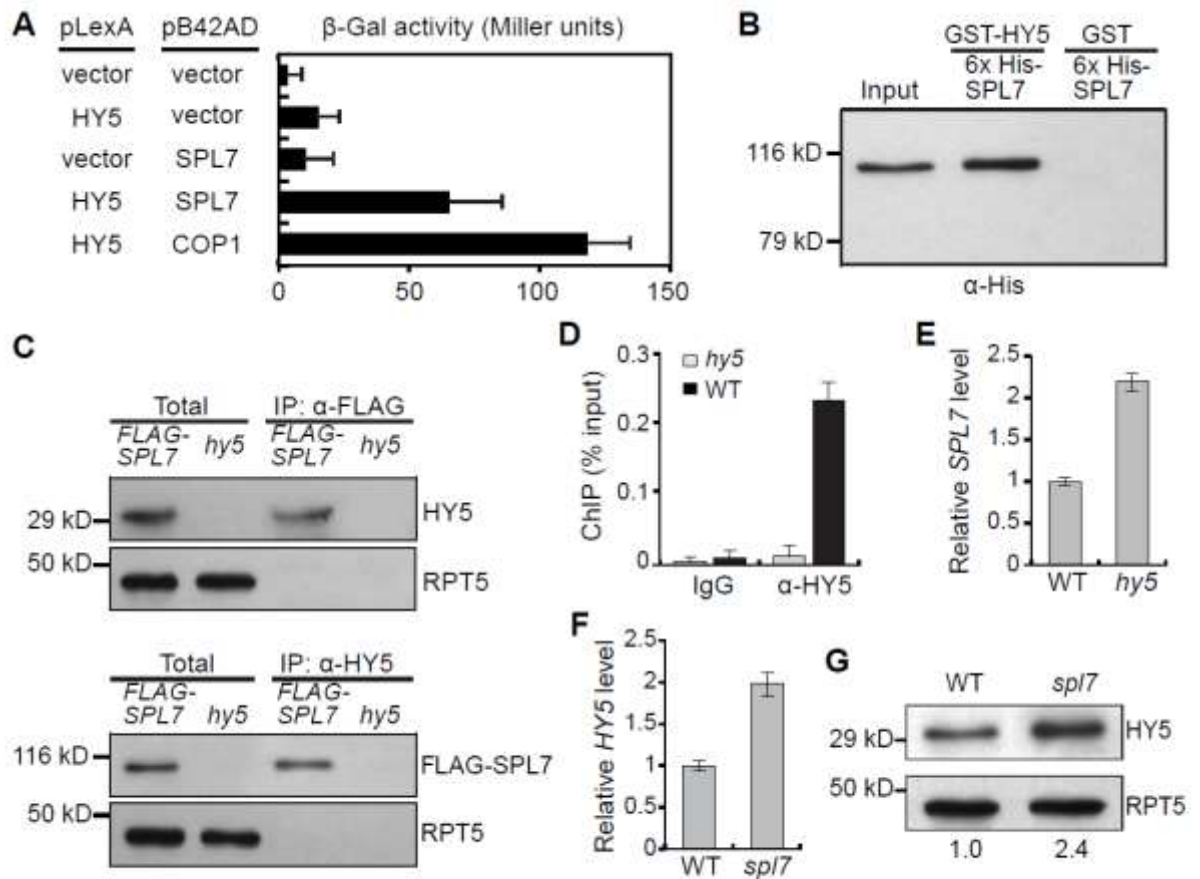


Figure 2. Genome-wide Analysis of the SPL7 Regulon.

- (A) Distribution of SPL7 binding peaks relative to the TSS. For genes with detected SPL7 binding, the regions 2,000 bp downstream and 1,000 bp upstream of the TSS were aligned and divided into 30 intervals. The number of genes with SPL7 binding (red) and the number of randomly selected genomic regions (blue) located in each interval were plotted.
- (B) ChIP-qPCR validation of 16 randomly selected SPL7 binding sites. ChIP was performed in *35S:FLAG-SPL7/spl7* and *spl7* seedlings with the anti-FLAG antibody. SPL7 binding profiles at these loci are depicted in Supplemental Figure 3. Values from qPCR analysis were normalized to their respective DNA inputs. Data are means \pm SD from three biological replicates.
- (C) Distribution of SPL7 binding sites across annotated genomic regions. Percentage of binding sites located in the 5' and 3' UTR (untranslated region), CDS (coding region), promoter (1 Kb region upstream of the TSS), and intergenic region is shown.
- (D) Both the GTAC tetranucleotide and the AGTACA/TGTACT hexanucleotide, but none of other related hexanucleotides, are overrepresented in the SPL7 binding sites compared to random genome sequences. *, $p < 0.01$ and **, $p < 0.00001$ by Student's *t* test.
- (E) Hierarchical clustering analysis of *SPL7* regulated genes. The heatmap is generated with differentially expressed genes from four pairwise comparisons that include *spl7*/HC vs WT/HC (column 1); WT/LC vs WT/HC (column 2); *spl7*/HC vs *spl7*/LC (column 3); and WT/LC vs *spl7*/LC (column 4). Each row represents a gene whose

scaled expression value, denoted as the row *Z*-score, is plotted in a color scale with blue indicating higher expression and red indicating lower expression. Grouping of the four major clusters is indicated on the far right. SPL7 binding pattern is shown as column 5 in which a horizontal line indicates SPL7 binding to a given gene.

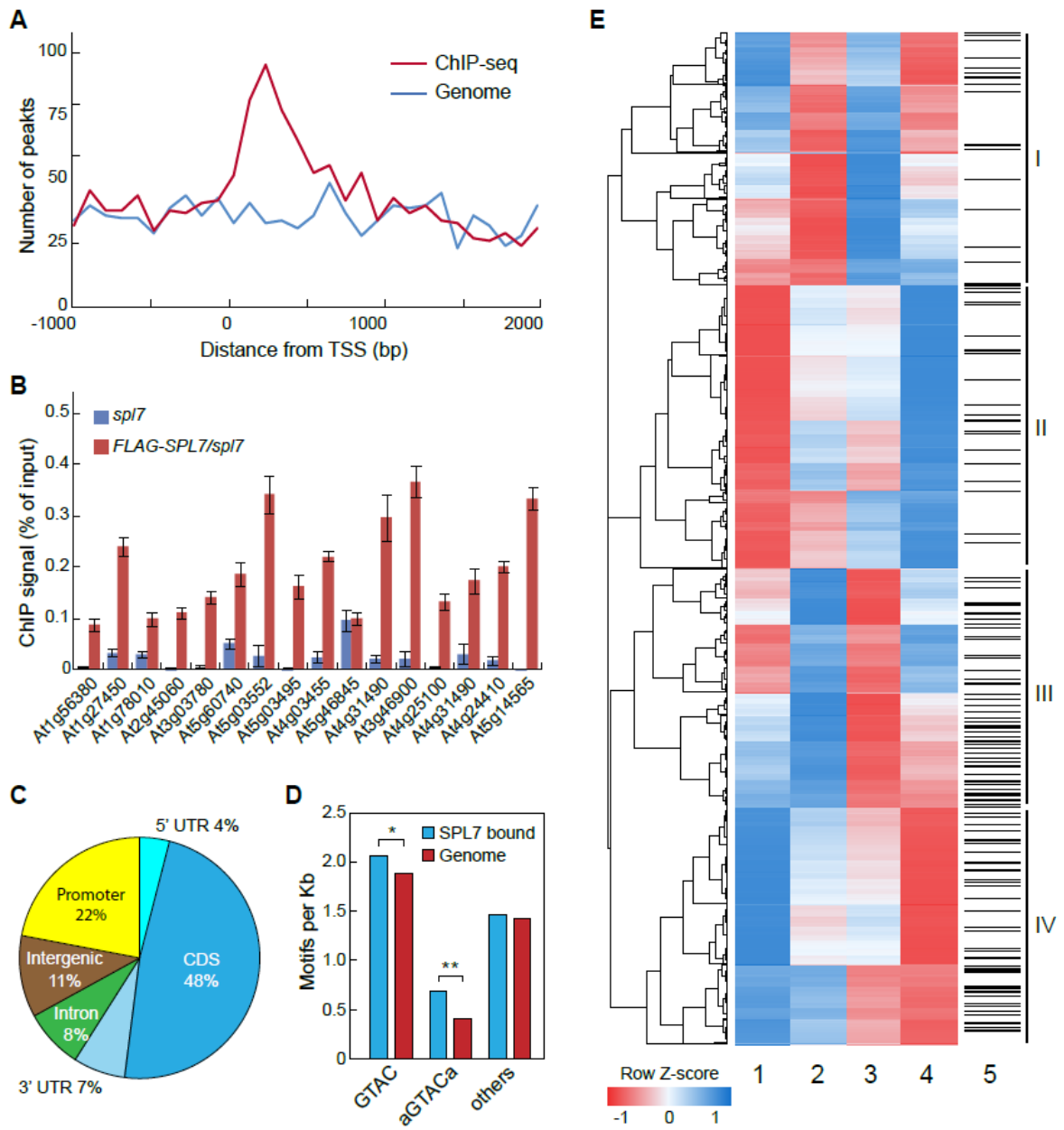


Figure 3. SPL7 and HY5 Co-regulate a Large Cohort of Genes.

- (A) G-box is significantly enriched in the SPL7 binding sites compared to random genomic sequences (posterior probability = 1).
- (B) Clustering of SPL7 and HY5 binding sites defined by the global ChIP data. Distance between neighboring SPL7 and HY5 binding sites was calculated and the number of sites as a function of distance in 250 bp intervals is plotted in blue. The control (red) summarizes 100 times simulation of randomly selected genomic loci by the same analysis with the error bars representing SD. The vertical dashed line indicates a cutoff distance (750 bp) below which SPL7 and HY5 binding sites show significant clustering based on hypergeometric test ($p < 0.001$).
- (C) Venn diagram showing the overlap of *SPL7* and *HY5* targeted genes. Using the cutoff illustrated in the above panel, 586 genes were considered to be bound by both *SPL7* and *HY5*.
- (D) Venn diagram on the left shows the overlap of *SPL7*- and *HY5*-dependent genes under the LC/HL condition. Heatmaps on the right illustrate expression changes of the 1,090 genes in *spl7* and *hy5* compared to wild type. Top, 582 genes influenced by *SPL7* and *HY5* in the opposite direction. Bottom, 508 genes influenced in the same direction.
- (E) Anthocyanin biosynthesis as a typical pathway co-regulated by *SPL7* and *HY5*. Biochemical steps leading to anthocyanin production are depicted on the left with genes encoding the relevant enzymes listed. Boxes on the right represent differentially expressed genes in either *spl7* or *hy5* that are involved in the pathway. Relative expression levels of these genes in either *spl7* or *hy5* are shaded with different colors

with red indicating reduced expression, blue indicating increased expression, and blank indicating no significant change in *spl7* or *hy5* compared to wild type.

