Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Marko Bajic Date

Establishment and Utilization of INTACT-ATAC-seq to Map Cell Type-Specific Gene Regulatory Networks in Plants.

By
Marko Bajic
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

Roger B. Deal, Ph.D.
Advisor

Tamara Caspary, Ph.D.
Committee Member

David J. Katz, Ph.D.
Committee Member

William G. Kelly, Ph.D.
Committee Member

Paula M. Vertino, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Establishment and Utilization of INTACT-ATAC-seq to Map Cell Type-Specific Gene Regulatory Networks in Plants.

By

Marko Bajic
B.S. University of Tennessee at Chattanooga, 2010

Advisor: Roger B. Deal, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfilment of the requirements for the degree of
Doctor of Philosophy
In Genetics and Molecular Biology
2019

Abstract

Establishment and Utilization of INTACT-ATAC-seq to Map Cell Type-Specific Gene Regulatory Networks in Plants.

By Marko Bajic

Differential transcription of protein-coding genes is the basis for cellular diversity and responses to environmental conditions. The precise control of timing and to what extent specific protein-coding genes are transcribed is regulated by transcription factors and chromatin organization. Nucleosomes, the fundamental units of chromatin compaction, can impede the ability of transcription factors (TFs) to bind to DNA. Chromatin regions where TFs bind to DNA are typically depleted of nucleosomes, either because TFs opportunistically bind to nucleosome-free DNA and reconfigure chromatin or because TFs bind nucleosomal DNA leading to nucleosomal repositioning or eviction. By virtue of this nucleosome depletion, TF binding sites can be identified by their increased sensitivity to nuclease cleavage or chemical modifications compared to the rest of the genome. The Assay for Transposase-Accessible Chromatin followed by high-throughput sequencing (ATAC-seq) utilizes a hyperactive Tn5 transposase to cleave DNA at these accessible sites and insert preloaded sequencing adapters into the cleaved DNA. The DNA fragments sequenced in ATAC-seq represent a genome-wide chromatin accessibility profile for the nuclei or cells that are assayed. By using nuclei isolated through Isolation of Nuclei TAgged in specific Cell Types (INTACT) for ATAC-seq, together referred to as INTACT-ATAC-seq, we demonstrated the ATAC-seq technique in *Arabidopsis thaliana*, which had not been done before. Additionally, we expanded INTACT-ATAC-seq for use in four other plant species, in 7 specific cell types, and during response to submergence stress. We found that ATAC-seq can be performed in plants with as little as 2000 nuclei as the starting material. In the diverse plant species examined, accessible chromatin sites were found predominantly within the 3 kb region upstream of the transcription start site, indicating the predominant compartmentalization of regulatory elements to those regions in plants, in contrast to their wide distribution in animal genomes. By coupling chromatin accessibility and transcriptome data, we built gene regulatory networks (GRNs) for two differentiated cell types of the *Arabidopsis* root epidermis, *Arabidopsis* shoot stem cells and multiple differentiating leaf cell types, as well as in the root tips of four diverse species under control conditions and during response to submergence stress. Of particular note, we identified 68 Submergence-UpRegulated gene Families (SURFs) that were upregulated during submergence stress in all four species analyzed, and we found that the same set of four TF families regulate these genes in all species. Interestingly, even though the four TFs and the 68 SURFs are all active in each of the plants during submergence, the connectivity between them varies among the plant species and is indicative of evolutionary adaption that allows rice to survive temporary submergence while the dryland-adapted tomato dies when flooded. In summary, I adapted ATAC-seq to multiple plant species, comprehensively profiled chromatin accessibility and transcription in cells of different plants and environmental conditions to build GRNs that connect chromatin accessibility with gene expression. These findings expand the research toolkit as well as our view of how chromatin is organized and gene regulation is maintained in plants, which is of biological significance for development and environmental stress response.

Establishment and Utilization of INTACT-ATAC-seq to Map Cell Type-Specific Gene Regulatory Networks in Plants.

By

Marko Bajic
B.S. University of Tennessee at Chattanooga, 2010

Advisor: Roger B. Deal, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfilment of the requirements for the degree of
Doctor of Philosophy
In Genetics and Molecular Biology
2019

Acknowledgements

This work is dedicated to my mother and father who fought, endured, and persevered through chaos and the hardships of starting over in a new country all so I could have a chance at a peaceful life. The effort contained within would not have been possible without the support, patience, and companionship of my best friend and partner, Christina Renaud Brosius. I would also like to thank the following people for all their help and support during graduate school.

To all the members of the Deal lab, past or present, I could not have asked for a better family to develop my scientific training with. Kelsey, Paja, and Dylan, it was a pleasure to work with you and to publish together. Ellen, we'll get that stomatal lineage work finished up. Shannon, you were one of the founding members of the Deal lab and I simply could not have figured out so many things without you. Roger, I cannot express how grateful I am to you for giving me a chance to do research in your lab. Thank you for being a caring, approachable, and supportive mentor to me. I could not have asked for a better person to spend time and do research with. Everything I have learned and presented here is only possible thanks to your guidance.

I also want to acknowledge and thank everyone in the plant plasticity team that I collaborated with. Mauricio, you mentored me through some really difficult parts of our data analysis and writing, and I am grateful for your patience and clarity. Kaisa, you are the role model for what a person in the field of science should be. I will always strive to be as helpful and brilliant as the two of you were to me. Donnelly, you helped me figure out how to keep up with all the plasticity work. Julia, Siobhan, and Neelima, thank you for guiding me and for believing in me as I worked on this excellent project with you.

Finally, I want to thank everyone in my graduate school cohort. Annie, Kelly, Kristie, Liz, and Alyx, you were a wonderful group to go through grad school with. Thank you for making the experience a more pleasant one, especially when I felt like I was the only one confused and struggling.

Table of Contents

CHAPTER 1: INTRODUCTION 1

Studying chromatin accessibility and TF binding in specific cell types 3

Scope of the dissertation 8

Figures 14

Literature Cited 18

CHAPTER 2: IDENTIFICATION OF OPEN CHROMATIN REGIONS IN PLANT GENOMES USING ATAC-SEQ 28

Abstract 28

Introduction 29

Materials 30

Methods 34

Notes 40

Acknowledgements 44

Tables and Figures 45

Literature Cited 55

CHAPTER 3: PROFILING OF ACCESSIBLE CHROMATIN REGIONS ACROSS MULTIPLE PLANT SPECIES AND CELL TYPES REVEALS COMMON GENE REGULATORY PRINCIPLES AND NEW CONTROL MODULES 56

Abstract 56

Introduction 57

Results and Discussion 60

Summary and Conclusions 81

Methods 84

Acknowledgements 91

Author Contributions 91

Figures 92

Literature Cited 103

CHAPTER 4: CHROMATIN ACCESSIBILITY CHANGES BETWEEN *ARABIDOPSIS* STEM CELLS AND MESOPHYLL CELLS ILLUMINATE CELL TYPE-SPECIFIC TRANSCRIPTION FACTOR NETWORKS 115

Summary 115

Introduction 116

Results 118

Discussion 128

Methods 132

Acknowledgements 139

Figures 141

Literature Cited 150

CHAPTER 5: EVOLUTIONARY FLEXIBILITY IN FLOODING RESPONSE CIRCUITRY IN ANGIOSPERMS 159

Abstract 159

Main Text 160

Acknowledgements 165

Materials and Methods 166

Figures 180

Literature Cited 187

CHAPTER 6: CHROMATIN ACCESSIBILITY CHANGES IN DIFFERENTIATING CELLS OF THE LEAF 195

Summary 195

Introduction 196

Results 198

Discussion 203

Methods 204

Tables and Figures 210

Literature Cited 232

CHAPTER 7: DISCUSSION – IMPLICATIONS AND FUTURE DIRECTIONS 237

Gene Regulatory Networks connect transcription factors and regulated gene targets 238

Technical improvements for INTACT-ATAC-seq 242

Final statements 244

Figures 246

Literature Cited 249

CHAPTER 1: INTRODUCTION

Chromatin organization and its role in regulating gene expression

Differential transcription of genes in distinct cell types or environmental conditions is the basis for morphological complexity and environmental adaptation of multicellular eukaryotes (Hajheidari et al. 2019). Chromatin organization within the nucleus regulates the extent to which genes are transcribed by RNA Polymerase II (Pol II) (Li et al. 2007). The nucleosome is the fundamental unit of chromatin in which about 145 bases of DNA are wrapped around a histone octamer, made up of an H3/H4 tetramer and two associated H2A/H2B dimers(Luger et al. 1997). Nucleosomes impede the ability of transcription factors (TFs) to bind to the DNA wrapped around the histone octamer (Bai et al. 2010). TF-binding sites are found on cis-regulatory elements (CREs) within promoters and enhancers. Promoters are short DNA regions around the transcription start site (TSS), and enhancers can be either distal or proximal sequences that physically interact with promoters (Figure 1.1A) (Ong et al. 2011, Wittkopp et al. 2011). TF-binding sequences that are wrapped by nucleosomes become temporarily accessible to TFs during brief moments when nucleosomes are partially or completely disrupted (Paranjape et al. 1994, Magnani et al. 2011, Zaret et al. 2011, Iwafuchi-Doi et al. 2014). TFs are typically associated with chromatin remodeling complexes that reconfigure the chromatin to allow for further TF binding. Reconfiguration of nucleosomes, particularly within promoters of genes that become activated, involves removal of promoter bound nucleosomes and a shifting of the +1 nucleosome into the gene to expose the TATA box and the transcription start site (Boeger et al. 2003). Transcription factors that are bound to the DNA sequences in a promoter also help recruit transcriptional machinery and RNA Pol II for stable transcription (Lee et al. 2000). Therefore, transcription is regulated by the availability and activity of transcription factors as well as the accessibility of regulatory elements that are distal or present in or around the promoter (Zou et al. 2011).

Chromatin penetration by Transcription Factors

The ability of TFs to bind their target DNA sequences is regulated by the chromatin structure around the binding site. Specific post translational modifications of histone tails, presence of histone variants, and the location and stability of nucleosomes all influence the binding of transcription factors as well as the initiation and elongation of transcription (Li et al. 2007). Exactly how a TF penetrates chromatin to bind to its DNA sequence is still not clearly understood, but at least four models have been proposed for how this process occurs (Figure 1.1B-E) (Voss et al. 2014). TF-binding sites may become accessible during nucleosome sliding, nucleosome displacement, or replacement of histone subunits with histone variants (Figure 1.1B) (Cairns 2007, Clapier et al. 2009, Korber et al. 2010, Glatt et al. 2011, Hargreaves et al. 2011). Alternatively, local nucleosome structure can be dislodged or reorganized by two TFs that cooperatively are able to overcome the many histone-DNA contacts in the nucleosome (Figure 1.1C) (Miller et al. 2003). For certain transcription factors, their structural similarity with histone H1 may allow them to wedge between DNA-histone contacts and recruit other TFs and the transcriptional machinery to that region (Figure 1.1D) (Zaret et al. 2011). Alternatively, the binding sites may become accessible during chromatin remodeling by chromatin remodeling complexes (CRCs) (Figure 1.1E) (Voss et al. 2011).