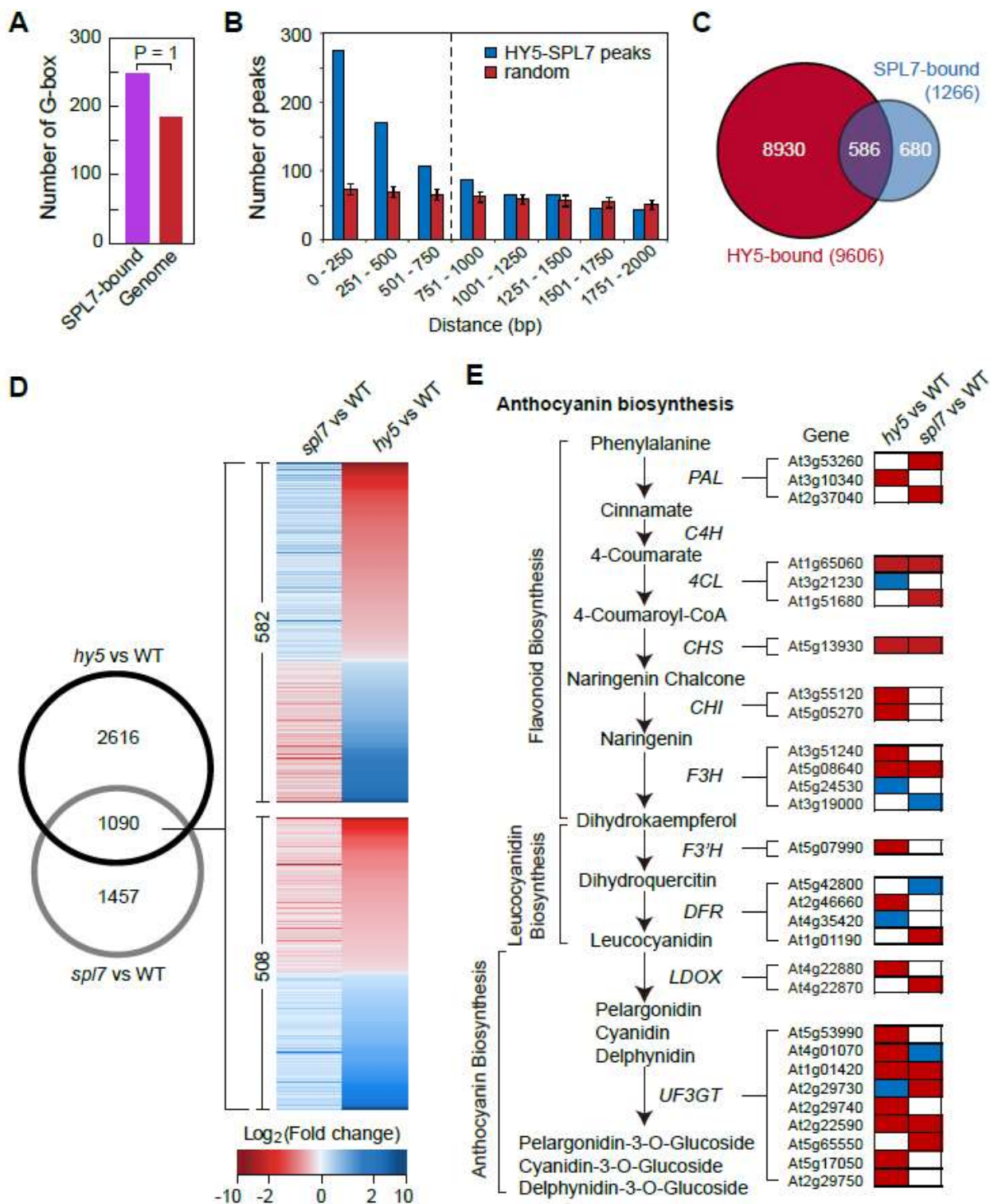


Figure 4. SPL7 and HY5 Co-bind to the *MIR408* Promoter

- (A) SPL7 and HY5 occupancy at the *MIR408* locus based on global ChIP data, which were mapped onto the *Arabidopsis* genome coordinates and visualized using the Affymetrix Integrated Genome Browser. Position of pre-miR408 is depicted as a black arrow. Orange and blue ovals in the promoter region represent the GTAC and G-box like motifs, respectively. Probes used in subsequent EMSA analyses are indicated.
- (B) Confirmation of HY5 and SPL7 binding to the *MIR408* promoter by ChIP-qPCR analysis. ChIP was performed in the indicated genotypes using either the anti-HY5 or anti-FLAG antibody. Data are means \pm SD from three biological replicates. The numbers 1 to 4 denote *hy5*, WT, *spl7*, and *35S:FLAG-SPL7*, respectively.
- (C) EMSA analysis of HY5 binding to the G-box in the *MIR408* promoter (Probe I in A). Lane 1, labeled probe alone; lane 2, labeled probe incubated with recombinant HY5; lanes 3 to 5, excessive unlabeled probe was added as competitor with the following competitor/probe ratios: 50 (lane 3), 100 (lane 4), and 200 (lane 5); lane 6 and 7, a mutated probe (CACGTG changed to CTGCAG) incubated with the same amount of HY5 as in lanes 2-5 or eight times more HY5, respectively.
- (D) Co-binding of SPL7 and HY5 to the *MIR408* promoter revealed by EMSA. Labeled Probe II was incubated without protein (lane 1), with HY5 alone (lane 2), SPL7 alone (lane 3), both HY5 and SPL7 (lane 4), HY5 plus SPL7 and 200-fold excessive mutated probe (CACGTG changed to CTGCAG) as a competitor (lane 5), and HY5 plus SPL7 and 200-fold excess mutated probe (GTAC changed to CATG) as a competitor (lane 6). FP, free probe.

(E) Yeast one-hybrid assay testing co-binding of *SPL7* and *HY5* to the *MIR408* promoter.

Yeast cells containing *pMIR408:LacZ* were transformed with *HY5*, *SPL7*, or *HY5* plus *SPL7* fused with the *Gal4* activation domain (AD) and grown on media containing X-gal. Cells expressing AD alone were used as the negative control.

(This work was done in collaboration with Dr. Huiyong Zhang)

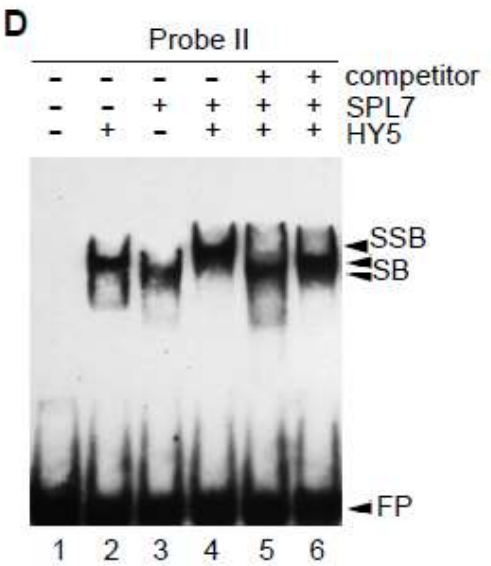
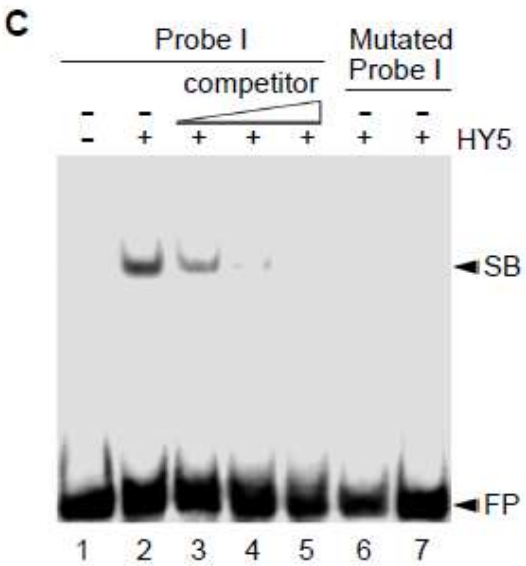
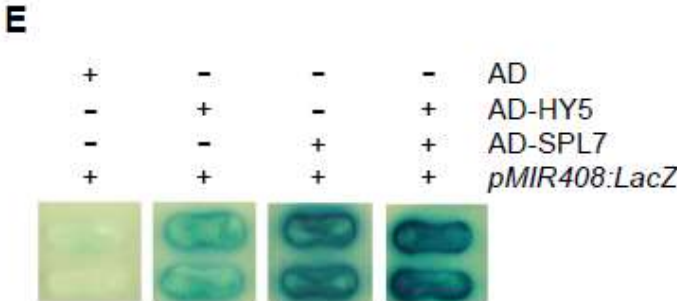
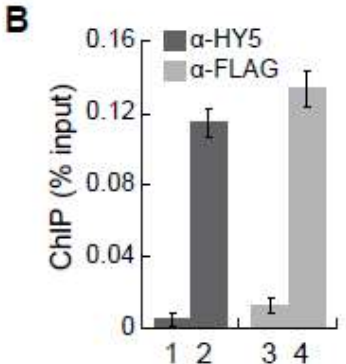
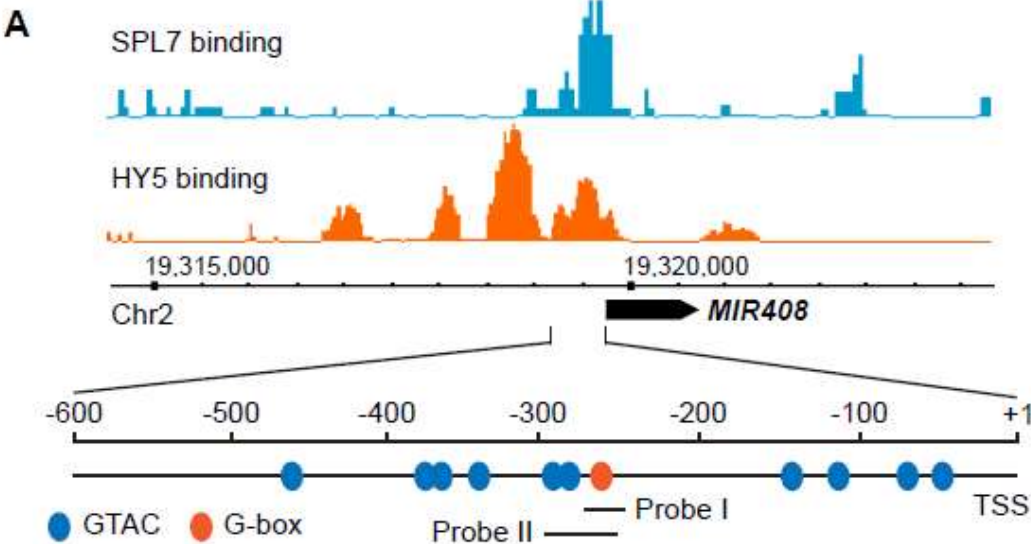


Figure 5. *HY5* and *SPL7* Coordinately Mediate *MIR408* Expression in Response to Changing Light and Copper Conditions

(A) RNA blot analysis of miR408 levels in WT, *hy5*, and *spl7* seedlings under four different light and copper regimes. U6 snRNA was used as the loading control.

(B) qRT-PCR analysis of miR408 levels and (C) the miR408 target gene *LAC13* in WT, *hy5*, and *spl7* seedlings. For comparison, levels of miR408 and *LAC13* in LC/HL were set as one. The star signs indicate statistical difference by Student's *t* test. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

(D) GUS activity in various transgenic plants. The *pMIR408:GUS* or *pMIR408m:GUS* reporter was expressed in the wild type or the indicated mutant background. Seedlings grown under different combinations of light and copper conditions were stained for GUS activity and visualized.

(E) qRT-PCR analysis of miR408 and pri-miR408 (top) and RNA blot analysis of miR408 (bottom) levels in WT, *spl7*, *hy5*, and *hy5 spl7* seedlings under the LC/HL condition.

(F) qRT-PCR analysis of *LAC13* transcript levels in the indicated genotypes under the LC/HL condition. Levels of miR408 and *LAC13* in wild type were set as one

(G) GUS activity in transgenic plants expressing *pMIR408:GUS* in the WT, *spl7*, *hy5*, and *hy5 spl7* backgrounds. Data for qRT-PCR are means \pm SD from three biological replicates. Bar = 1 cm.

(This figure was generated by Dr. Huiyong Zhang)

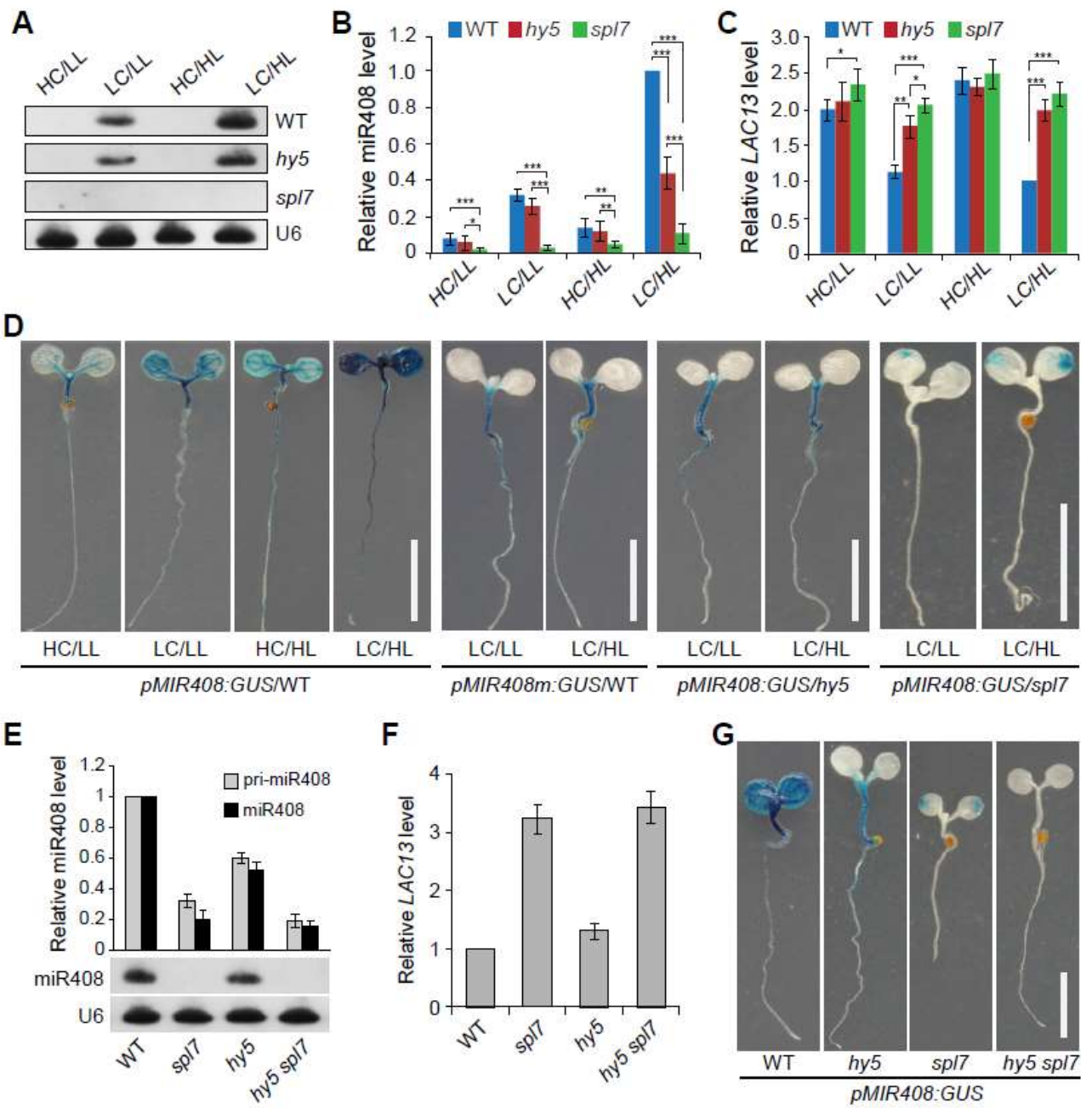


Figure 6. The miR408 Target Genes *LAC12* and *LAC13* Are Differentially Regulated in the *HY5-SPL7* Network.

- (A) Confirmation of miR408 targeting on *LAC12* and *LAC13* by 5'RNA ligase-mediated RACE. Gene structures of *LAC12* and *LAC13* are shown on the top with shaded boxes representing exons. The complementary mRNA and miRNA sequences are shown on the bottom. Perfect base pairing is shown as a vertical dash whereas G:U wobble pairing indicated by "o". Vertical arrows mark the sequenced cleavage sites with the frequency of clones shown.
- (B) Analysis of HY5 and SPL7 binding to the *LAC12* and *LAC13* promoters by ChIP-qPCR analysis. ChIP was performed in the indicated genotypes using either the anti-HY5 or anti-FLAG antibody. The resultant DNA was analyzed by qPCR with the values normalized to their respective DNA inputs.
- (C) Regulatory interactions among *HY5*, *SPL7*, *MIR408*, *LAC12*, and *LAC13*. Mutual inhibition between *HY5* and *SPL7* is based on molecular data presented in Figure 1. Coordinated transcriptional regulation of *MIR408* by *HY5* and *SPL7* is deduced from Figure 4. Regulation of *LAC12* and *LAC13* by miR408 at the post-transcriptional level and by *HY5* at the transcriptional level is as indicated by panel A and B, respectively.
- (D) Quantitative analysis of *HY5*, *SPL7*, miR408, pri-miR408, *LAC12*, and *LAC13* transcript levels in wild type and *hy5* seedlings. Seedlings were grown under the LL/LC condition and transferred to HL/LC at time zero, and assayed by qRT-PCR at indicated time points thereafter. Data for ChIP-qPCR or qRT-PCR are means \pm SD from three biological replicates.

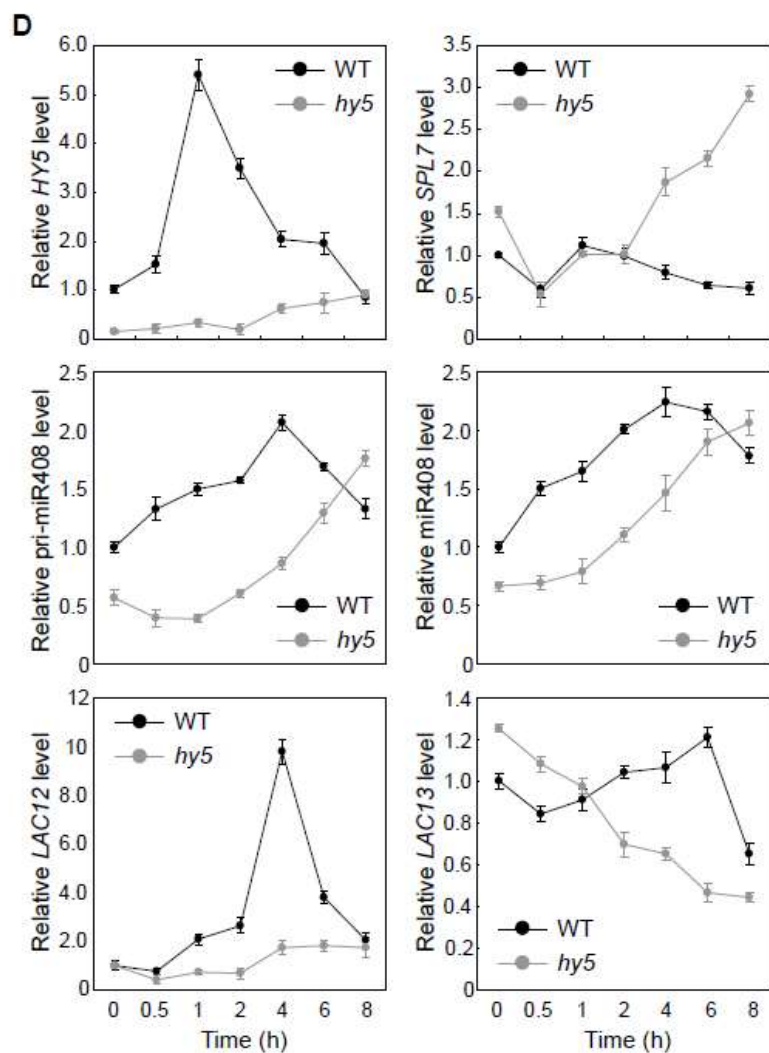
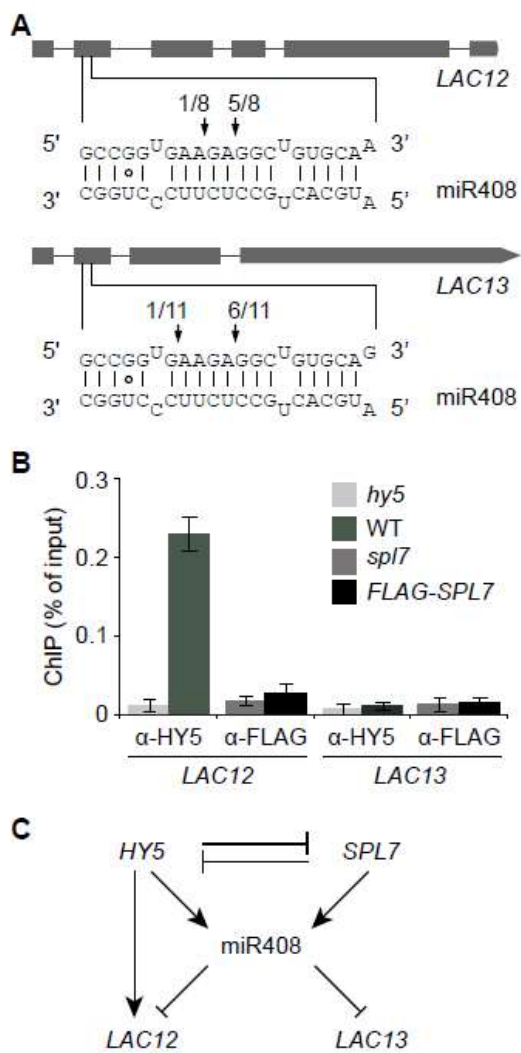


Figure 7. Complementation of the *hy5* and *spl7* Mutations by *MIR408* Overexpression.

(A) Seedling morphology of eight genotypes as indicated in which miR408 has varied expression levels. Seedlings were grown under the HL/LC condition and photographed seven days after germination. Bar = 1 cm.

(B) Quantitative measurement of fresh weight, (C) hypocotyl length, (D) chlorophyll, and (E) anthocyanin content in seedlings of the eight genotypes. Genotypes labeled with different numbers of the star sign are statistically different ($p < 0.01$) by ANOVA test. Data are means \pm SD from n biological replicates where $n \geq 30$ for (B) and (C), $n = 3$ for (D) and (E).

(This figure was generated by Dr. Huiyong Zhang)

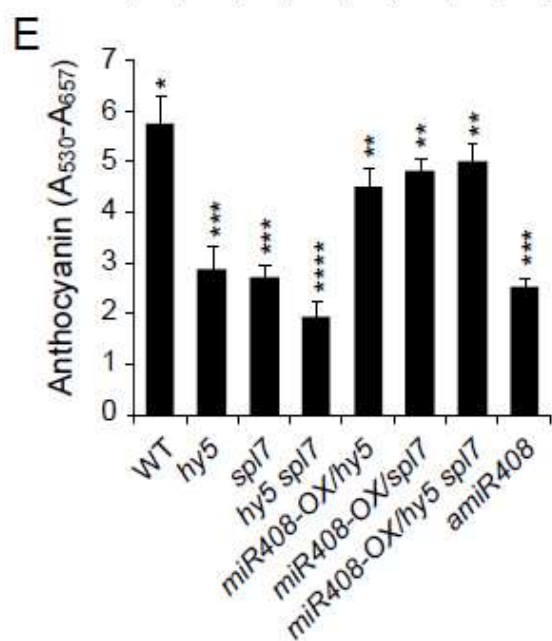
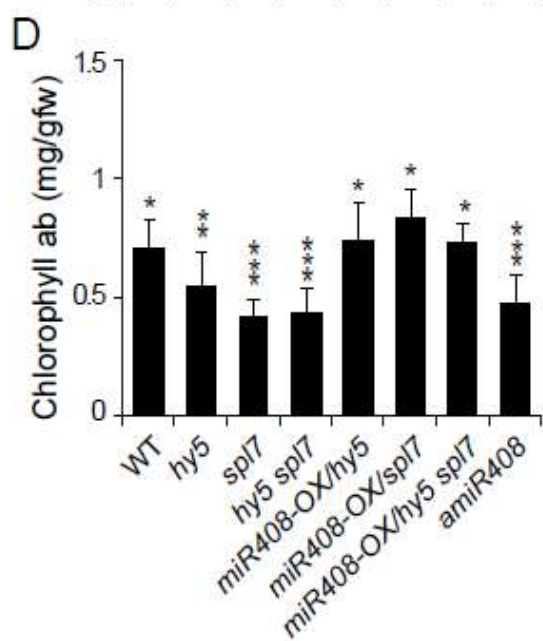
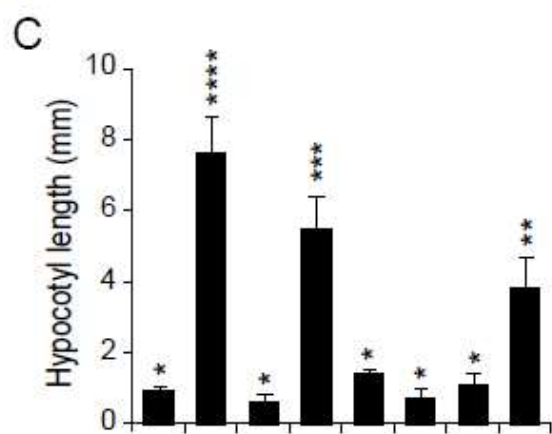
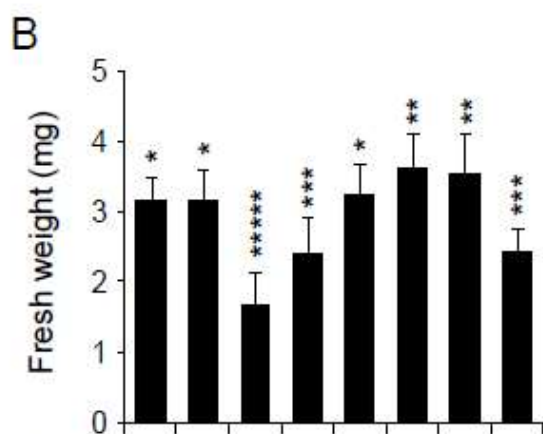
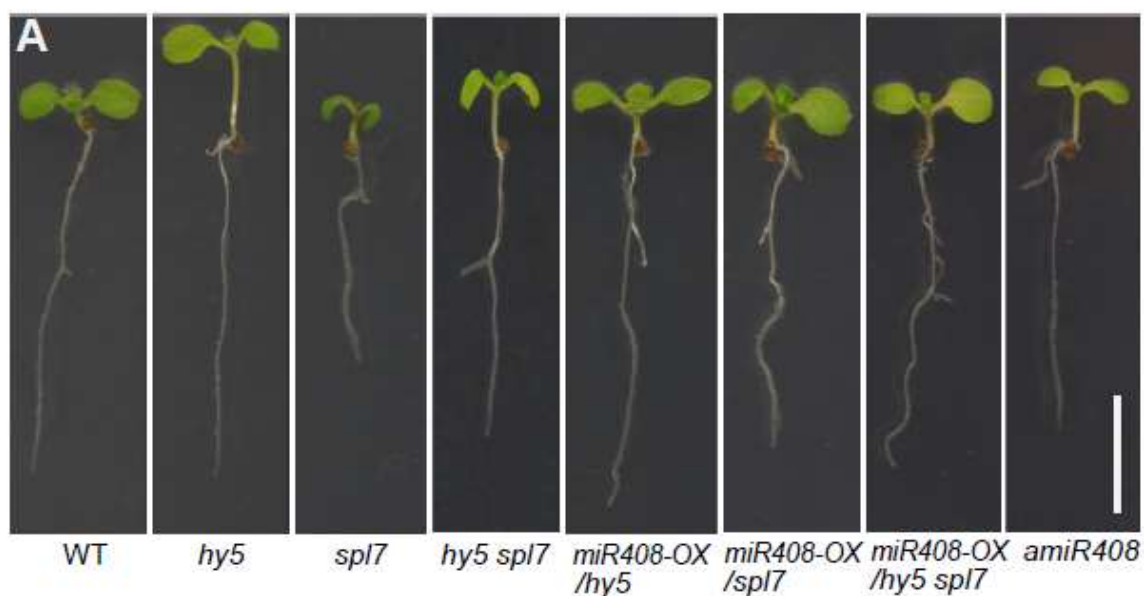


Figure 8. Cellular Function of miR408.