TF binding influences transcription

Whether transcription factors bind their target sequences from indirect processes, such as nucleosome turnover or chromatin remodeling, or from targeted approaches such as coordinated dislodging of DNA from histones by two TFs, they can have profound effects on chromatin and transcription upon binding. Some of these transcription factors positively regulate transcription and are referred to as activators (Li et al. 2007). When activators bind to regulatory elements they can promote binding of general transcription factors (GTs) and have other positive effects on transcription initiation. Additionally, the activators can recruit coactivator complexes, such as chromatin-remodeling complexes and histone-modifying enzymes (Li et al. 2007). The modified histones often decrease the barrier to transcription directly or through recruitment of other factors (Pokholok et al. 2005).

Studying chromatin accessibility and TF binding in specific cell types

Techniques for profiling chromatin accessibility

Chromatin regions that are bound by transcription factors are also more sensitive to nuclease cleavage or chemical modifications (Gross et al. 1988). One technique for detecting these accessible sites utilizes DNase I to selectively cleave within nucleosome-depleted DNA regions around TF binding sites. Isolated nuclei are used as the starting material for the DNase I treatment. This method originally utilized Southern blots and indirect end-labeling to identify DNase I hypersensitive sites at loci of interest, but now the mapping of hypersensitive sites can be done genome wide using high-throughput sequencing of DNase I cleavage fragments (Figure 1.2A). This process of treating DNA with DNase I followed by high throughput sequencing is called DNase-seq and can be used to identify all types of regulatory elements, such as promoters, enhancers, silencers, insulators, and locus control regions (Song et al. 2010). The characterization of DNase Hypersensitive Sites (DHSs) has been done in many human cells and tissues (Boyle et al. 2008, Boyle et al. 2011, John et al. 2011, Song et al. 2011, Thurman et al. 2012), several other model animals (Gerstein et al. 2010, Roy et al. 2010, Stamatoyannopoulos et al. 2012, Quillien et al. 2017), as well as two plant model organisms, *Arabidopsis thaliana* and rice, *Oryza sativa* (Zhang et al. 2012, Zhang et al. 2012). Interestingly, DNase-seq data comparisons between plants and humans revealed major differences in the genomic localization of DHSs between plants and animals. There were significantly more DHSs present in the introns of humans compared to plants, which is most likely caused by the much larger average intron size in humans compared to plants (Zhang et al. 2014). Additionally, these findings also revealed the simpler organization of CREs in *Arabidopsis*, with the majority of its DHSs located in the promoter regions and fewer DHSs located in introns and intergenic space compared to humans (Zhang et al. 2014).

While DNase-seq is able to identify CREs genome-wide, the actual methodology for performing the DNase I treatment and amplifying libraries for high-throughput sequencing is labor-intensive, and a large amount of starting nuclei are needed for DNase I treatment. A newer technique called Assay for

Transposase-Accessible Chromatin followed by high-throughput sequencing (ATAC-seq) was developed as a simpler alternative to DNase-seq (Buenrostro et al. 2013). This technique utilizes a hyperactive Tn5 transposase to simultaneously cleave DNA and insert preloaded sequencing adapters into the cleaved DNA (Figure 1.2 B). This process is referred to as 'tagmentation', and the cleaved fragments that are the end product can be amplified into a high-throughput sequencing library by Polymerase Chain Reaction (PCR). The regions of high signal relative to the background correspond to places where frequent tagmentation occurred, which is reflective of chromatin sites of high accessibility. ATAC-seq data are highly similar to those of DNase-seq, both in the location and the amount of accessibility detected throughout the genome (Buenrostro et al. 2015). However, the benefits of using ATAC-seq are not only that it is a quicker and simpler protocol compared to DNase-seq, but it also requires far fewer starting nuclei. The combined simplicity of protocol and low starting material requirement have made ATAC-seq the ideal methodology for studying chromatin accessibility, even compared to other techniques such as FAIRE-seq, NOMe-seq, and NicE-seq (Giresi et al. 2007, Boyle et al. 2008, Song et al. 2010, Ponnaluri et al. 2017, Shashikant et al. 2019).

Isolation of nuclei from specific cell types

ATAC-seq has been used to profile chromatin accessibility in *C. elegans* (Daugherty et al. 2017, Janes et al. 2018), the human malaria parasite *Plasmodium falciparum* (Ruiz et al. 2018), *Xenopus* (Bright et al. 2019), black rockfish (He et al. 2019), *Drosophila* blastoderm (Bozek et al. 2019), several cell types in mouse (Hilliard et al. 2019, Liu et al. 2019, Wilkerson et al. 2019, Yoshida et al. 2019), and several cell types and species of plants (Lu et al. 2017, Bajic et al. 2018, Maher et al. 2018). Tagmentation on pure nuclei is imperative because organellar genomes, such as mitochondrial or chloroplast genomes, can also be tagmented and end up dominating the sequencing library if present. Using organelle-free nuclei as the starting material for ATAC-seq thus greatly improves the cost per read that is returned after sequencing. Additionally, isolation of nuclei from specific cell types is essential for studying development and response to environmental stimuli. Finally, isolation of nuclei from specific cell types is more informative

of transcriptional regulation by CREs because there is less heterogeneous mixing of different regulatory networks that arises from studying transcriptional regulation in a pool of many different cell types.

There are primarily three different technologies for the isolation of nuclei from individual cell types: Laser Microdissection (LM), Fluorescence-Activated Cell Sorting (FACS), and Isolation of Nuclei TAgged in specific Cell Types (INTACT; Figure 1.2C-E) (Wang et al. 2012). While whole cells are isolated using LM and FACS, INTACT specifically purifies nuclei from specific cell types, which minimizes organellar DNA contamination. For laser microdissection, the tissue is chemically fixed, embedded in a solid material, sliced into thin sections, and then the cell type of interest is excised using microdissection with a laser (Figure 1.2C) (DeCarlo et al. 2011). In FACS, the cell type of interest is fluorescently labeled, either by promoter-driven expression of a fluorescent protein or by treating cells with a fluorescent antibody, and cells are streamed past a laser that applies a charge to fluorescently labeled cells, collecting them in different receptacle from the unlabeled cells (Figure 1.2D) (Harkins et al. 1990, Herzenberg et al. 2002). INTACT utilizes a cell-type-specific promoter to drive the expression of a nuclear targeting fusion (NTF) protein made up of a nuclear envelope-associating domain, green fluorescent protein, and a biotin ligase recognition peptide, which can be biotinylated by the *E. coli* biotin ligase, BirA. BirA is expressed constitutively, allowing for isolation of tagged nuclei from a specific cell type using streptavidin-coated magnetic beads (Figure 1.2E) (Deal et al. 2011). INTACT requires the establishment of transgenic lines that express the transgenes that direct cell type-specific targeting of NTF to nuclei, but the isolation of nuclei from these lines is significantly faster and easier compared to LM and FACS. Additionally, as more cell type-specific promoters are identified over time the more cell type-specific nuclei can be isolated and categorized for chromatin profiling and transcriptional responses using INTACT. Since its initial publication for isolating nuclei in *Arabidopsis thaliana* (Deal et al. 2010), INTACT has been used to isolate nuclei from specific cell types in flies (Henry et al. 2012), *Xenopus* (Wasson et al. 2019), mouse (Amin et al. 2014, Bhattacharyya et al. 2019), as well as additional cell types of *Arabidopsis* and other plant species (Ron et al. 2014, Park et al. 2016, Moreno-Romero et al. 2017, Reynoso et al. 2018). By using the nuclei isolated through INTACT as the starting material for ATAC-

seq, together referred to as INTACT-ATAC-seq, genome-wide chromatin accessibility libraries for a specific cell type can be generated from whole tissue starting material in less than a day.

Gene regulatory networks: connecting accessible chromatin regions to regulated genes

Isolation of nuclei from specific cell types yields nuclear DNA and RNA molecules that can be assayed for chromatin accessibility and the transcriptome using ATAC-seq and RNA-seq, respectively. Identification of specific accessible chromatin sites and precise levels of transcripts are greatly improved by analyzing specific cell types compared to the heterogeneous mixing of multiple cell types that averages out multiple networks of regulation. This is especially true during development, where cis-regulatory elements are accessed by specific transcription factors to confer precise control of transcription (Lenhard et al. 2012). An overall schematic of transcription factors binding to DNA and the genes whose expression is regulated by those TFs is referred to as a Gene Regulatory Network (GRN) (Lowe et al. 2019). Within a GRN, each node represents a gene (responder) or a TF (director) and the two are connected by lines or arrows denoting which genes are regulated by which TFs. Because TFs are proteins encoded by DNA, intricate levels of regulation can be depicted where a TF can regulate the level of expression of its own or another TF's transcription, creating self-looping or cross-cluster connections within a GRN.

Composition of a GRN is reliant on proper identification of which genomic sites are occupied by which TFs, what the transcript level is for the target genes, and whether the DNA-bound TF actually regulates the expression of its target gene (Spitz et al. 2012). Identifying where TFs bind DNA throughout the genome can be done using ChIP-seq if antibodies for specific TFs exist, but ATAC-seq offers a more comprehensive approach, defining all TF binding sites simultaneously, that is cheaper and requires less starting material. Expression of target genes can be measured using RNA-seq, but it is most informative if the starting material for the construction of the RNA-seq libraries is the same as that used to assay chromatin accessibility. The nuclei isolated using INTACT can be used to identify multiple potential TF-

binding sites using ATAC-seq as well as the expression of their target genes using RNA-seq, from the same starting material.

Identifying multiple potential TF-binding sites using ATAC-seq is done by first identifying transposase hypersensitive sites (THSs) specific to the cell type or condition being studied. THSs represent genomic regions where chromatin is highly accessible and TFs can bind their target DNA sequence, or motif. These motif sequences can be identified as potentially active TF-binding sites if they are found to be overrepresented in the THS sequences being analyzed. If experimentally-derived prior information exists for the motif sequences bound by TFs of interest, represented in the form of positional weight matrices (PWMs), then it is possible to infer the potential identity and location of the different TFs that are binding DNA within the accessible sites analyzed.

Knowing the identity and locations of active TFs is not enough to construct a GRN because the connection between the TF and its target gene may not be known. While this is straightforward when the TF-binding sequence is within a promoter of the target gene, it is harder to connect TF-binding sequences found in distal enhancers to target genes without chromatin conformation data, such as Hi-C contact maps that specifically identify chromatin sites that are proximal to each other within the nucleus. The difficulty of connecting enhancers with their target genes arises from the fact that the enhancers can be significantly distant from their target genes, in either the upstream or downstream direction (Ong et al. 2011).