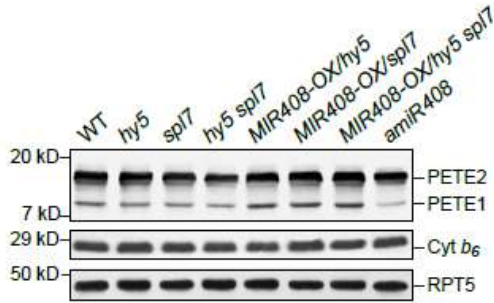
- (A) Detection of PC isoforms in wild type *Arabidopsis* seedlings and various genotypes with altered *MIR408* expression. Total protein prepared from seedlings grown in the LC/HL condition was fractionated and probed with a commercial PC antibody. Two PC isoforms, PETE1 and PETE2, were detected with PETE2 being the more abundant isoform. Chloroplast cytochrome *b*₆ (Cyt *b*₆) and RPT5 were used as controls. Blots shown are from one representative of three independent experiments.
- (B) Quantification of total PC protein levels in seedlings of various genotypes. Intensity of the bands corresponding to PETE1 and PETE2 was acquired using Image J to determine the total PC level, which was then normalized against RTP5 and set to one for wild type. Cyt *b*₆ level was also quantified as a control. Data are means \pm SD from three biological replicates. Samples denoted with the same letter are statistically different (ANOVA test; $p < 0.01$).
- (C) Relative copper contents in the chloroplast of various genotypes as indicated. Copper content in isolated chloroplasts and whole seedlings of the same genotype was determined separately. Values shown are percentages of chloroplastic copper contents over those of the whole seedlings. Data are means \pm SD from four biological replicates. Samples denoted with the same letter are statistically different (ANOVA test; $p < 0.01$).
- (D) A model for *SPL7-HY5* regulated *MIR408* activation in copper homeostasis and plant development. Depicted on the left are simplified copper transport and utilization pathways that include the main copper transporter COPT1 and metallochaperones CCS, CCH, and ATX1. These chaperones deliver copper to different internal transporters

such as Golgi-localized RNA1 and chloroplast-specific importers PAA1 and PAA2. Together the chaperones and transporters mediate copper delivery to specific protein targets such as CSD1 in the cytosol, CSD2 in chloroplast, PC in the photosynthetic electron transport chain, and LAC13 in the secretory pathway. Elevated miR408, which is promoted by HY5 and SPL7, represses several genes in the copper secretory pathway.

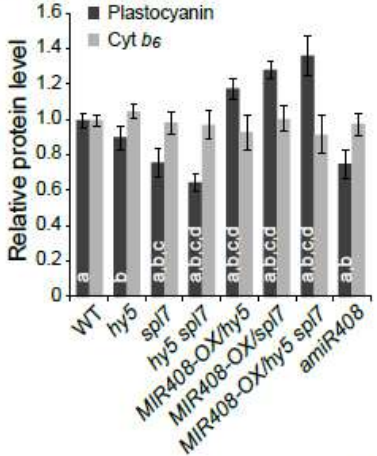
(E) Genetically, as shown in the three panels on the right, silencing miR408 expression (*amiR408*) suppresses PC, reduces chloroplast copper, and compromises seedling development, phenotypes reminiscent of the *hy5 spl7* double mutant. Conversely, constitutive action of miR408 in the mutant backgrounds complements such phenotypes. Thus cellular function of miR408 is to promote copper allocation to as well as abundance of PC, thereby constituting one specific regulatory route downstream of *SPL7* and *HY5*, though its impact on other components of copper homeostasis remains to be determined.

(This work was done in collaboration with Dr. Huiyong Zhang)

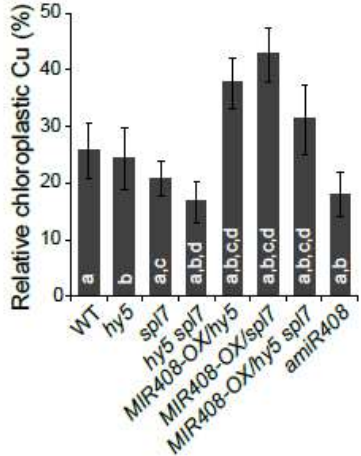
A



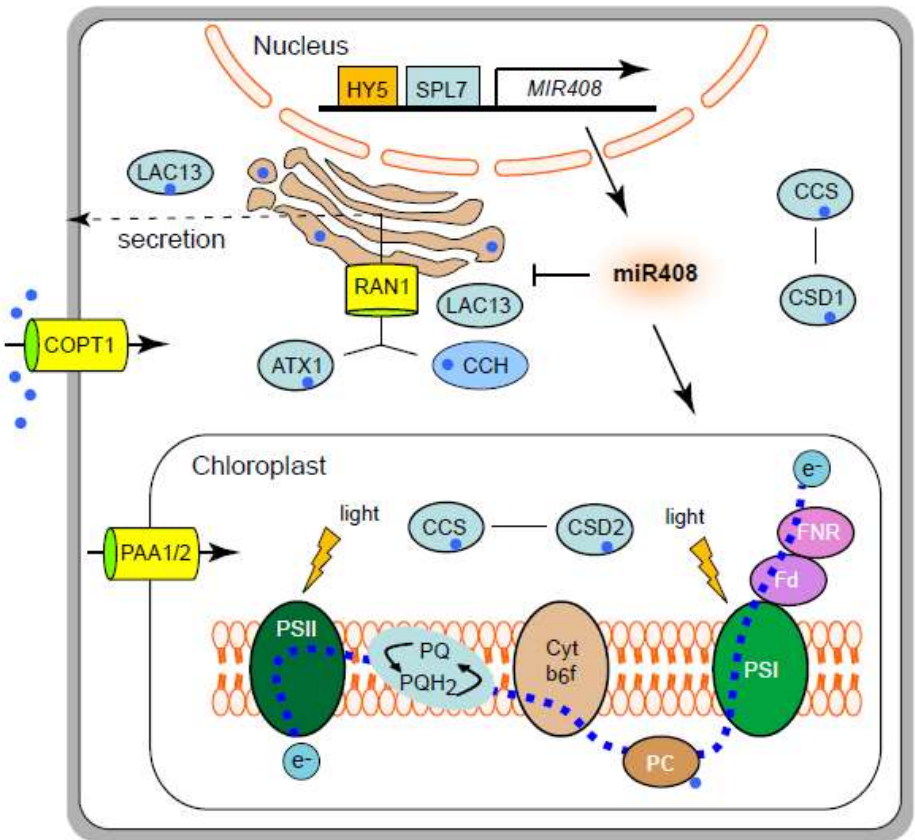
B



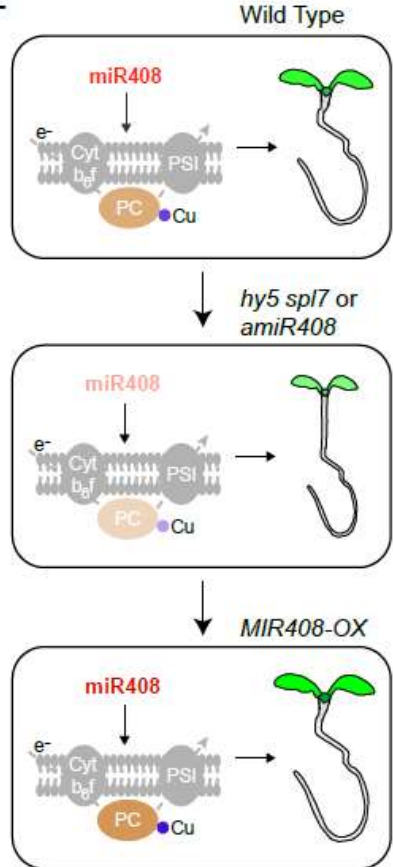
C



D



E



Supplementary materials

Supplemental Data

Supplemental Figure 1. Analysis of the Light-copper Crosstalk in *Arabidopsis*.

Supplemental Figure 2. HY5 Directly Binds to the *SPL7* promoter.

Supplemental Figure 3. Analysis of *SPL7*-targeted Genes.

Supplemental Figure 4. Identification and Validation of Enriched DNA Elements in the *SPL7* Occupied Sites.

Supplemental Figure 5. Global Analysis of *SPL7* Regulated Genes.

Supplemental Figure 6. Validation of RNA-sequencing Data by qRT-PCR.

Supplemental Figure 7. Analysis of Genes Co-bound by *SPL7* and HY5.

Supplemental Figure 8. *SPL7* and *HY5* Coordinately Regulate Photosynthesis.

Supplemental Figure 9. Functional Analysis of miR408.

Supplemental Figure 10. Growth Phenotypes of Adult Plants with Altered *MIR408* Levels.

Supplemental Figure 11. Copper Content in Whole Seedlings and the Chloroplast Fractions of Various Genotypes with Altered *MIR408* Expression.

Supplemental Table 1. Summary of ChIP-sequencing and RNA-sequencing Data.

Supplemental Table 2. List of Genes for ChIP-sequencing and RNA-sequencing Data.

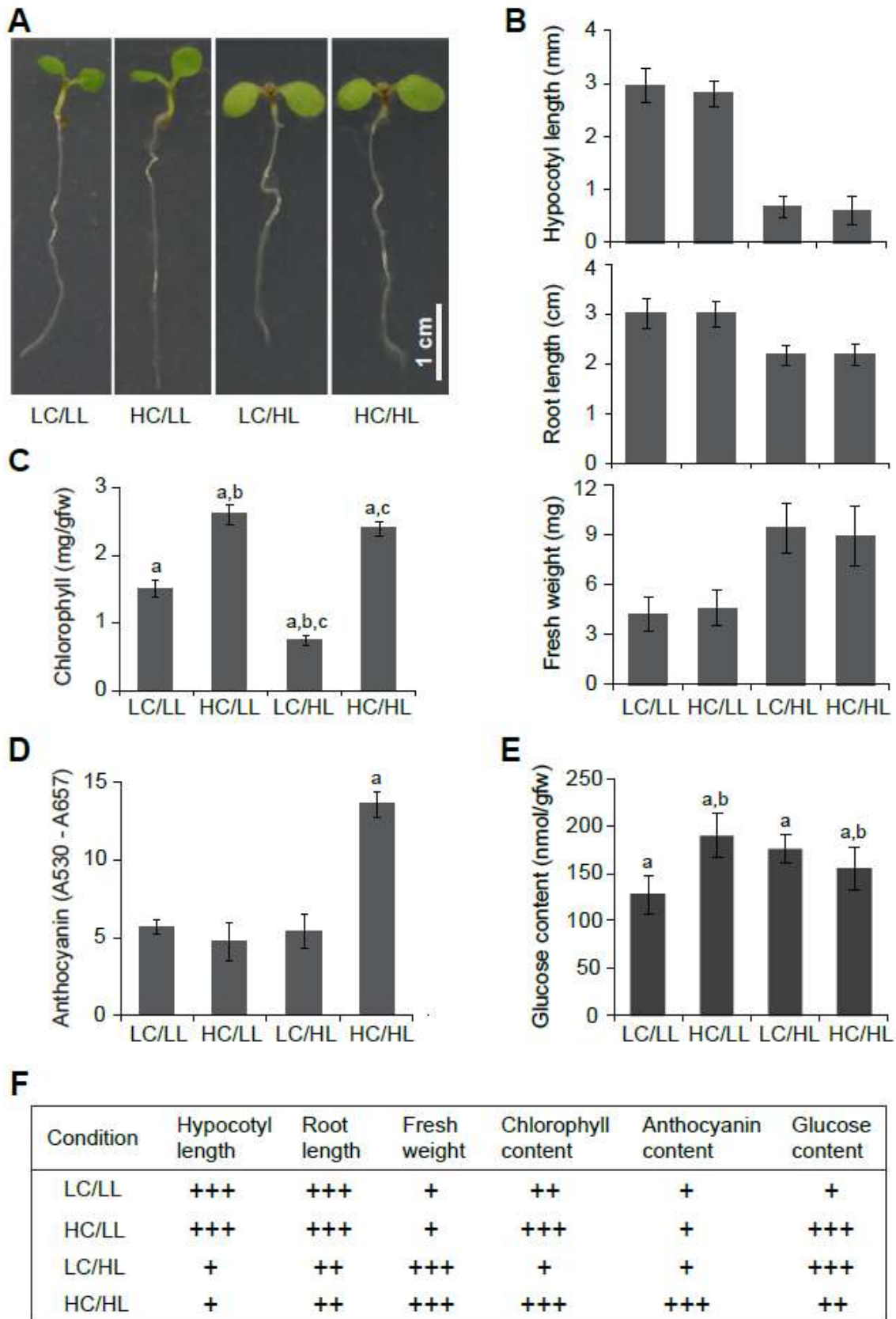
Supplemental Table 3. Oligonucleotide Sequences for the Primers and Probes Used in this Study.

Supplementary Figures

Supplemental Figure 1. Analysis of the Light-copper Crosstalk in *Arabidopsis*.

- (A) Morphology of wild type seedlings grown under the LC/LL, LC/HL, HC/LL, and HC/HL conditions.
- (B) Quantitative measurement of hypocotyl length, root length, and fresh weight of seedlings grown in the four combinations of light and copper regimes. Data are means \pm SD ($n \geq 30$).
- (C) Total chlorophyll and (D) anthocyanin contents in wild type seedlings grown under the indicated conditions. Data are averages of three independent replicates. Letters indicate statistically significant groups (Student's t test, $p < 0.01$).
- (E) Glucose content in the shoots of seedlings grown in four indicated conditions. Data are means \pm SD of four biological replicates. Letters indicate statistically significant groups (Student's t test, $p < 0.05$).
- (F) Summary of the aforementioned measurements showing that seedlings display overall distinctive morphological and metabolic profiles under the four combinations of light and copper conditions. Numbers of the plus sign indicate statistic difference as determined in the previous panels.