Fortunately, the chromatin organization in *Arabidopsis thaliana* is relatively straightforward with majority of the CREs compartmentalized into domains that are centralized around genes, with few distal chromatin conformation interactions (Rowley et al. 2017). This makes the study of gene regulation by cis-regulatory elements relatively straightforward in this model organism. Furthermore, *Arabidopsis* has a small genome (120 Mb), is easy to transform, and there are multitudes of publicly available datasets for chromatin immunoprecipitation of different histone marks and transcription factors, as well as RNA-seq experiments in different cell types, mutants, and conditions. Finally, the study of transcriptional regulation in the plant model organism *Arabidopsis thaliana* is important not only because it is an efficient way to study chromatin regulation and its direct role in controlling the expression of the

information contained within DNA, but also because understanding the regulation of development and response to environmental stresses in plants will help us build hardier and more productive crops.

Scope of the Dissertation

By studying chromatin accessibility we can define the regulatory regions of a genome, identify gene regulatory networks (GRNs), and examine changes among cell states and environmental conditions. While there are other technologies that can be used to study chromatin organization and TF binding throughout the genome, ATAC-seq has the distinct advantages of being quick and easy, while requiring low starting material. To that end, we defined the use of ATAC-seq in five different plant species, two of which have limited chromatin sequencing datasets, seven distinct cell types of the *Arabidopsis thaliana* root and shoot tissue, and in the root tips of four species responding to submergence stress, a rising environmental threat to agriculture.

In Chapter 2, I discuss the work done by myself and another graduate student in the Deal lab, Kelsey Maher, to apply the ATAC-seq technique to *Arabidopsis thaliana*, which had not been previously reported. We showed that ATAC-seq libraries, with expected sequenced fragment sizes, could be generated using nuclei isolated from specific cell types of the root or the whole root tip. Additionally, we demonstrated two different ways for generating nuclei that can be used as input for ATAC-seq: crude nuclei isolated using sucrose sedimentation and isolation of nuclei from specific cell types using INTACT. Furthermore, besides providing a reproducible protocol for isolating nuclei in *Arabidopsis thaliana*, performing ATAC-seq, amplifying libraries, and visualizing sequencing results, we also demonstrated the higher signal over background in sequenced libraries that is obtained by performing ATAC-seq in INTACT-isolated nuclei from specific cell types compared to the heterogeneous mix of cell types in the root tip. Overall, Chapter 2 serves as the foundational technical work that established the use of ATAC-seq in the model organism *Arabidopsis thaliana*.

Chapter 3 builds upon the techniques established in Chapter 2, applying INTACT-ATAC-seq in several different plant species, as well as two specific cell types of the *Arabidopsis thaliana* root

epidermis. Kelsey Maher and I performed INTACT-ATAC-seq in *Arabidopsis thaliana* and *Medicago truncatula*, whereas our collaborators performed similar experiments in rice, *Oryza sativa*, and tomato, *Solanum lycopersicum*. Utilizing the same growth conditions and starting material, we were able to compare the chromatin accessibility profiles of four divergent plant species whose last common ancestor existed approximately 123 million years ago. Surprisingly, even though the genomic composition of the four plant species varied greatly, the chromatin accessibility was found predominantly within the 3 kb upstream region of the transcription start site in all species, indicating that regulatory elements are found proximal to genes in all four plants. Our ATAC-seq data also had excellent overlap with accessibility data generated using DNase-seq from similar tissue, confirming the reliability of this technique in the new model organisms in which it was being used. Additionally, we found a common set of four TFs whose binding-sequences were overrepresented in the root tip set of THSs across all species. This core group of TFs regulated similar target genes, representing a root-specific gene regulatory network that is conserved among all four plant species.

This chapter also demonstrated the first analyses of chromatin accessibility profiles in distinct cell types of the root. The two cell types analyzed were the root epidermal hair cells and the root epidermal non-hair cells, both representing terminally differentiated cell types of the root epidermis. The chromatin profiles between the two differentiated cell types appeared highly similar, but specific sites of differential accessibility were identified, and these were often found near genes that had RNA expression profiles unique to one cell type or the other. We used the DNA sequences found in sites more accessible in one cell type or the other to identify TF-binding sequences that are enriched in each cell type. By mapping the motifs identified as enriched in the hair cell and by utilizing RNA expression data for this cell type, we built a GRN for the hair cell. The GRN demonstrated two different sets of genes regulated by two MYB TFs, one of which most likely regulates hair development and differentiation while the other regulates the cell's response to water and phosphate deprivation. Overall, this chapter built upon the groundwork for utilizing INTACT-ATAC-seq in plant tissue established in Chapter 2, it discovered evolutionarily

conserved gene regulatory networks specific to roots, and it pioneered the differential analyses of chromatin accessibility levels in distinct cell types in plants.

In Chapter 4, I discuss work done alongside a postdoc in the Deal lab, Paja Sijacic, where we expanded on the differential analyses of chromatin accessibility changes between distinct cell types established in Chapter 3. We used INTACT-ATAC-seq to assay chromatin accessibility profiles in the stem cells of the shoot apical meristem and the differentiated photosynthetic mesophyll cells of the leaf. For practical purposes, the ATAC-seq datasets generated from this work reflected the beginning and the end point of mesophyll cell differentiation. This work reflects a contrast to a part of the work done in Chapter 3 where two endpoint differentiated cell types were compared to each other. The THSs that were more enriched in one cell type or the other were clearly defined and statistically validated, and they represent the progression of the best methods for ATAC-seq data analyses that were being cultivated in the Deal lab. The differential THSs identified in the two cell types were found predominantly around the TSS of protein-coding genes. These genes were annotated with gene ontology (GO) terms that are specific to either the stem cells or mesophyll cells. There were many overrepresented TF-binding sequence motifs found in the two cell type-specific groups of THSs. We utilized publicly available transcriptome datasets to rank the expression of TFs that bind the overrepresented sequence motifs. The 23 TFs more highly expressed in stem cells and the 129 TFs more highly expressed in mesophyll cells were evaluated using the STRING network (Szklarczyk et al. 2017) of connections curated by publicly available data to identify 4 TFs in each cell type that regulate gene networks specific to functions of that cell type. Besides illuminating gene regulatory networks for each cell type, we also built regulatory pathways for key transcription factors that orchestrate transcriptional regulation in each cell type. Overall, the first in-depth quantitative analyses of chromatin accessibility changes in differentiating cells of the leaf using INTACT-ATAC-seq successfully identified gene regulatory networks composed of TFs known to be important for each cell type and regulated genes whose physiological role is specific to that cell type. However, the reliance on RNA expression data that was not generated from the same starting

material as the ATAC-seq libraries, as well as unfinished validation of TF mutant lines represent limitations of the approach outlined in this chapter.

The culmination of my work on chromatin accessibility profiling and gene regulation control is highlighted in Chapter 5 where I worked alongside collaborators from three other labs to characterize the different levels of gene expression regulation, from chromatin accessibility changes to differential ribosome occupancy on translating mRNAs, in the roots of four different plant species undergoing submergence stress. I was primarily involved with performing the experiments and generating libraries for sequencing in barrel clover, *Medicago truncatula*, as well as doing all the ATAC-seq analyses for all four species. Our collaborators performed similar experiments in rice, *Oryza sativa*, and two tomato species, the domesticated *Solanum lycopersicum* and the dryland-adapted wild relative *Solanum pennellii*. Sequencing results for the different levels of RNA expression regulation identified a conserved set of 68 Submergence-UpRegulated gene Families (SURFs) that are upregulated in response to submergence stress in all species. These submergence-upregulated genes had elevated chromatin accessibility around the TSS and 3' of the polyA sites. By analyzing the DNA sequences of THSs that become more accessible during submergence and are localized within the promoter regions of SURF genes, we identified overrepresented TF-binding sequence motifs in *Medicago truncatula* and *Oryza sativa*. These sequence motifs, due to their localization within regulatory regions of submergence upregulated genes, represent probable regulatory motifs specific to submergence stress. After defining a clear validation pipeline for evaluating motifs that specifically regulate SURF genes, we identified four sequence motifs within SURF promoters in all four species. These included a Hypoxia Responsive Promoter Element (HRPE) that was previously shown to upregulate genes specific to flooding survival in *Arabidopsis thaliana* (Mustroph et al. 2010, Lee et al. 2011, Gasch et al. 2016). The motif occurrences were used to build gene regulatory networks between the four motifs and the SURF genes in all four species. Interestingly, the different preference for which TFs regulate which gene families provided insight into differential regulatory responses among the four species. For example, the higher prevalence of HRPE in rice SURF genes compared to the dryland-adapted tomato may explain rice's ability to better tolerate

temporary flooding. Furthermore, the benefit of having RNA expression and chromatin accessibility data generated from the same starting material, INTACT-isolated nuclei, allowed us to make conclusions about the chromatin state around a regulatory motif and the expression level of the gene nearest to that regulatory motif. It was observed that the accessibility of a motif within a SURF gene was most indicative of increased transcription for that gene, compared to the lesser correlative property of having one or more inaccessible motifs within that SURF gene's promoter region. Overall, the work in this chapter connected the multiple targets of gene expression regulation that are observed at the level of chromatin accessibility changes, nuclear transcripts, total transcripts, and mRNA translation. Furthermore, the careful work done to keep the experiments consistent among the plant species helped identify not only submergence-specific gene regulatory networks, but specifically those that are conserved across at least 123 million years of evolution.

In Chapter 6, I discuss unpublished research that builds upon the chromatin changes that occur during cell differentiation, as discussed in Chapter 4, between the start and end points of the stem cells of the shoot apical meristem and the differentiated mesophyll cells of the leaf. Research presented in this chapter incorporates three additional cell states along the *Arabidopsis* leaf stomatal lineage in addition to the two cell types discussed in Chapter 4. These three cell states are the self-renewing meristemoid cells, the guard mother cells (GMCs) that arise from an asymmetric division of a meristemoid cell, and the guard cells that GMCs give rise to. The ATAC-seq libraries for each of these cell types were generated by a graduate student who rotated in the Deal lab, Liz Dreggors, and I performed data analysis on the sequenced libraries. Utilizing the ATAC-seq pipelines I built throughout my graduate work, I identified reproducible THSs in all five cell types and found cell type-enriched THSs that were found primarily in promoters of genes whose GO terms were reflective of the physiology of the cell type in which the THS was most accessible. Chapter 6 also expands upon my initial attempts to find overrepresented DNA sequence motifs in each cell type-enriched set of THSs, evaluation of these results, as well as future improvements for the discovery of overrepresented motifs. Overall, this chapter lays the groundwork for building gene regulatory networks across a differentiation timeline by following chromatin accessibility

changes between a start point, three sequential time points, as well as an alternate end point. Additionally, this Chapter points to the limitations encountered in trying to build gene regulatory networks using only ATAC-seq data and publicly available RNA expression data that may not match the starting material well.

Throughout the experiments and analyses presented within this thesis I show how INTACT-ATAC-seq was established in plants, as well as how it was applied in different plant species to study evolutionary conservation of gene expression in roots, response to submergence stress, as well as leaf cell differentiation in *Arabidopsis*. In addition to being valuable resources to the scientific community, the datasets generated here produced meaningful results that are important for understanding regulation of gene expression at the chromatin level, and may aid in manipulating crops to have increased tolerance to stress. Future directions, data analysis optimizations, and greater implications for the work presented within this thesis are discussed further in Chapter 7.

FIGURES

Fig 1.1

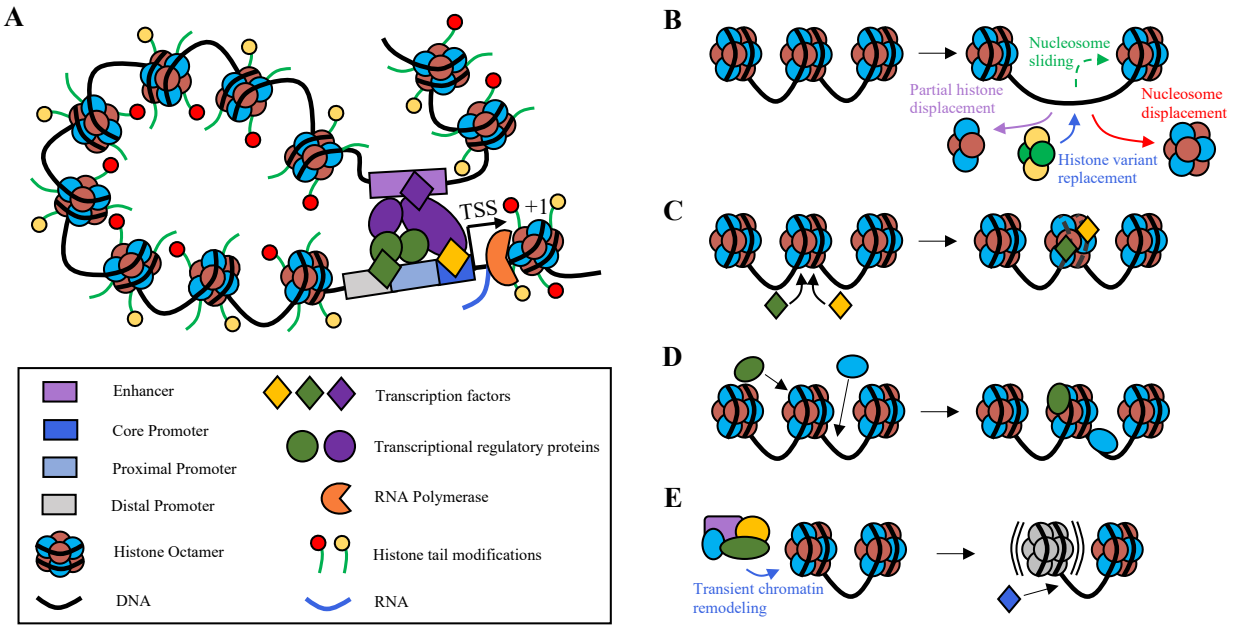


Figure 1.1. Chromatin organization and penetration by pioneer transcription factors. (A) The 3D structure of chromatin is reorganized during transcription. Transcription factors bind to their sequence motifs within cis-regulatory elements such as enhancers and promoters. Additional transcriptional regulatory proteins create a bridge that connects distal enhancers to promoters and regulate the level of transcription. The ability of transcription factors to bind to DNA is influenced by modification of the histone tail, and these marks are further modified after transcriptional machinery is recruited. DNA information encoding proteins is transcribed to RNA molecules by RNA Polymerase II. One of the primary barriers to productive transcription is the +1 nucleosome that is encountered by the RNA Polymerase. **(B-E**) The ability of pioneer transcription factors to overcome the complex structure of chromatin and bind TF-binding sequences is explained through four models. **(B)** Increased accessibility of TF-binding sites occurs during nucleosome sliding (green dashed arrow), partial histone displacement (purple arrow), histone variant replacement (blue arrow), and nucleosome displacement (red arrow). **(C)** Cooperative nucleosome attack by two transcription factors that act in tandem can generate sufficient free

energy to overcome the DNA-histone contacts and bind to the sequence motifs. **(D)** In certain instances, transcription factors have structural similarity to the H1 histone, which allows them to wedge themselves between the DNA-histone contacts. **(E)** Transient remodeling of chromatin by chromatin remodeling complexes leaves brief windows of DNA access wherein a TF can bind to its sequence motif.

Fig 1.2

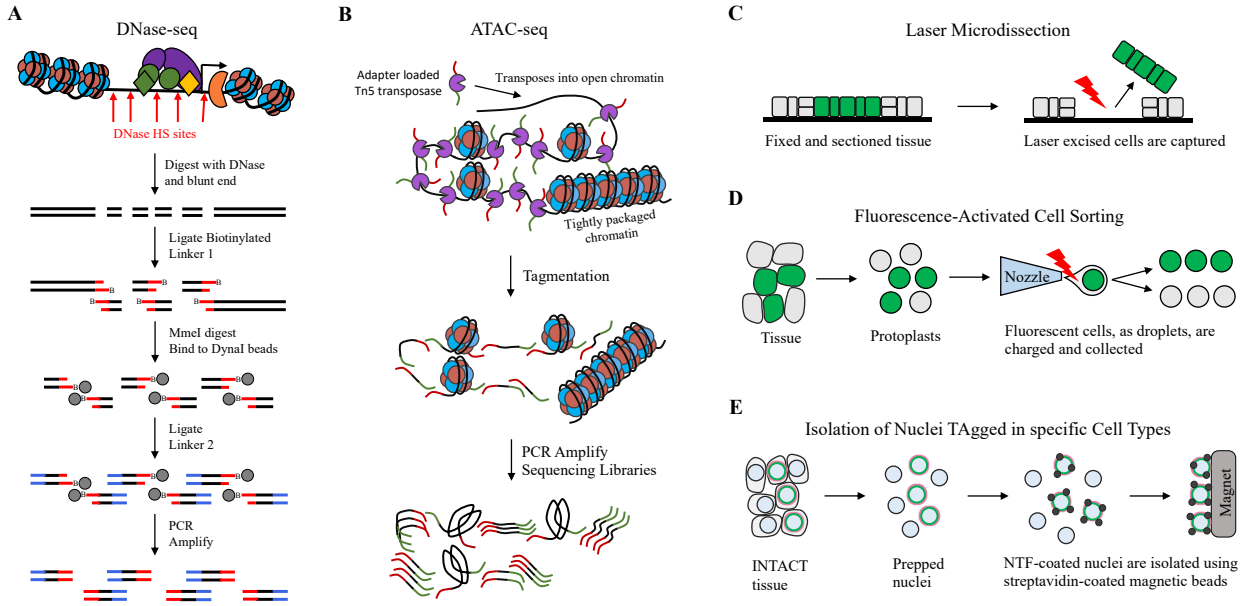


Figure 1.2. Chromatin accessibility profiling in specific cell types. (A) Overview of the DNase-seq protocol. DNA from isolated nuclei or cells is digested with DNase I. Chromatin regions depleted of nucleosomes are preferentially digested and represent DNase hypersensitive (HS) sites (red arrows). The digested DNA is blunt-ended, purified from the reaction, and then ligated to biotinylated linker 1. The biotinylated fragments are digested by MmeI and captured by streptavidin-coated Dynabeads. The 2 base overhangs generated by MmeI are used to ligate linker 2 onto the fragments. Finally, the DNA fragments are PCR amplified and sequenced using next-gen sequencing. **(B)** Overview of the ATAC-seq protocol. DNA from isolated nuclei or cells is treated with sequencing adapter-loaded Tn5 transposase. The sequencing adapters are inserted and the DNA is cut, through a process referred to as tagmentation, at nucleosome depleted regions. The fragmented DNA is amplified using primers that correspond to fractions of the sequencing adapters, and the amplified DNA is sent for next-gen sequencing. **(C)** Overview of the Laser Microdissection (LM) protocol. Sectioned tissue is fixed to a slide and the specific cells of interest, depicted using a green color, are excised using a laser and captured for downstream applications. **(D)** Overview of the Fluorescence-Activated Cell Sorting (FACS) protocol. Cells from the tissue of interest are digested out to isolate protoplasts. The protoplasts are passed through a nozzle that

produces a stream of droplets, with each droplet containing a single protoplast. Protoplasts of fluorescently labeled cells, which are fluorescently labeled either by promoter driven expression of a fluorescent protein or by fluorescent antibody treatment, are charged by a laser and collected in a different receptacle from the unlabeled cells. **(E)** Overview of the Isolation of Nuclei TAgged in specific Cell Types (INTACT) protocol. Nuclei are isolated from the tissue of interest either by grinding frozen tissue or chopping fresh tissue. Cell type-specific nuclei that express the nuclear targeting fusion (NTF) in specific cell types are separated from the rest of the nuclei using streptavidin-coated magnetic beads and a magnet. The streptavidin interacts with the biotinylated biotin ligase recognition peptide of the NTF, which is localized to the outer nuclear membrane.

LITERATURE CITED

Amin, N. M., T. M. Greco, L. M. Kuchenbrod, M. M. Rigney, M. I. Chung, J. B. Wallingford, I. M. Cristea and F. L. Conlon (2014). "Proteomic profiling of cardiac tissue by isolation of nuclei tagged in specific cell types (INTACT)." Development **141**(4): 962-973.

Bai, L. and A. V. Morozov (2010). "Gene regulation by nucleosome positioning." Trends in Genetics **26**(11): 476-483.

Bajic, M., K. A. Maher and R. B. Deal (2018). "Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq." Methods Mol Biol **1675**: 183-201.

Bhattacharyya, S., A. A. Sathe, M. Bhakta, C. Xing and N. V. Munshi (2019). "PAN-INTACT enables direct isolation of lineage-specific nuclei from fibrous tissues." PLoS One **14**(4): e0214677.

Boeger, H., J. Griesenbeck, J. S. Strattan and R. D. Kornberg (2003). "Nucleosomes Unfold Completely at a Transcriptionally Active Promoter." Molecular Cell **11**(6): 1587-1598.

Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey and G. E. Crawford (2008). "High-resolution mapping and characterization of open chromatin across the genome." Cell **132**(2): 311-322.

Boyle, A. P., J. Guinney, G. E. Crawford and T. S. Furey (2008). "F-Seq: a feature density estimator for high-throughput sequence tags." Bioinformatics **24**(21): 2537-2538.

Boyle, A. P., L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey (2011). "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells." Genome Res **21**(3): 456-464.