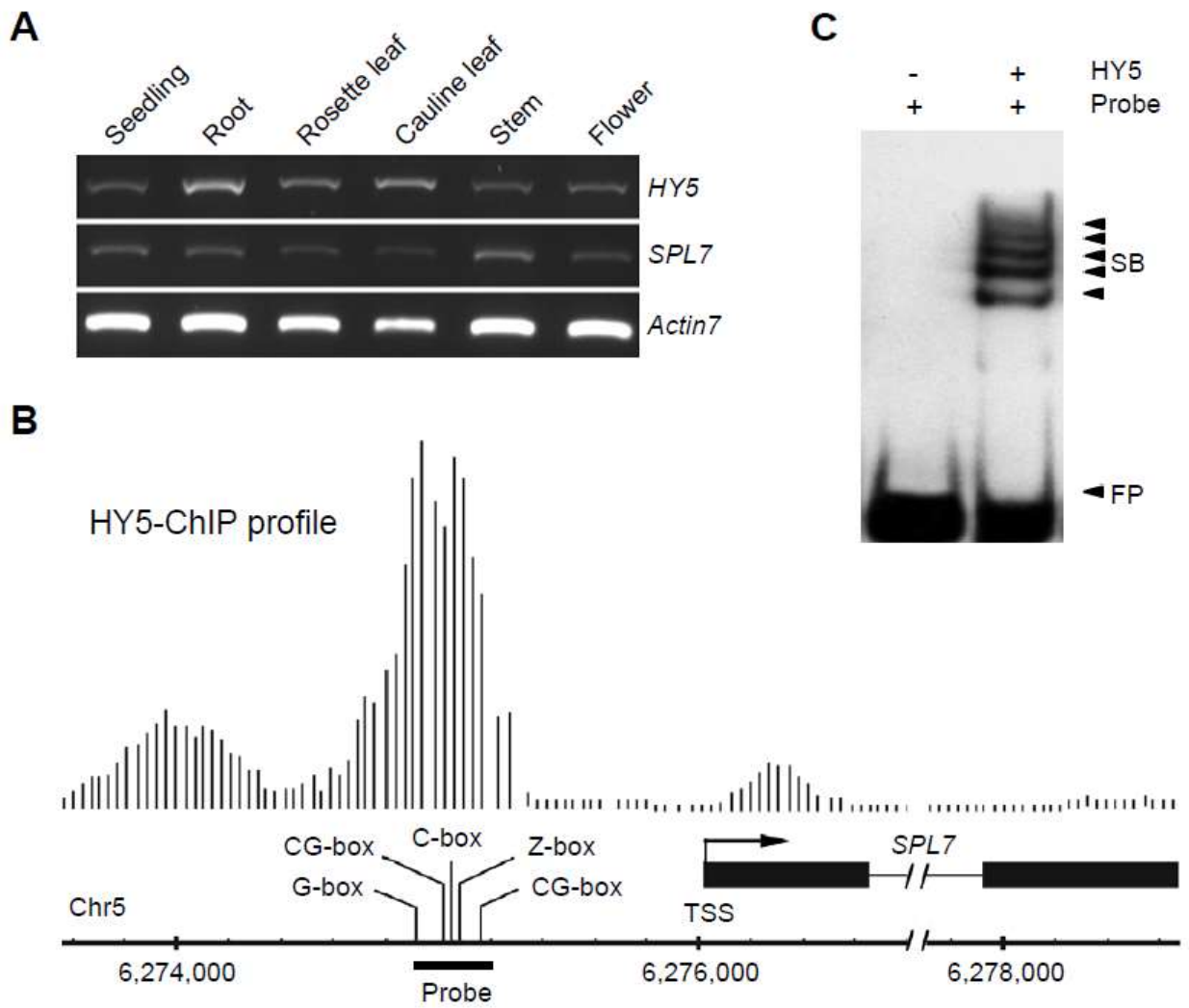
(This figure was generated by Dr. Huiyong Zhang)



Supplemental Figure 2. HY5 Directly Binds to the *SPL7* Promoter.

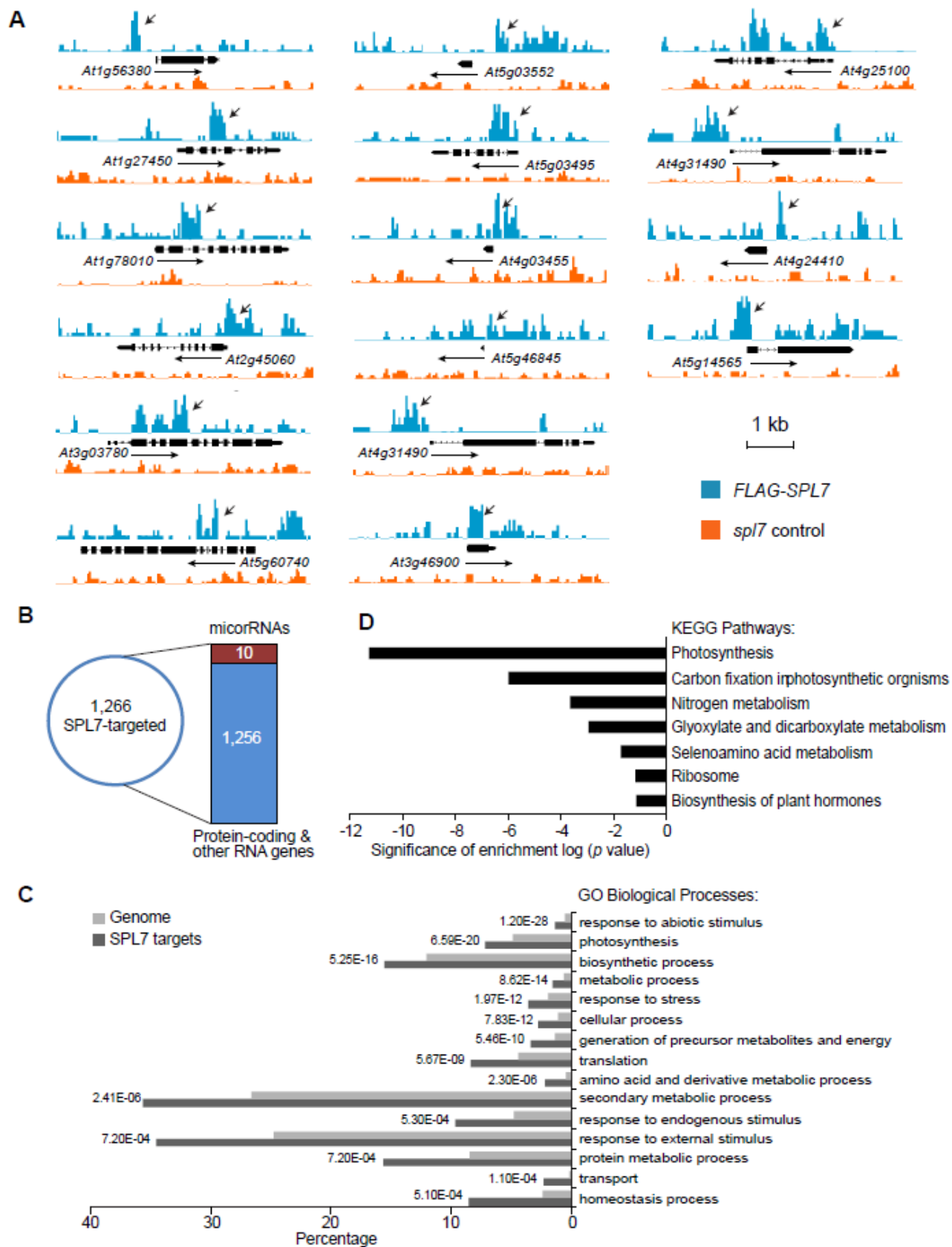
- (A) RT-PCR analysis of the transcript levels of *HY5* and *SPL7* in young seedlings and different tissues of adult plants. *Actin7* was used as a control.
- (B) HY5 occupancy pattern at the *SPL7* locus. Raw ChIP-chip data was obtained from Zhang et al. (2011) and visualized using the Affymetrix Integrated Genome Browser. Vertical lines represent individual probes from the genome tiling microarray with height of the lines proportional to the signal intensity. Gene structure, transcriptional direction, and chromosomal location of *SPL7* are depicted. The G-box like motifs, TSS, and position of the probe used for EMSA are indicated as well.
- (C) EMSA analysis of HY5 binding to the *SPL7* promoter. The first lane shows only the labeled probe (-1146 to -819 from the TSS) that contains all five G-box like motifs. The second lane shows the labeled probe together with recombinant HY5 protein. Arrows indicate the free probe (FP) and the shift bands (SB).

(This figure was generated by Dr. Huiyong Zhang)



Supplemental Figure 3. Analysis of SPL7-targeted Genes.

- (A) SPL7 binding profiles validated by ChIP-qPCR of *At1g56380*, *At5g03552*(*MIR822A*), *At4g25100*, *At1g27450*, *At5g03495*, *At4g31490*, *At1g78010*, *At4g03455*(*MIR447B*), *At4g24410*, *At2g45060*, *At5g46845* (*MIR160C*), *At5g14565* (*MIR398C*), *At3g03780*, *At4g31490*, *At5g60740* and *At3g46900*. ChIP-sequencing data at the selected loci were visualized using the Affymetrix Integrated Genome Browser. Vertical lines (SPL7 ChIP in blue and the negative control in orange) represent individual sequencing reads with height of the lines proportional to the copy number. Structure and transcriptional direction of the loci are depicted. Arrows indicate the binding sites selected for ChIP-qPCR validation.
- (B) SPL7 binding is associated with 1,266 genes including 10 *MIR* genes.
- (C) Enriched GO terms in the Biological Process category that are associated with the 1,266 SPL7-targeted genes. Frequencies of the GO terms for randomly selected genes from the genome and the corresponding P values are indicated.
- (D) Enriched pathways associated with the SPL7 targeted genes by KEGG (Kyoto encyclopedia of genes and genomes) analysis.



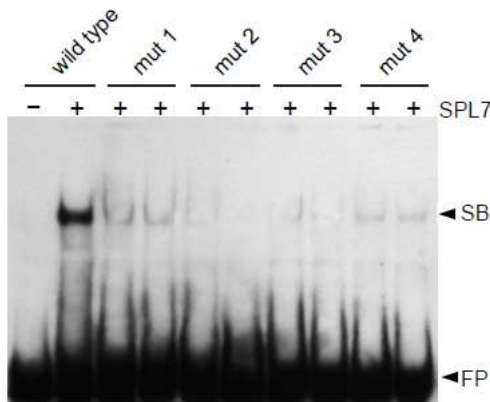
Supplemental Figure 4. Identification and Validation of Enriched DNA Elements in the *SPL7* Occupied Sites.

- (A) Effects of nucleotides immediately flanking the GTAC core sequence on the binding affinity for *SPL7*. EMSA assay was performed with recombinant *SPL7* and five synthesized probes, which differ only in the two nucleotides flanking the GTAC core motif. The symmetric nucleotides (A/T) showed enhanced affinity for *SPL7*.
- (B) A novel consensus sequence for *SPL7* binding, which was discovered in *SPL7* occupied sites using the motif finder MEME.
- (C) EMSA validation of *SPL7* binding to the novel motif. A probe corresponding to the consensus sequence shown in the previous panel (positions 1 to 17, TCTTCTTCTCCTTCCTC) was labeled (lane 1) and incubated with recombinant *SPL7* alone (lane 2) or in the presence of unlabeled probe at different concentrations (10-fold, 100-fold, and 200-fold; lanes 3-5).
- (D) DNA elements resembling other transcription factor binding sites in *SPL7*-occupied regions. The *SPL7* binding sites were scanned for motifs listed in the AGRIS and Transfac databases. Significantly overrepresented (left) and under-represented (right) motifs are shown. Numbers represent posterior probability for $P_{spl7} > P_{genome}$ which was calculated using 10,000 times Monte Carlo simulation in Matlab.

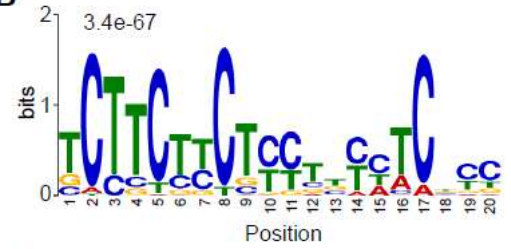
(This work was done in collaboration with Dr. Huiyong Zhang)

A

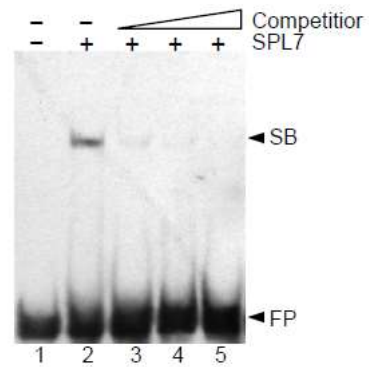
wild type: 5'-TCGC**AGTACA**AGACTTTGCAAAACAAGGG**AGTACA**AAATA-3'
 mut 1: 5'-TCGC**CGTACC**AGACTTTGCAAAACAAGGG**CGTACC**AAATA-3'
 mut 2: 5'-TCGC**CGTACA**AGACTTTGCAAAACAAGGG**CGTACA**AAATA-3'
 mut 3: 5'-TCGC**GGTACT**AGACTTTGCAAAACAAGGG**GGTACT**AAATA-3'
 mut 4: 5'-TCGC**CGTACG**AGACTTTGCAAAACAAGGG**CGTACG**AAATA-3'