Bozek, M., R. Cortini, A. E. Storti, U. Unnerstall, U. Gaul and N. Gompel (2019). "ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm." Genome Res **29**(5): 771-783.

Bright, A. R. and G. J. C. Veenstra (2019). "Assay for Transposase-Accessible Chromatin-Sequencing Using Xenopus Embryos." Cold Spring Harb Protoc **2019**(1): pdb prot098327.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang and W. J. Greenleaf (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nat Methods **10**(12): 1213-1218.

Buenrostro, J. D., B. Wu, H. Y. Chang and W. J. Greenleaf (2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." Curr Protoc Mol Biol **109**: 21 29 21-29.

Cairns, B. R. (2007). "Chromatin remodeling: insights and intrigue from single-molecule studies." Nat Struct Mol Biol **14**(11): 989-996.

Clapier, C. R. and B. R. Cairns (2009). "The biology of chromatin remodeling complexes." Annu Rev Biochem **78**: 273-304.

Daugherty, A. C., R. W. Yeo, J. D. Buenrostro, W. J. Greenleaf, A. Kundaje and A. Brunet (2017). "Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans." Genome Res **27**(12): 2096-2107.

Deal, R. B. and S. Henikoff (2010). "The INTACT method for cell type–specific gene expression and chromatin profiling in Arabidopsis thaliana." Nature Protocols **6**: 56.

Deal, R. B. and S. Henikoff (2011). "The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana." Nat Protoc **6**(1): 56-68.

DeCarlo, K., A. Emley, O. E. Dadzie and M. Mahalingam (2011). "Laser Capture Microdissection: Methods and Applications." Methods in Molecular Biology **755**.

Gasch, P., M. Fundinger, J. T. Muller, T. Lee, J. Bailey-Serres and A. Mustroph (2016). "Redundant ERF-VII Transcription Factors Bind to an Evolutionarily Conserved cis-Motif to Regulate Hypoxia-Responsive Gene Expression in Arabidopsis." Plant Cell **28**(1): 160-180.

Gerstein, M. B., Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M. S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dose, J.

Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff, S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D. M. Miller, 3rd, A. Muroyama, J. I. Murray, S. L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, C. Slightam, R. Smith, W. C. Spencer, E. O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N. L. Washington, C. M. Whittle, B. Wu, K. K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, E. C. mod, J. Ahringer, S. Strome, K. C. Gunsalus, G. Micklem, X. S. Liu, V. Reinke, S. K. Kim, L. W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J. D. Lieb and R. H. Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer and J. D. Lieb (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." Genome Res **17**(6): 877-885.

Glatt, S., C. Alfieri and C. W. Muller (2011). "Recognizing and remodeling the nucleosome." Curr Opin Struct Biol **21**(3): 335-341.

Gross, D. S. and W. T. Garrard (1988). "Nuclease Hypersensitive Sites in Chromatin." Ann Rev Biochem **57**: 159-197.

Hajheidari, M., C. Koncz and M. Bucher (2019). "Chromatin Evolution-Key Innovations Underpinning Morphological Complexity." Front Plant Sci **10**: 454.

Hargreaves, D. C. and G. R. Crabtree (2011). "ATP-dependent chromatin remodeling: genetics, genomics and mechanisms." Cell Res **21**(3): 396-420.

Harkins, K. R., R. A. Jefferson, T. A. Kavanagh, M. W. Bevan and D. W. Galbraith (1990). "Expression of photosynthesis-related gene fusions is restricted by cell type in transgenic plants and in transfected protoplasts." Proceedings of the National Academy of Sciences of the United States of America **87**(2): 816-820.

He, Y., Y. Chang, L. Bao, M. Yu, R. Li, J. Niu, G. Fan, W. Song, I. Seim, Y. Qin, X. Li, J. Liu, X. Kong, M. Peng, M. Sun, M. Wang, J. Qu, X. Wang, X. Liu, X. Wu, X. Zhao, X. Wang, Y. Zhang, J. Guo, Y. Liu, K. Liu, Y. Wang, H. Zhang, L. Liu, M. Wang, H. Yu, X. Wang, J. Cheng, Z. Wang, X. Xu, J. Wang, H. Yang, S. M. Lee, X. Liu, Q. Zhang and J. Qi (2019). "A chromosome-level genome of black rockfish, Sebastes schlegelii, provides insights into the evolution of live birth." Mol Ecol Resour.

Henry, G. L., F. P. Davis, S. Picard and S. R. Eddy (2012). "Cell type-specific genomics of Drosophila neurons." Nucleic Acids Res **40**(19): 9691-9704.

Herzenberg, L. A., D. Parks, B. Sahaf, O. Perez, M. Roederer and L. A. Herzenberg (2002). "The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford." Clinical Chemistry **48**(10): 1819.

Hilliard, S., R. Song, H. Liu, C. H. Chen, Y. Li, M. Baddoo, E. Flemington, A. Wanek, J. Kolls, Z. Saifudeen and S. S. El-Dahr (2019). "Defining the dynamic chromatin landscape of mouse nephron progenitors." Biol Open **8**(5).

Iwafuchi-Doi, M. and K. S. Zaret (2014). "Pioneer transcription factors in cell reprogramming." Genes Dev **28**(24): 2679-2692.

Janes, J., Y. Dong, M. Schoof, J. Serizay, A. Appert, C. Cerrato, C. Woodbury, R. Chen, C. Gemma, N. Huang, D. Kissiov, P. Stempor, A. Steward, E. Zeiser, S. Sauer and J. Ahringer (2018). "Chromatin accessibility dynamics across C. elegans development and ageing." Elife **7**.

John, S., P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager and J. A. Stamatoyannopoulos (2011). "Chromatin accessibility pre-determines glucocorticoid receptor binding patterns." Nat Genet **43**(3): 264-268.

Korber, P. and P. B. Becker (2010). "Nucleosome dynamics and epigenetic stability." Essays Biochem **48**(1): 63-74.

Lee, S. C., A. Mustroph, R. Sasidharan, D. Vashisht, O. Pedersen, T. Oosumi, L. A. C. J. Voesenek and J. Bailey-Serres (2011). "Molecular characterization of the submergence response of the Arabidopsis thaliana ecotype Columbia." The New Phytologist **190**(2): 457-471.

Lee, T. I. and R. A. Young (2000). "Transcription of Eukaryotic Protein-Coding Genes." Annu Rev Genet **34**: 77-137.

Lenhard, B., A. Sandelin and P. Carninci (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." Nat Rev Genet **13**(4): 233-245.

Li, B., M. Carey and J. L. Workman (2007). "The role of chromatin during transcription." Cell **128**(4): 707-719.

Liu, C., M. Wang, X. Wei, L. Wu, J. Xu, X. Dai, J. Xia, M. Cheng, Y. Yuan, P. Zhang, J. Li, T. Feng, A. Chen, W. Zhang, F. Chen, Z. Shang, X. Zhang, B. A. Peters and L. Liu (2019). "An ATAC-seq atlas of chromatin accessibility in mouse tissues." Sci Data **6**(1): 65.

Lowe, E. K., C. Cuomo, D. Voronov and M. I. Arnone (2019). "Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity." Methods Cell Biol **151**: 115-126.

Lu, Z., B. T. Hofmeister, C. Vollmers, R. M. DuBois and R. J. Schmitz (2017). "Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes." Nucleic Acids Res **45**(6): e41.

Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent and T. J. Richmond (1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution." Nature **389**(6648): 251-260.

Magnani, L., J. Eeckhoute and M. Lupien (2011). "Pioneer factors: directing transcriptional regulators within the chromatin environment." Trends Genet **27**(11): 465-474.

Maher, K. A., M. Bajic, K. Kajala, M. Reynoso, G. Pauluzzi, D. A. West, K. Zumstein, M. Woodhouse, K. Bubb, M. W. Dorrity, C. Queitsch, J. Bailey-Serres, N. Sinha, S. M. Brady and R. B. Deal (2018). "Profiling of Accessible Chromatin Regions across Multiple Plant Species

and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules." Plant Cell **30**(1): 15-36.

Miller, J. A. and J. Widom (2003). "Collaborative competition mechanism for gene activation in vivo." Mol Cell Biol **23**(5): 1623-1632.

Moreno-Romero, J., J. Santos-Gonzalez, L. Hennig and C. Kohler (2017). "Applying the INTACT method to purify endosperm nuclei and to generate parental-specific epigenome profiles." Nat Protoc **12**(2): 238-254.

Mustroph, A., S. C. Lee, T. Oosumi, M. E. Zanetti, H. Yang, K. Ma, A. Yaghoubi-Masihi, T. Fukao and J. Bailey-Serres (2010). "Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses." Plant Physiol **152**(3): 1484-1500.

Ong, C. T. and V. G. Corces (2011). "Enhancer function: new insights into the regulation of tissue-specific gene expression." Nat Rev Genet **12**(4): 283-293.

Paranjape, S. M., R. T. Kamakaka and J. T. Kadonaga (1994). "ROLE OF CHROMATIN STRUCTURE IN THE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II." Annual Review of Biochemistry **63**(1): 265-297.

Park, K., J. M. Frost, A. J. Adair, D. M. Kim, H. Yun, J. S. Brooks, R. L. Fischer and Y. Choi (2016). "Optimized Methods for the Isolation of Arabidopsis Female Central Cells and Their Nuclei." Mol Cells **39**(10): 768-775.

Pokholok, D. K., C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford and R. A. Young (2005). "Genome-wide map of nucleosome acetylation and methylation in yeast." Cell **122**(4): 517-527.

Ponnaluri, V. K. C., G. Zhang, P. O. Esteve, G. Spracklin, S. Sian, S. Y. Xu, T. Benoukraf and S. Pradhan (2017). "NicE-seq: high resolution open chromatin profiling." Genome Biol **18**(1): 122.

Quillien, A., M. Abdalla, J. Yu, J. Ou, L. J. Zhu and N. D. Lawson (2017). "Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq." Cell Rep **20**(3): 709-720.

Reynoso, M. A., G. C. Pauluzzi, K. Kajala, S. Cabanlit, J. Velasco, J. Bazin, R. Deal, N. R. Sinha, S. M. Brady and J. Bailey-Serres (2018). "Nuclear Transcriptomes at High Resolution Using Retooled INTACT." Plant Physiol **176**(1): 270-281.

Ron, M., K. Kajala, G. Pauluzzi, D. Wang, M. A. Reynoso, K. Zumstein, J. Garcha, S. Winte, H. Masson, S. Inagaki, F. Federici, N. Sinha, R. B. Deal, J. Bailey-Serres and S. M. Brady (2014). "Hairy root transformation using Agrobacterium rhizogenes as a tool for exploring cell type-specific gene expression and function using tomato as a model." Plant Physiol **166**(2): 455-469.

Rowley, M. J., M. H. Nichols, X. Lyu, M. Ando-Kuri, I. S. M. Rivera, K. Hermetz, P. Wang, Y. Ruan and V. G. Corces (2017). "Evolutionarily Conserved Principles Predict 3D Chromatin Organization." Mol Cell **67**(5): 837-852 e837.

Roy, S., J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M.

MacAlpine, L. D. Stein, K. P. White and M. Kellis (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science **330**(6012): 1787-1797.

Ruiz, J. L., J. J. Tena, C. Bancells, A. Cortes, J. L. Gomez-Skarmeta and E. Gomez-Diaz (2018). "Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum." Nucleic Acids Res **46**(18): 9414-9431.

Shashikant, T. and C. A. Ettensohn (2019). "Genome-wide analysis of chromatin accessibility using ATAC-seq." Methods Cell Biol **151**: 219-235.

Song, L. and G. E. Crawford (2010). "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." Cold Spring Harb Protoc **2010**(2): pdb prot5384.

Song, L., Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B. K. Lee, N. C. Sheffield, S. Graf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb and T. S. Furey (2011). "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity." Genome Res **21**(10): 1757-1767.

Spitz, F. and E. E. Furlong (2012). "Transcription factors: from enhancer binding to developmental control." Nat Rev Genet **13**(9): 613-626.

Stamatoyannopoulos, J. A., M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie, G. Jain, A. Sanyal, K.-B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski,

S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H. Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi, M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon and L. B. Adams (2012). "An encyclopedia of mouse DNA elements (Mouse ENCODE)." Genome biology **13**(8): 418-418.

Szklarczyk, D., J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen and C. von Mering (2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." Nucleic Acids Res **45**(D1): D362-D368.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.

Voss, T. C. and G. L. Hager (2014). "Dynamic regulation of transcriptional states by chromatin and transcription factors." Nat Rev Genet **15**(2): 69-81.

Voss, T. C., R. L. Schiltz, M. H. Sung, P. M. Yen, J. A. Stamatoyannopoulos, S. C. Biddie, T. A. Johnson, T. B. Miranda, S. John and G. L. Hager (2011). "Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism." Cell **146**(4): 544-554.

Wang, D., E. S. Mills and R. B. Deal (2012). "Technologies for systems-level analysis of specific cell types in plants." Plant Sci **197**: 21-29.

Wasson, L., N. M. Amin and F. L. Conlon (2019). "INTACT Proteomics in *Xenopus*." Cold Spring Harb Protoc **6**.

Wilkerson, B. A., A. D. Chitsazan, L. S. VandenBosch, M. S. Wilken, T. A. Reh and O. Bermingham-McDonogh (2019). "Open chromatin dynamics in prosensory cells of the embryonic mouse cochlea." Sci Rep **9**(1): 9060.

Wittkopp, P. J. and G. Kalay (2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence." Nat Rev Genet **13**(1): 59-69.

Yoshida, H., C. A. Lareau, R. N. Ramirez, S. A. Rose, B. Maier, A. Wroblewska, F. Desland, A. Chudnovskiy, A. Mortha, C. Dominguez, J. Tellier, E. Kim, D. Dwyer, S. Shinton, T. Nabekura, Y. Qi, B. Yu, M. Robinette, K. W. Kim, A. Wagers, A. Rhoads, S. L. Nutt, B. D. Brown, S. Mostafavi, J. D. Buenrostro, C. Benoist and P. Immunological Genome (2019). "The cis-Regulatory Atlas of the Mouse Immune System." Cell **176**(4): 897-912 e820.

Zaret, K. S. and J. S. Carroll (2011). "Pioneer transcription factors: establishing competence for gene expression." Genes Dev **25**(21): 2227-2241.

Zhang, W., Y. Wu, J. C. Schnable, Z. Zeng, M. Freeling, G. E. Crawford and J. Jiang (2012). "High-resolution mapping of open chromatin in the rice genome." Genome Res **22**(1): 151-162.

Zhang, W., T. Zhang, Y. Wu and J. Jiang (2012). "Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis." Plant Cell **24**(7): 2719-2731.

Zhang, W., T. Zhang, Y. Wu and J. Jiang (2014). "Open chromatin in plant genomes." Cytogenet Genome Res **143**(1-3): 18-27.

Zou, C., K. Sun, J. D. Mackaluso, A. E. Seddon, R. Jin, M. F. Thomashow and S.-H. Shiu (2011). "Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana." Proceedings of the National Academy of Sciences **108**(36): 14992.

CHAPTER 2: IDENTIFICATION OF OPEN CHROMATIN REGIONS IN PLANT GENOMES USING ATAC-SEQ

Marko Bajic, Kelsey A. Maher, and Roger B. Deal

This work is published in *Methods in Molecular Biology* (2018) 1675:183-201. doi: 10.1007/978-1-4939-7318-7_12.

ABSTRACT

Identifying and characterizing highly accessible chromatin regions assists in determining the location of genomic regulatory elements and understanding transcriptional regulation. In this chapter we describe an approach to map accessible chromatin features in plants using the Assay for Transposase Accessible Chromatin, combined with high throughput sequencing (ATAC-seq), which was originally developed for cultured animal cells. This technique utilizes a hyperactive Tn5 transposase to cause DNA cleavage and simultaneous insertion of sequencing adapters into open chromatin regions of the input nuclei. The application of ATAC-seq to plant tissue has been challenging due to the difficulty of isolating nuclei sufficiently free of interfering organellar DNA. Here we present two different approaches to purify plant nuclei for ATAC-seq: the INTACT method (Isolation of Nuclei TAgged in Specific Cell Types) to isolate nuclei from individual cell types of the plant, and tissue lysis followed by sucrose sedimentation to isolate sufficiently pure total nuclei. We provide detailed instructions for transposase treatment of nuclei isolated using either approach, as well as subsequent preparation of ATAC-seq libraries. Sequencing-ready ATAC-seq libraries can be prepared from plant tissue in as little as one day. The procedures described here are optimized for *Arabidopsis thaliana* but can also be applied to other plant species.

Key words: ATAC-seq, INTACT system, chromatin, nucleus, transposition, nucleosome, transcription factor, enhancer

INTRODUCTION

Plants are sessile organisms that must precisely regulate their transcription in response to their environment, as well as for proper development, growth, and homeostasis. Transcription is associated with regions of relatively open chromatin, in which cis-regulatory elements such as enhancers and promoters can recruit transcription factors and RNA polymerase II to transcribe DNA (Li et al., 2007). Binding of transcription factors to DNA generally results in the depletion of nucleosomes, rendering these regions hypersensitive to nucleases. Characterizing such regulatory regions throughout the genome has therefore relied on methods that combine enzymatic digestion of nuclear DNA and high-throughput sequencing, such as microccocal nuclease sequencing (MNase-seq, see Chapter 10) and DNase I Hypersensitivity sequencing (DNase-seq) (Song and Crawford, 2010; Ken, 2005). Alternatively, regulatory regions can be inferred by Chromatin Immunoprecipitation sequencing (ChIP-seq, see Chapter 5) where antibodies are used to pull down transcription factors or histone marks associated with active transcription (Park, 2009).

An improved method for identifying accessible regions of chromatin and transcription factor binding is the Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) (Buenrostro et al., 2013; Buenrostro et a., 2015). This method uses a hyperactive Tn5 transposase to integrate preloaded sequencing adapters into regions of open chromatin (Fig 2.1A). ATAC-seq is a fast protocol with simple library amplification steps and requires very small amounts of starting material, making it a vast improvement over alternative methods. However, a drawback of this protocol is that the hyperactive Tn5 transposase also targets sources of extranuclear genetic material, including the genomes of mitochondria and chloroplasts. This decreases the proportion of reads that map to the nuclear genome, reducing the amount of information that can be used to identify regulatory regions of open chromatin. Such extranuclear reads must be discarded at the start of the data analysis process, diminishing the efficiency of the assay both in terms of cost and in effective use of materials. To gain the maximum

efficiency of this powerful procedure, input material free from extranuclear genetic material, such as purified nuclei, is the ideal input for ATAC-seq

In this chapter, we describe the use of two different methods to isolate either total nuclei from tissues or nuclei from specific cell types of *Arabidopsis thaliana* (Fig 2.1B). To isolate total nuclei from plant tissue we use extraction buffers with a non-ionic detergent to lyse organelles, followed by sucrose sedimentation to further purify the nuclei (Gendre et al., 2005). This method of nuclei isolation can be done in any lab on most plant tissues. However, these partially purified nuclei still contain some organellar DNA in addition to nuclear DNA, which reduces the efficiency of Tn5 transposition to nuclear DNA and results in fewer sequencing reads that map to nuclear DNA. In addition, we describe the Isolation of Nuclei TAgged in specific Cell Types (INTACT) method to isolate nuclei from tissue or from specific cell types (Deal and Henikoff, 2010). This system uses two transgenes for nuclear targeting for affinity purification: 1) the Nuclear Tagging Fusion (NTF) construct, which encodes a fusion of WPP nuclear envelope-targeting domain, a Green Fluorescent Protein (GFP), and the Biotin Ligase Recognition Peptide (BLRP); and 2) an E. coli biotin ligase (BirA), which biotinylates the BLRP tag. The BirA is expressed from a constitutive promoter while the NTF is expressed either from a constitutive or cell type-specific promoter. The specificity of the NTF promoter determines which cell types will have biotinylated nuclei and can then be isolated by affinity purification with streptavidin-coated magnetic beads (Wang and Deal, 2015). A key advantage of the INTACT approach is not only that the isolated nuclei have less organellar DNA contamination, but also that this method can be used to selectively isolate nuclei from specific cell types. While INTACT is a powerful technique, it does require that stable transgenic lines containing BirA and NTF cassettes for the cell type of interest are available, which are time-consuming to generate and can be limiting for many species. Even so, the protocol described here, particularly ATAC-seq using sucrose sedimentation-purified nuclei, can readily be adapted for chromatin profiling in any plant species.

MATERIALS

2.1 Equipment

1. Porcelain 50 mL mortar and pestle, or equivalent.
2. Liquid nitrogen.
3. Metal lab spoon.
4. DynaMag 2 magnetic rack for 1.5 mL tubes (e.g. Life Technologies, catalog no. 12321D).
5. DynaMag 15 magnetic rack for 15 mL tubes (e.g. Life Technologies, catalog no. 12301D).
6. MagWell 96 well magnetic separator plate (e.g. EdgeBio, catalog no. 57624)
7. Nylon cell strainers with 70 μm pores.
8. Long-stem analytical funnel.
9. Pipet-Aid.
10. Sterile 10 mL plastic serological pipettes.
11. Eppendorf tubes, 1.5 mL.
12. PCR tubes, 0.2 mL.
13. Falcon tubes, 15 and 50 mL.
14. Nutator platform rotator.
15. Hemocytometer (e.g. Hausser Bright Line hemocytometer, Fisher Scientific)
16. Microcentrifuge and refrigerated centrifuge with rotor for 15 mL tubes.
17. Cold room, 4 °C.
18. Molecular biology grade water.
19. Sterile disposable filter unit, 500 mL.
20. Sterile 0.2 μm syringe filter.
21. Sterile 10 mL plastic syringe.
22. Thermal cycler
23. Real-Time PCR machine

24. A 64-bit computer with at least 1 TB hard disk and 16 Gb of memory for ATAC-seq data analysis.

25. Fluorescent microscope.

2.2 Stock Solutions and Reagents

1. Complete, EDTA-free Protease Inhibitors (e.g. Roche).

2. Stock solution of 2 M spermidine. Prepare by dissolving 2.904 g spermidine powder in 10 mL water. Aliquot 1mL of solution per 1.5 mL Eppendorf tube and store at -20 °C.