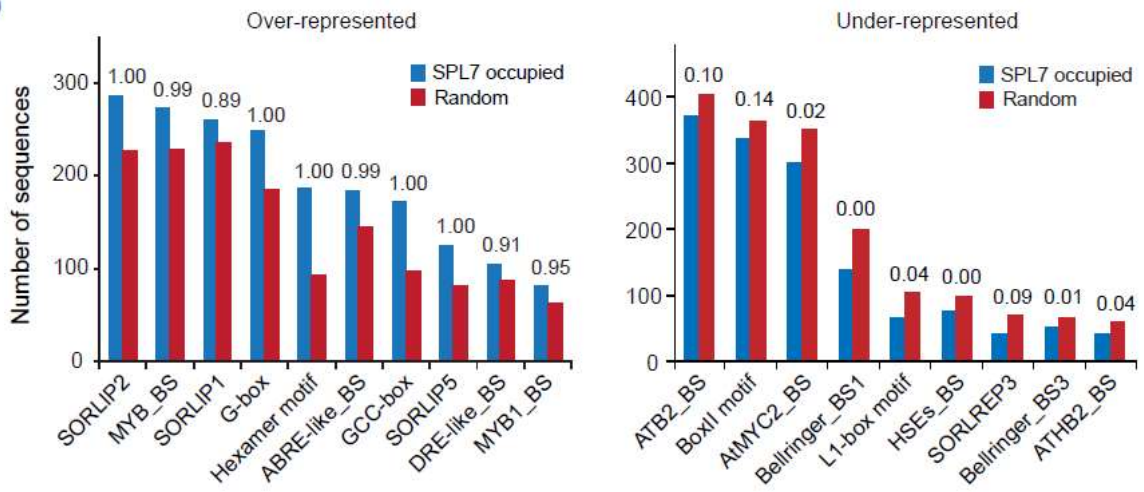
B



C



D

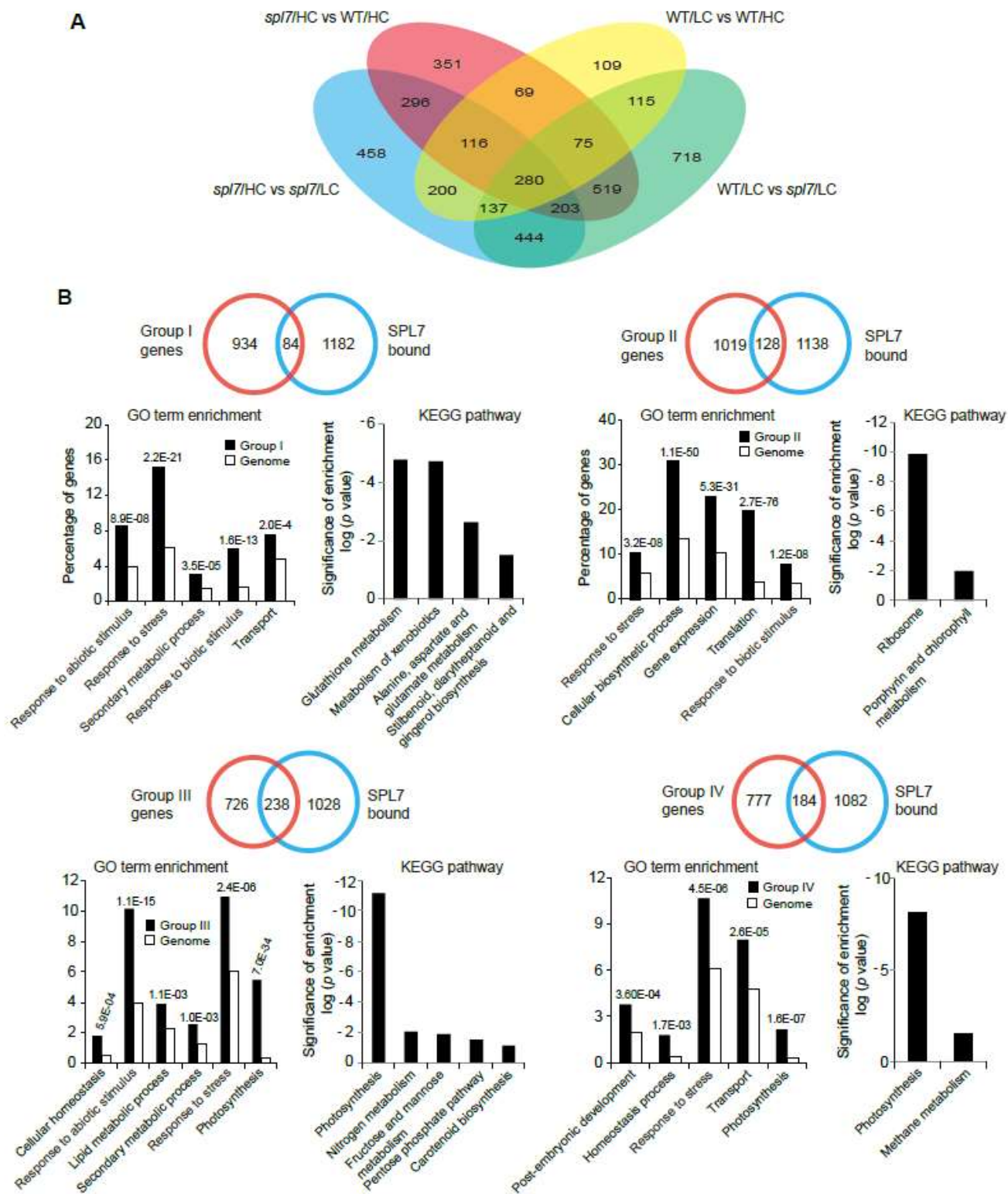


Supplemental Figure 5. Global Analysis of *SPL7* Regulated Genes.

(A) Venn diagram depicting differentially expression genes among the four pairwise comparisons as indicated.

(B) Detailed analysis of the four groups of genes as indicated in Figure 2 E in the main text.

For each group, three panels are shown. The top panel is Venn diagram showing the overlapping between that group and *SPL7* bound genes. Bottom left panel shows frequency of significantly enriched GO terms associated with genes in that group (black bar). Randomly selected genes from the genome (open bar) were used as control and the corresponding P values indicated. Bottom right panel shows KEGG pathway analysis.

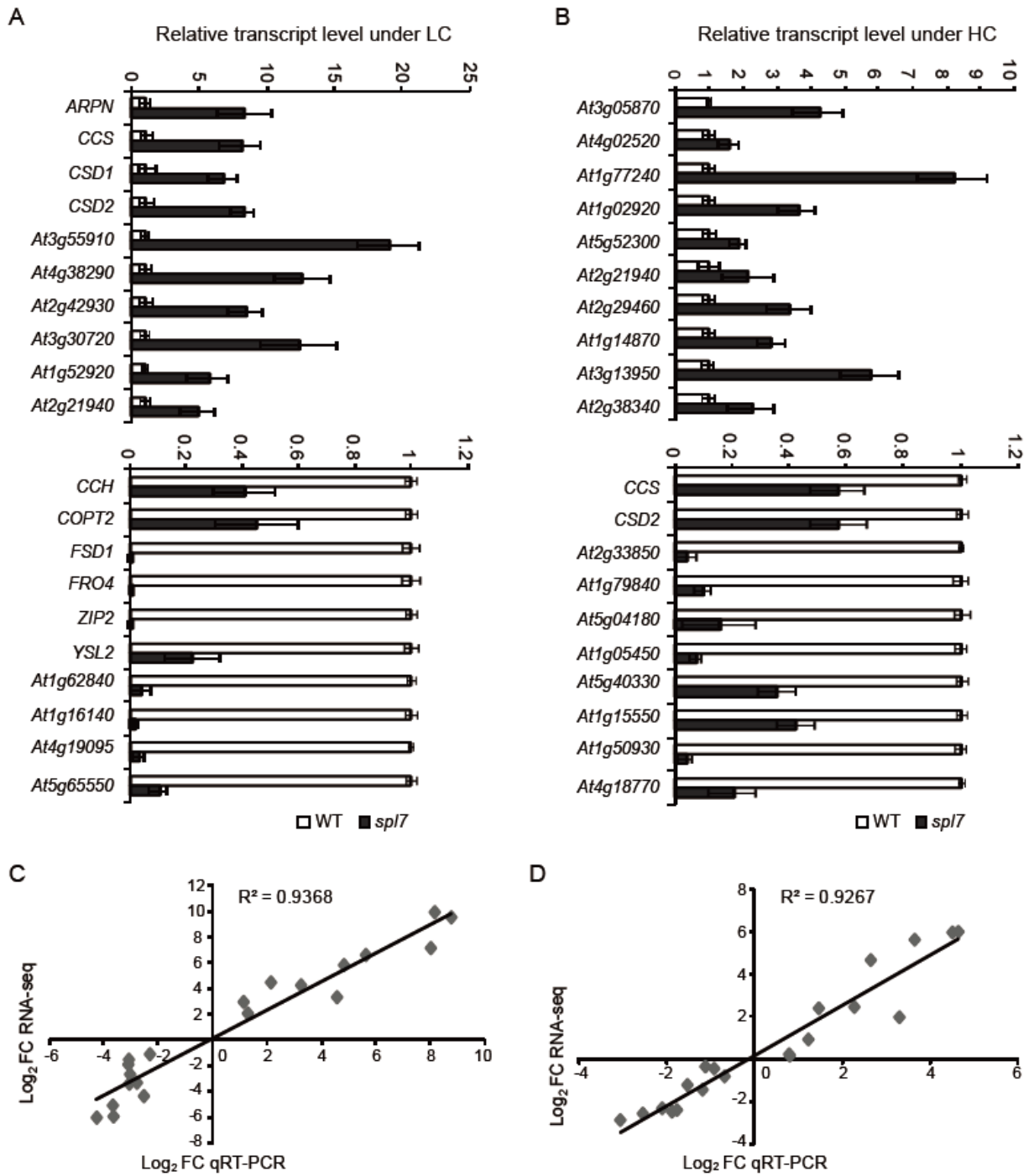


Supplemental Figure 6. Validation of RNA-sequencing Data by qRT-PCR.

(A) and (B) qRT-PCR analysis of the relative transcript levels of copper-responsive as well as randomly selected genes in WT and *spl7* seedlings grown under the LC (A) and HC (B) conditions. Data shown are transcript levels relative to *Actin7* and set to one for WT. Data are means \pm SD of three biological replicates.

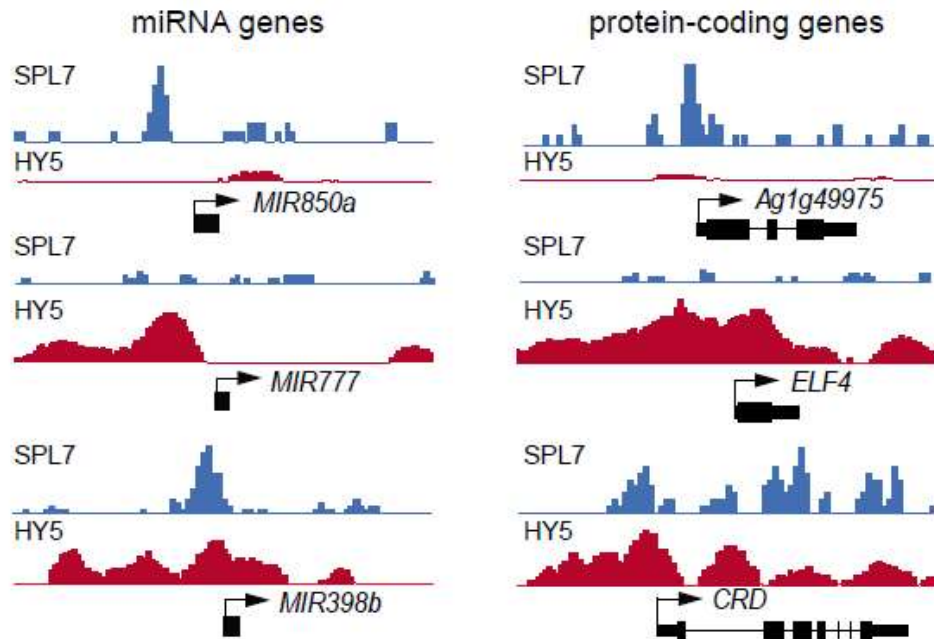
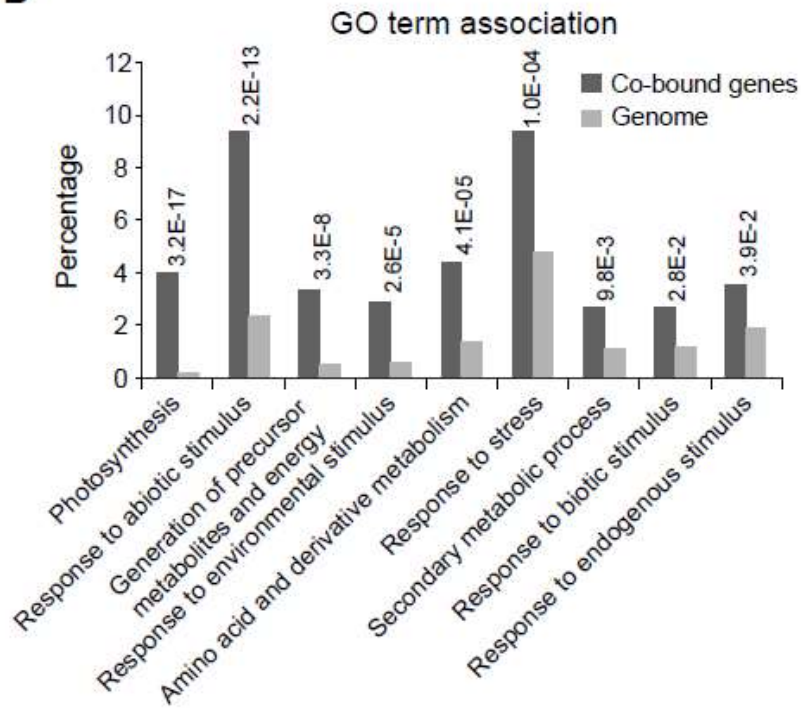
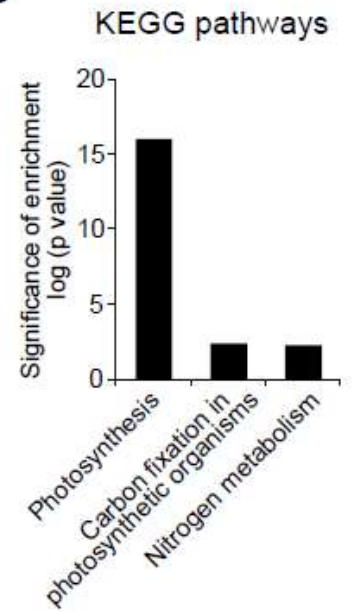
(C) and (D) Correlation between RNA-sequencing and qRT-PCR data on the selected genes under the LC (C) and HC (D) conditions. Pearson correlation was calculated using data points representing Log_2 transformed transcript level ratios between WT and the *spl7* mutant.