3. Stock solution of 200 mM spermine. Prepare by dissolving 0.4047 g spermine powder in 10 mL of water. Aliquot 1 mL of solution per 1.5 mL Eppendorf tube and store at -20 °C.

4. Stock solution of incomplete Nuclei Purification Buffer (NPBi): 20 mM MOPS, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, adjusted to pH 7 with 2M KOH. Filter sterilize the solution and degas under vacuum for 10 minutes. Store at 4 °C for up to 3 months.

5. Stock solution of 10% Triton X-100.

6. Stock solution of 10X DAPI. Prepare by dissolving 10 mg DAPI powder in 5 mL water, for a final concentration of 2 μg/μL. Filter sterilize the solution and store at 4 °C in the dark for several months. To stain nuclei with DAPI, dilute the 10X DAPI solution to 1X using water (final concentration of 0.2 μg/μL), and use within 2-3 hours.

2.3 Purification of Tagged Nuclei using INTACT

1. Plant material: tissue from transgenic plants expressing both NTF and BirA in the cell type of interest. INTACT transgenic lines targeting the root epidermal hair and non-hair cell types, as well as INTACT plasmid vectors are available from the Arabidopsis Biological Resource Center at Ohio State University.

2. M-280 Streptavidin Dynabeads (e.g. Life Technologies).

3. Nuclei Purification Buffer (NPB): 20 mM, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine, 0.2 mM spermine, 1X Roche Complete protease inhibitors, adjusted to pH 7 with 2M KOH. Prepare by adding spermidine, spermine, and Roche Complete protease inhibitors to NPBi just before starting the INTACT nuclei purification procedure. Keep solution on ice, and use within 1 hour of preparation.

4. Nuclei Purification Buffer containing 0.1% Triton X-100 (NPBt): 20 mM MOPS pH 7, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine, 0.2 mM spermine, 0.1% (v/v) Triton X-100. Prepare by adding spermidine, spermine, and Triton X-100 to NPBi just before starting the INTACT nuclei purification procedure. Keep solution on ice, and use within 1 day of preparation.

2.4 Purification of Total Nuclei using Sucrose Sedimentation

1. Plant material: fresh or frozen plant tissue.
2. Stock solution of 1M Tris-HCl pH 8
3. Stock solution of 1M MgCl2
4. Stock solution of 2M sucrose.
5. Nuclei Purification Buffer (NPB): 20 mM MOPS pH7, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine, 0.2 mM spermine, 1X Roche Complete protease inhibitors. Prepare by adding spermidine, spermine, and Roche Complete protease inhibitors to NPBi just before starting the nuclei purification procedure. Keep solution on ice, and use within 1 hour of preparation.
6. Nuclei Extraction Buffer 2 (NEB 2): 0.25 M Sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl2, 1% Triton X-100, 1X Roche Complete Protease Inhibitors. Prepare solution just before use, keep on ice, and use within 1 hour of preparation.
7. Nuclei Extraction Buffer 3 (NEB 3): 1.7 M Sucrose, 10 mM Tris-HCl pH 8, 2 mM MgCl2, 0.15% Triton X-100, 1X Roche Complete Protease Inhibitors. Prepare solution just before use, keep on ice, and use within 1 hour of preparation.

2.5 Tagmentation of Chromatin by Tn5 transposase

1. Nextera Library Kit (Illumina, FC-121-1030).
2. MinElute PCR Purification kit (Qiagen).

2.6 Sequencing Library Preparation

1. ATAC Primer 1 (AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG)
2. ATAC barcoded Primer 2 (CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTCTCGTGGGCTCGGAGATGT); N's indicate the 8-base index sequence. Each library to be pooled for sequencing should be amplified with a different barcoded primer 2. See Supplementary Table 2.1 of (Buenrostro et al., 2013) for all primer sequences.
3. NEBNext High-Fidelity 2X PCR Master Mix (NEB).
4. Solution of 20X EvaGreen dye (Biotium).
5. Solution of 50X ROX dye (Invitrogen).
6. MinElute PCR Purification kit (Qiagen).
7. Agencourt Ampure XP PCR Purification beads (Beckman Coulter).
8. 100% ethanol.
9. Horizontal electrophoresis gel box and power source.
10. 302 nm ultraviolet transilluminator.
11. NEBNext Library Quantification kit for Illumina (NEB)

METHODS

Users should either begin at section 3.1 for affinity purification of nuclei using INTACT, or at section 3.2 for isolation of total nuclei. In either case, the purified nuclei are used for tagmentation by Tn5

transposase in step 3.3. All procedures are carried out at room temperature (25 °C) unless otherwise specified.

3.1 Purification of Tagged Nuclei Using INTACT

1. Grind tissue (3 g of roots or 0.5 g of leaves) to a fine powder in liquid nitrogen using a mortar and pestle. Using a nitrogen-cooled metal lab spoon, quickly transfer the frozen tissue powder to another mortar containing 10 mL of ice-cold Nuclei Purification Buffer (NPB). Thoroughly resuspend the powder in NPB by grinding it with a new, cold pestle (see Note 1).
2. Use a 10 mL serological pipette to draw up the tissue suspension and filter it through a 70 μm nylon cell strainer, placed in the center of a long stemmed funnel. Collect the flow-through in a chilled 15 mL tube on ice.
3. Spin down the nuclei at 1,200 x g for 10 minutes at 4 °C. Use a 10 mL serological pipet and then a 1 mL pipette tip to carefully remove as much of the supernatant as possible without disturbing the pellet.
4. Gently resuspend the pellet in 1 mL of ice-cold NPB. Transfer the crude nuclei suspension to a 1.5 mL tube. Keep on ice.
5. Wash the appropriate amount of Streptavidin M280 Dynabead suspension (25 μL for nuclei from 3 g of roots or 10 μL for 0.5 g of leaves) with 1 mL of ice-cold NPB in a 1.5 mL tube. Collect the beads on the DynaMag2 magnetic rack. Discard the supernatant and resuspend the beads with ice-cold NPB to their original volume (e.g. 25 μL). Keep on ice.
6. Add the washed and resuspended beads to the 1 mL of resuspended nuclei from Step 4. Rotate on a nutator in a 4 °C cold room for 30 minutes. Work in the 4 °C cold room for Steps 7-14.
7. Transfer the 1 mL bead-nuclei mixture to a 15 mL tube and slowly add to it 13 mL of ice-cold NPBt. Mix gently and place on a nutator for 30 seconds.
8. Place the 15 mL tube in the DynaMag 15 magnetic rack for 2 minutes to capture the nuclei-beads along the walls of the tube.

9. Slowly remove the NPBt supernatant with a serological pipette, making sure not to disturb the beads on the side walls of the tube. Gently resuspend the beads with 14 mL of ice-cold NPBt, mix gently, and place on a nutator for 30 seconds.

10. Place the 15 mL tube in the DynaMag 15 magnetic rack for 2 minutes to capture the nuclei and beads.

11. Repeat Steps 9 and 10 one more time, for a total of three washes.

12. Slowly remove the NPBt supernatant with a serological pipette. Resuspend the beads in 1 mL of ice-cold NPBt. Remove 25 µL of this nuclei-bead suspension to a 0.6 ml tube on ice for counting captured nuclei with a hemocytometer.

13. Transfer the remaining nuclei-bead suspension to an ice-cold 1.5 mL tube. Place the 1.5 mL tube in the DynaMag 2 magnetic rack to capture the beads along the walls of the tube.

14. Carefully remove the NPBt supernatant and resuspend the nuclei-beads in 20 µL of ice-cold NPB. Keep on ice until the nuclei are counted and ready for tagmentation. (see Note 2).

15. To view and quantify nuclei under a light microscope, add 1 µL of diluted DAPI solution (0.2 µg/µL) to each 25 µL aliquot of nuclei from Step 12. Mix well, and place on ice for 5 minutes in the dark.

16. Use a hemocytometer to count the DAPI-stained, bead-bound nuclei and determine the total yield. Purified nuclei should appear as shown in Figure 2.1C (see Note 3).

17. Use the calculated total yield to determine the volume of resuspended nuclei from Step 14 needed to obtain 50,000 nuclei for the ATAC-seq reaction. Transfer this volume of resuspended nuclei to a new 0.2 mL tube, and keep on ice. Immediately proceed to Section 3.3.

3.2 Purification of Total Nuclei Using Sucrose Sedimentation

1. Grind 0.1 to 1 g of plant tissue to a fine powder in liquid nitrogen using a mortar and pestle (see Note 4).

2. Using a nitrogen-cooled metal lab spoon, quickly transfer the frozen tissue powder to another mortar containing 10 mL ice-cold NPB. Thoroughly resuspend the powder in NPB by grinding it with a new, cold pestle.

3. Use a 10 mL serological pipette to draw up the tissue suspension and filter it through a 70 μm nylon cell strainer, placed in the center of a long stemmed funnel. Collect the flow-through into a 15 mL tube on ice.

4. Centrifuge the tube at 1,200 x g for 10 minutes at 4 °C.

5. Gently remove the supernatant and gently but thoroughly resuspend the pellet in 1 mL of ice-cold NEB2 buffer. Transfer this suspension to a new 1.5 mL microcentrifuge tube.

6. Spin the resuspended nuclei at 12,000 x g for 10 minutes at 4 °C.

7. Carefully remove the supernatant and resuspend the pellet thoroughly in 300 μL of NEB3 buffer.

8. Add 300 μL of ice-cold NEB3 to a new 1.5 mL microcentrifuge tube. Carefully layer the resuspended pellet from Step 7 on top of the fresh NEB3. Centrifuge at 16,000 x g for 10 minutes at 4 °C (see Note 5).

9. Carefully remove the supernatant and resuspend the nuclei pellet in 1 mL of cold NPB. Keep these nuclei on ice.

10. Remove 25 μL of this nuclei suspension and move to a fresh 0.6 ml tube on ice. To this add 1 μL of diluted DAPI solution (0.2 μg/μL). Mix well and place on ice for 5 minutes in the dark.