(This work was done in collaboration with Dr. Huiyong Zhang)



Supplemental Figure 7. Analysis of Genes Co-bound by SPL7 and HY5.

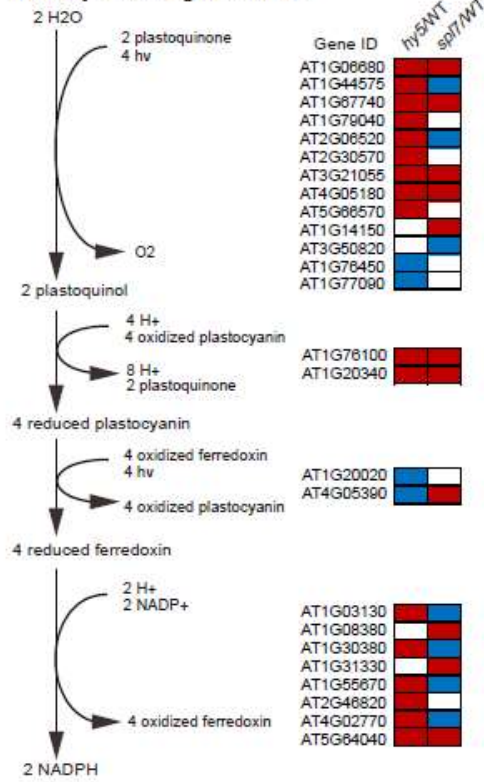
- (A) Representative examples of the 586 genes co-bound by SPL7 and HY5 as shown in Figure 3C in the main text. For each gene, SPL7 (blue) and HY5 (red) binding patterns are visualized using the Affymetrix Integrated Genome Browser. The left and right column includes three *MIR* and three protein-coding genes, respectively. For each column, the gene on top is bound primarily by SPL7, the gene in the middle is primarily bound by HY5, and the gene on the bottom is co-bound by both SPL7 and HY5.
- (B) Enriched GO terms in association with the 586 genes with both SPL7 and HY5 binding. Frequencies of the GO terms for randomly selected genes from the genome and the corresponding P values are indicated.
- (C) KEGG pathway enrichment analysis of SPL7-HY5 co-bound genes.

A**B****C**

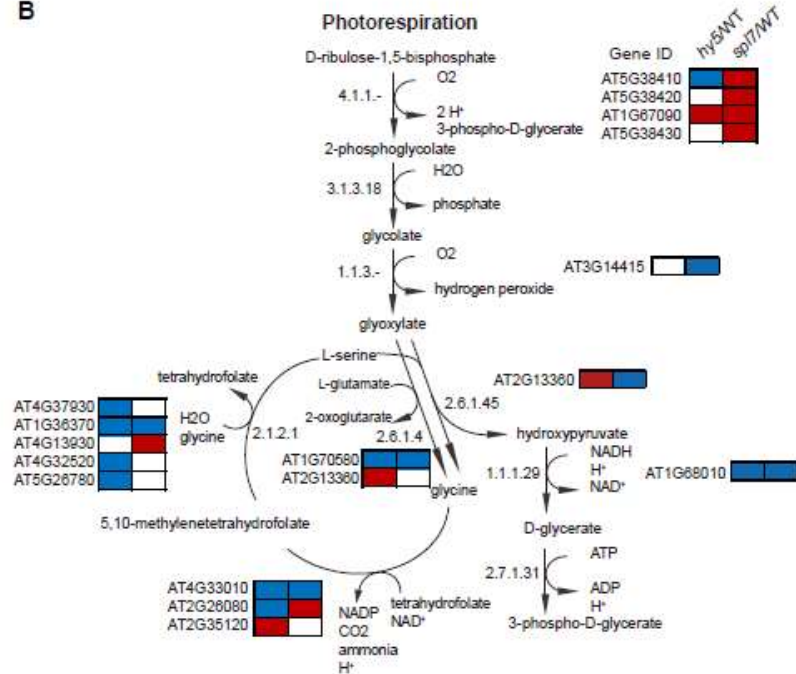
Supplemental Figure 8. *SPL7* and *HY5* Coordinately Regulate Photosynthesis.

(A) Genes involved in the light reactions of photosynthesis, (B) photorespiration, and (C) Calvin-Benson-Bassham cycle, are coordinately regulated by *SPL7* and *HY5*. In each panel, biochemical steps in the process as well as the corresponding enzymes are depicted. Boxes represent differentially expressed genes in either *spl7* or *hy5* that are involved in the pathway. Relative expression levels of these genes in either *spl7* or *hy5* are shaded with different colors with red indicating reduced expression in *spl7* or *hy5*, blue indicating increased expression in *spl7* or *hy5*, and blank indicating no significant change in *spl7* or *hy5* compared to wild type.

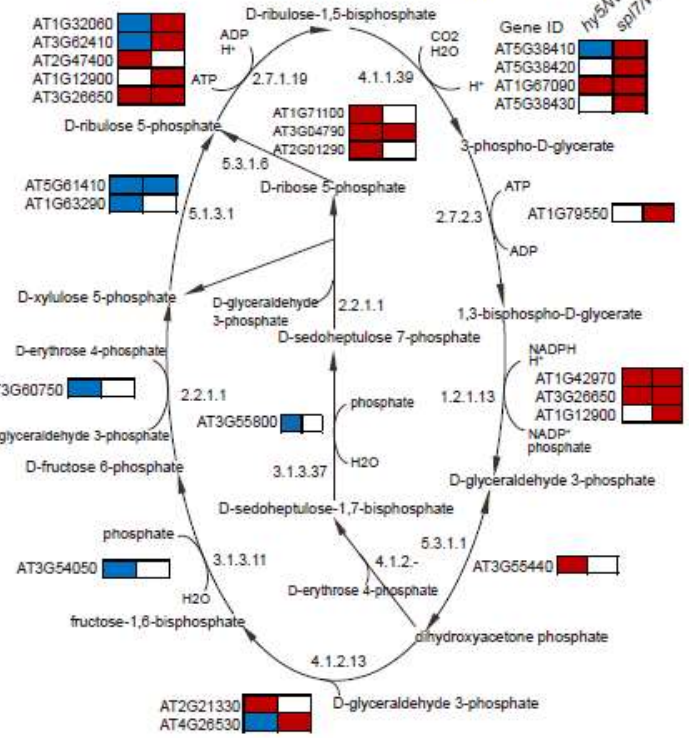
A Photosynthesis Light Reactions



B Photorespiration



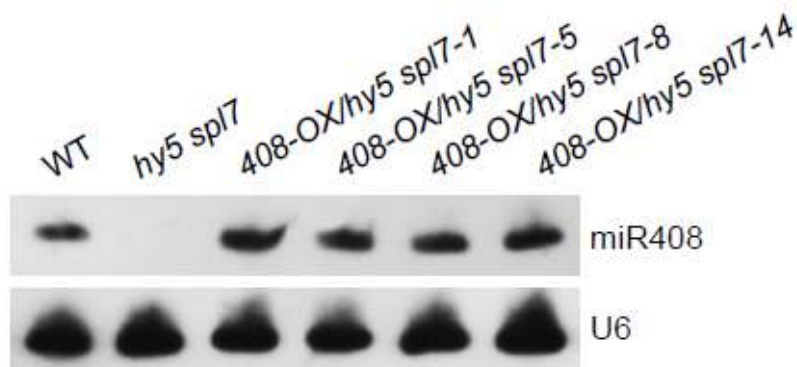
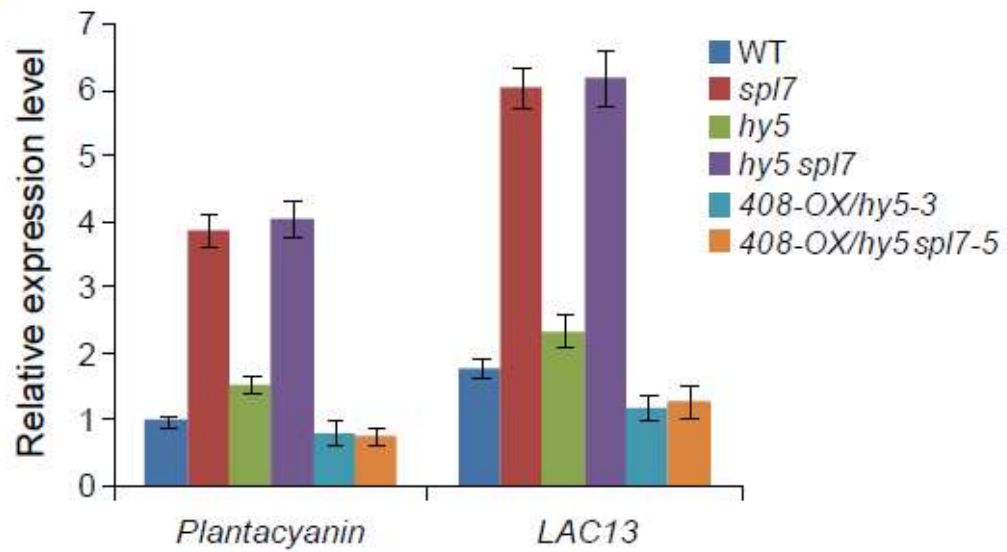
C Calvin-Benson-Bassham cycle



Supplemental Figure 9. Functional Analysis of *MIR408*.

- (A) Expression of miR408 is constitutively activated in *hy5* (top panel) and *hy5 spl7* (bottom panel) by introducing the *35S:pre-MIR408* transgene in the corresponding genetic backgrounds. Over accumulation of miR408 in multiple independent transgenic lines is validated by Northern blotting. U6 was used as a loading control.
- (B) qRT-PCR analysis of two miR408 target genes, *Plantacyanin* and *LAC13*, in various mutants and representative transgenic lines over-expressing *MIR408* grown in the HL/LC condition. Level of *Plantacyanin* in wild type was set as one. Data are means \pm SD from three biological replicates.

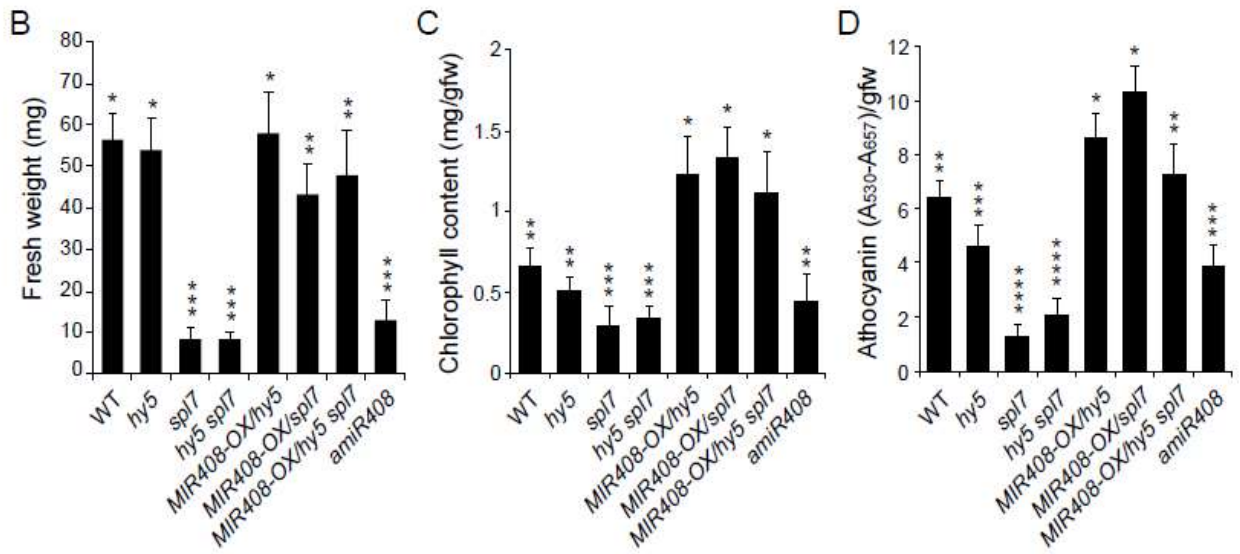
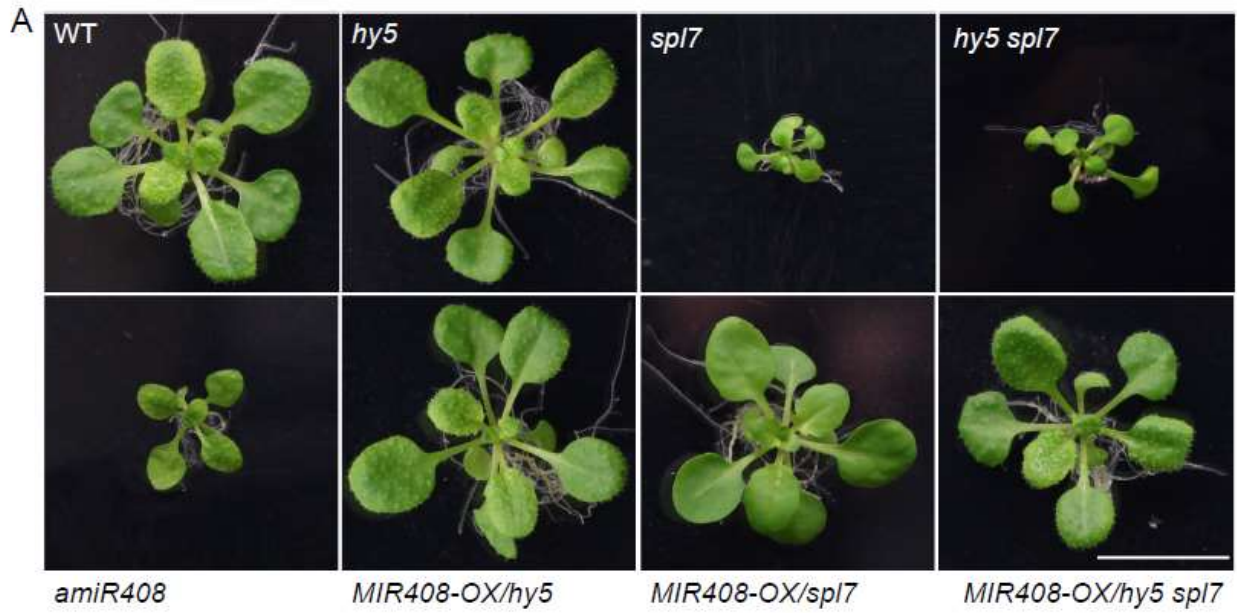
(This figure was generated by Dr. Huiyong Zhang)

A**B**

Supplemental Figure 10. Growth Phenotypes of Adult Plants with Altered *MIR408* Levels.

- (A) Morphology of adult plants with genotypes as indicated in which *MIR408* has varied expression levels. Plants were allowed to grow on MS medium containing 1% sucrose under the Light:Dark = 16h:8h condition for three weeks before photographed. Note that the *spl7* mutant contains the *gll* mutation that inhibits trichome formation and gives the plants glabrous appearance. This recessive mutation was allowed to segregate freely from *spl7* in genetic crosses to create the *hy5 spl7* double mutant. Bar = 1 cm.
- (B) Quantitative measurement of fresh weight (n = 7), (C) chlorophyll (n = 3), and (D) anthocyanin (n = 3) of adult plants with the indicated genotypes. Data are means \pm SD. Genotypes labeled with different numbers of the star sign are statistically different ($p < 0.01$) by ANOVA test.

(This figure was generated by Dr. Huiyong Zhang)

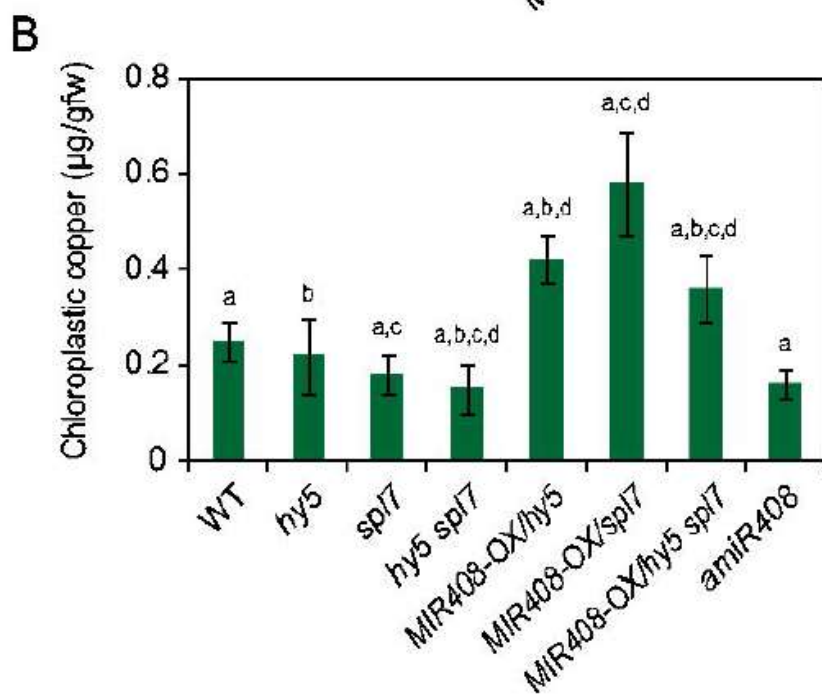
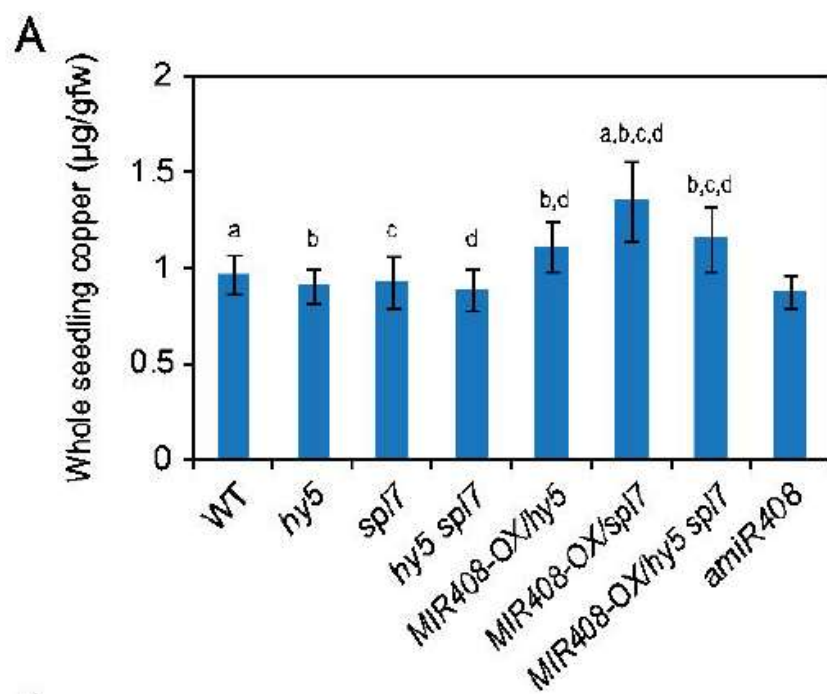


Supplemental Figure 11. Copper Content in Whole Seedlings and the Chloroplast Fractions of Various Genotypes with Altered *MIR408* Expression.

(A) Seedlings of the indicated genotypes were grown on MS media supplemented with 1% sucrose under the LC/HL condition for ten days, weighted, desiccated, digested in 1% nitric acid, and used directly for measuring copper content by inductively coupled plasma-atomic emission spectroscopy.

(B) Intact chloroplasts from seedlings of the indicated genotypes were prepared and measured for copper content as well. Values indicate amounts of copper expressed as $\mu\text{g}\cdot\text{g}^{-1}$ fresh weight. Data are means \pm SD of four biological replicates. Letters above the columns indicate statistically significant groups (ANOVA test; $p < 0.01$). Samples labeled with the same letter are significantly different.

(This figure was generated by Dr. Huiyong Zhang)



Supplementary Tables

Table S1. Summerization of ChIP-sequencing and RNA-sequencing data.

	Sample	Total reads	Unique mapped reads
ChIP-sequencing	<i>FLAG-SPL7</i>	17,790,335	12,280,165 (69%)
	<i>spl7</i>	14,793,925	9,657,327 (62%)
RNA-sequencing	WT-LC	39,493,611	29,818,770 (75%)
	WT-HC	38,883,416	29,522,234 (76%)
	<i>spl7</i> -LC	43,828,358	33,304,095 (76%)
	<i>spl7</i> -HC	36,451,770	27,467,445 (75%)

FLAG-SPL7, 35S:*FLAG-SPL7* transgene; *spl7*, *spl7* mutant; WT, wild type; LC, Low copper concentration (MS); HC, high copper concentration (MS plus 5 μ M CuSO₄).

Supplemental Table 2. List of Genes for ChIP-sequencing and RNA-sequencing Data.

This supplemental table includes 8 sheets and more than 12,000 lines, is available from the author upon request

Supplemental Table 3. Oligonucleotide sequences for the primers and probes used in this study

Oligo name	Oligo sequence (5' to 3')
For quantitative RT-PCR	
SPL7-F	GAGCTGGAGGGCTATATCCG
SPL7-R	GGAAGAGGCTCGATGACTGT
Plantacyanin-F	GAGGCAGTGCATCATGGTCG
Plantacyanin-R	GAGGTCCGTTTGAATCTTCCA
LAC12-F	TCGGCTTCATTGATTATCGCCAAAG
LAC12-R	TTGTGGTGGCACGCTCACAGCTC
LAC13-F	TTCACCTGTCAATGCAGAAGTTAC
LAC13-R	TCTCATTATCCGCCCTCCGCTCTC
HY5-F	CCATCAAGCAGCGAGAGGTCATCAA
HY5-R	CGCCGATCCAGATTCTCTACCGGAA
pre-MIR408-F	TGCAATGAAAGAAGACAAAGCG
pre-MIR408-R	GAGAGGTAGACCAAACCCAAAAAC
Atactin7-F	GGTGTCTATGGTTGGTATGGGTC
Atactin7-R	CCTCTGTGAGTAGAACTGGGTGC
5SRNA-F	GATGCGATCATACCAGCACTAA
5SRNA-R	GATGCAACACGAGGACTTCCC
For plasmid construction	
AD-SPL7-R	CGGCTCGAGTCAAATTTTGTGTACCAATCTCA
FLAG-SPL7-F	AGTCTAGATGGACTACAAGGACGACGATGACAAATCTTCTCTGTGCGCAATCG
FLAG-SPL7-R	TAACTAGTTCAAATTTTGTGTACCAATCTCATTG
p408mut-BF	GGGTCTACCTCGAGGCAGCTAAATTATTTCT
p408mut-BR	TTAGCTGCCTCGAGGTAGGACCCAAAGTACAT
For EMSA	
Probe G-box-R	GGGTCCTACCACGTGGCAGCTAAA
Probe G-box-F	TTTAGCTGCCACGTGGTAGGACCC
Mutated probe G-box-F	GGGTCCTACCTGCAGGCAGCTAAA
Mutated probe G-box-R	TTTAGCTGCCTGCAGGTAGGACCC
probe Vi-F	AAAGTGTACTTTCGATGTACTTTGGGTCCTACCACGTGGCAGC
probe Vi-R	GCTGCCACGTGGTAGGACCCAAAGTACATCGCAAGTACACTTT
SPL7-pF	TTAGTAAGCACATGGTGGATG
SPL7-pR	GAGCTATGTAGGAGGAAGTG
MEME-motif-F	TCTTCTTCTCCTTCTCCTC
MEME-motif-R	GAGGAAGGAGAAGAAGA

For genotyping

T- spl7-F	TTGGAAATTC AAGCTGATTCCG
T- spl7-R	TCCACCTGTCAAACCAAGAC
LBb1	GCGTGGACCGCTTGCTGCAACT
HY5-GF	GTCATCAAGCTCTGCTCCACAT
HY5-GR	AAGACACCTCTTCAGCCGCTTG

For ChIP-qPCR

miR408pF	TGCAACCAATACTGAACCAATCCAA
miR408pR	AGTCTTGACTGCGATCTGGCTAA
SPL7pF	ATAGTAACTTAGTAAGCACATGGTG
SPL7pR	GAGGAGGTGTTGAGCTATGTAG
LAC12pF	TGATTTGTGTATTTGGTAAGACAGGA
LAC12pR	AGAAGTTGAGACATTGAAAGGGAGC
LAC13pF	TGTCTTTGAAGCACTCAAAGTCAC
LAC13pR	CCTACCAGCCTACTTTAGTACCT
At1g58100-FP	ACAGTGGTATGTTCCAATTCTCATC
At1g58100-RP	ATATTATTTTTAGTTTGTAGTCGGTGC
At1g56380-FP	CATGTTTAGTGATTTTTCTTCTTGCT
At1g56380-RP	AAATCGATTTTATCTCTTCTTTCTGTT
At1g27450-FP	CCAAAACATCACCAGCACTTCCT
At1g27450-RP	TGAAAATCCACTCTTGCCTCTCC
At2g45060-FP	GAGCATCGGTTTCATTGCTTTC
At2g45060-RP	AGTCACTTAGGGTGGCATTGTCTT
At3g03780-FP	ACAACACACACCAACTCTCCTCC
At3g03780-RP	GAGAATTATACGTATAACGACGGTTTAC
At4g36040-FP	TGCTTATTATATGGAGGGGAGGAGG
At4g36040-RP	ATAGAGATCAAATCTGAAATGACGAGT
At5g03495-FP	TCTTCATTTGTCTCTGTTTCTCTG
At5g03495-RP	GAGCCTAGTGTATATCCCTCCA
At5g60740-FP	GCAAATCCTTATCAAAAAGCACTCTC
At5g60740-RP	TTCTTCGTCTCTCCTCTGCCATC
At5g03552-FP	ATGAGCATATTGGGATCTTAAATAGC
At5g03552-RP	TCTAGATACAGTTCGTTTCGTGACAG
At5g46845-FP	CACCATTGTCTTTTTCTATTATGTGCT
At5g46845-RP	TATCTTTATACGTGGATCTTGGCTTG
At2g01120-FP	AGAGACTCCGGCGGAGAAATCC
At2g01120-RP	ATACTGAATCAGGAAACTGCTCC
At5g48840-FP	TACATTAGTTTGCTAGTTAATTC
At5g48840-RP	CCAGAAAACGAGTGAGCCATAAAG
At3g21060-FP	AACCATGAGAGGTTCTCGCAC
At3g21060-RP	GAAGGCGTGAGAACGAAATCG
At4g25100-FP	TACCGGATTGGCTGATCCACT
At4g25100-RP	GAGAACTCGAACATGAACTAAG
At4g21270-FP	CGTGTAAATCCACCCATCGAAG
At4g21270-RP	ACTTCTGGAATTATCATAACG
At5g14565-RP	GGCATCTTTGGAACACTTCCTAG
At5g14565-FP	CACAACAAATGATGAAAGGATTGTG