11. Use a hemocytometer to quantify the DAPI-stained nuclei and determine the total yield. Purified nuclei should appear as shown in Fig 2.1C (see Note 6).

12. Use the calculated total yield to determine the volume of resuspended nuclei from Step 9 needed to obtain 50,000 nuclei for the ATAC-seq reaction. Transfer this volume of the resuspended nuclei to a new 0.2 mL tube, and keep on ice. Immediately proceed to Section 3.3.

3.3 Tagmentation with Tn5 Transposase

1. Prepare the transposition reaction master mix in a 0.2 mL PCR tube on ice according to Table 2.1 and mix well. The volumes given in Table 2.1 are for a single reaction with 50,000 nuclei.

2. If the nuclei were isolated using the Sucrose Sedimentation procedure, pellet 50,000 nuclei from Subheading 3.2 Step 9 by spinning the appropriate volume of nuclei at 1,500 x g for 7 minutes at 4 °C. Remove the supernatant, and resuspend the nuclei in 50 µL of ice-cold transposition reaction mix prepared in step 1. Move the reaction to a 0.2 mL PCR tube on ice. If the nuclei were isolated using the INTACT procedure, move 50,000 bead-bound nuclei from Subheading 3.1 Step 14 into a 0.2 mL tube and capture the beads on the tube wall in a MagWell 96 well magnetic plate on ice. Remove the supernatant, and resuspend the bead-bound nuclei in 50 µL of ice-cold transposition reaction mix. Keep on ice.

3. Place the transposition reaction in a thermal cycler block pre-warmed to 37 °C and incubate for 30 minutes with occasional gentle mixing to keep the nuclei in suspension.

4. Purify the transposed DNA using the Qiagen MinElute PCR purification kit according the manufacturer's instructions. Elute DNA in 11 µL of elution buffer EB, provided in the kit. DNA can now be stored at -20 °C until future use, or used immediately for PCR amplification.

3.4 PCR Amplification of the DNA Library

1. Prepare the PCR amplification mix in a 0.2 mL tube on ice according to Table 2.2. Mix well, and perform PCR cycling as described in Table 2.3 (see Note 7).

2. Once the thermal cycler reaches 4 °C, remove the samples and place them on ice.

3. To determine the number of additional PCR cycles needed to adequately amplify the DNA library, prepare the qPCR Library Amplification Mix described in Table 2.4 in a 0.2 mL PCR tube. Keep the mixture on ice.

4. Perform thermal cycling in the qPCR machine according to Table 2.5.

5. To determine the optimal number of cycles needed to amplify the remaining 45 µL of each library from Step 2, view the linear fluorescence versus cycle number plot on the qPCR machine once the

reaction is finished. The cycle number at which the fluorescence for a given reaction is at 1/3 of its maximum is the number of additional cycles (N) that each library requires for adequate amplification (see Note 8).

6. Run the remaining 45 µL of each PCR reaction from Step 2 according to Table 2.6.

7. Purify the libraries by mixing Ampure XP beads with the reaction products at a 1.5:1 ratio of beads:PCR sample by volume (see Note 9). Incubate at room temperature for 5 minutes.

8. Place the 0.2 mL tube on the MagWell 96 well magnetic plate for 1 minute to capture the Ampure beads, and discard the supernatant.

9. With the tube still in the magnetic plate, wash the beads twice for 30 seconds each with 200 µL of 80% ethanol without disturbing the bead pellet. After the last wash, allow the beads to dry for 5 minutes to remove all traces of ethanol (see Note 10).

10. Remove the tube from the magnet and resuspend the bead pellet in 20 µL 10 mM Tris pH 8. Incubate at room temperature for 2 minutes, capture the beads on the magnet, and transfer the supernatant into a fresh 0.2 mL PCR tube on ice. A small aliquot of the library, 1-2 µL, can be run on a 2% agarose gel to visualize the abundance and size distribution of amplified libraries (Fig 2.2A) (see Note 11). The purified libraries can now be stored at -20 °C.

11. Quantify the molar concentrations of the libraries using the NEBNext Library Quantification kit for Illumina, according to manufacturer's directions. Alternatively, other qPCR-based library quantification kits can be used to determine the concentration of the amplified libraries.

12. Once quantified, the libraries are ready for pooling and high-throughput sequencing on the Illumina platform (see Note 12).

13. The quality of the sequencing reads, alignment to the genome, fragment size distribution (Fig 2.2B), and downstream analyses can be performed as described in Note 13. A genome browser shot of the typical Arabidopsis ATAC-seq data from libraries made using the procedures described here can be seen in Fig 2.2C.

NOTES

1. This protocol is optimized for 3 g of root or 0.5 g of leaf tissue from *Arabidopsis thaliana*. Ground leaf tissue contains more debris, relative to roots, and therefore requires a lower amount of starting material to obtain highly purified nuclei. INTACT may also be performed on fresh tissue by chopping the tissue in NPB as opposed to grinding to a fine powder using liquid nitrogen. However, this approach does require the use of fresh tissue. The number of samples that can be run through INTACT purification simultaneously is mainly limited by the capacity of the DynaMag 15 magnetic rack used for nuclei capture. Up to four separate samples can be processed in parallel using one DynaMag 15 magnetic rack.

Using an INTACT line with nuclei labeled in the root epidermal non-hair cell type, approximately 200,000 purified nuclei can be obtained from 3g of roots. Larger amounts of tissue can be used for purifying nuclei from less abundant cell types, and this generally only requires adjustments to the amount of streptavidin beads used and the volume of solution used for bead capture. See (Wang and Deal, 2015) for more details on variations in the INTACT procedure.

2. After isolating the bead bound nuclei, keep the sample on ice while quantifying the nuclei from the aliquot in Subheading 3.1 Step 12. Do not freeze the isolated nuclei before doing tagmentation and library preparation. Freezing and thawing of isolated nuclei can disrupt protein-DNA interactions.

3. After DAPI staining, nuclei purified by INTACT can be easily identified and counted using a hemocytometer. The ideal setup for visualizing nuclei is under a mix of dim white light and DAPI channel fluorescence. The dim white light allows for visualization of the hemocytometer grid and the beads, and the DAPI fluorescence allows for the visualization of nuclei. A sample image of isolated bead-bound nuclei is shown in Fig 2.1C. A nucleus is identified as a circle that fluoresces in the DAPI channel and has several beads clustered around it. Minimal cellular debris or contaminating unbound nuclei

should be observed in the final product. These contaminants may be further reduced by using fewer beads and by increasing the volumes of NPB and NPBt used during purification as described in Note 1. We have successfully used as few as 20,000 to as many as 200,000 INTACT-purified nuclei in this procedure without altering any other parameters of the protocol presented here.

4. This protocol is optimized for less than 1 g of root or 0.5 g of leaf tissue. Ground leaf tissue contains more debris relative to roots, and therefore requires a lower amount of starting material to obtain purified nuclei. As with the INTACT protocol, sucrose sedimentation of nuclei may also be performed on fresh tissue by chopping the tissue in NPB as opposed to grinding to a fine powder using liquid nitrogen. However, this approach does require the use of fresh tissue. We recommend starting with the minimum amount of tissue needed to obtain the required number of nuclei (e.g. 50,000 per ATAC-seq reaction).

5. Proper separation of nuclei from other cellular debris requires the nuclei to pass through the sucrose cushion during centrifugation. The NEB3 resuspended nuclei should therefore be placed gently on top of NEB3 layer present in the tube. After centrifugation, the contaminating organelles and debris may be visible at the top of the tube. If leaf tissue was used, the top layer will become greener after centrifugation and the pellet will become noticeably less green than it was prior to centrifugation.

6. After DAPI staining, nuclei purified by sucrose sedimentation can be identified and quantified using a hemocytometer. A mixture of DAPI-channel fluorescence and white light illumination allows the stained nuclei and the hemocytometer grid to be seen simultaneously. A sample image of isolated nuclei is shown in Fig 2.1C. A nucleus is identified as a punctate circle with strong DAPI fluorescence. The nucleus is typically ~5 μm in size and can be easily identified at 200X and 400X magnifications. Cellular debris may be observed in the final preparation, but this generally does not affect the outcome of the ATAC-seq procedure. To reduce cellular debris contamination, starting tissue can be chopped with a

razor blade (see Note 4) and/or additional NEB3 wash steps may also be done by repeating Subheading 3.2 steps 7-9 for a second sucrose cushion centrifugation.

7. Ensure that all work surfaces, pipettes, and reagents needed for amplification and library preparation are free of DNA contamination. For library amplification, unique barcoded adapters are used for each sample if multiple libraries are to be sequenced in an individual flow cell lane. The sequences of all primers can be found in the supplementary material of (Buenrostro et al., 2013).

8. The number of PCR cycles needed to amplify ATAC libraries is determined by the PCR reaction in Subheading 3.4 step 5. We recommend using the minimum number of cycles necessary to obtain a sufficient molar amount of library for Illumina sequencing. This must be determined empirically and will also depend on the number of libraries to be pooled for sequencing.

9. The ratio of Ampure XP PCR Purification beads to PCR volume determines the size of purified DNA fragments isolated. The 1.5 Ampure bead to PCR reaction ratio results in the isolation of DNA fragments shown in Fig 2.2A. Using ratios that have higher proportions of beads may result in purification of sequencing adapters and PCR primers, which can negatively affect sequencing.

10. A drying time of 5 minutes is generally sufficient to remove all traces of ethanol from the beads, but this time may vary based on humidity and room temperature. Georgia is very humid in the summer. Ensure that all ethanol has evaporated before moving on to the next step. Do not allow beads to dry to the extent that the pellet begins to crack.

11. Libraries can generally be visualized by agarose gel electrophoresis followed by ethidium bromide staining. Sensitivity can be greatly increased by staining the gel with Sybr green stain or using an Agilent Bioanalyzer or equivalent instrument, if available.

The libraries that we have prepared using this method generally present as a DNA smear starting at ~180 bp and ranging to greater than 1 kb, with peak intensity between ~180 – 500 bp (See Figure 2.2A). The original publication on ATAC-seq (Buenrostro et al., 2013) reported a nucleosome-like periodicity in the library size distribution, but we have not observed this phenomenon as assayed by either electrophoresis or estimation of fragment size distribution based on distance between paired-end sequencing reads, as shown in Fig 2.2B. This lack of observed nucleosome fractions may be due to size selection of library fragments by Ampure XP beads and the low transposase to nuclei ratio described in this protocol.

12. Paired-end sequencing is recommended in order to maximize the number of transposase integration events that can be observed in a given sample and to allow measurement of the length of the sequenced fragments (Fig 2.2B).

To identify open chromatin regions in Arabidopsis, users should aim to obtain at least 10-20 million reads per library that map to the nuclear genome. For transcription factor footprinting the number of nuclear genome-mapping reads should be increased to at least 100 million per library.

When using sucrose sedimentation for nuclei purification, users should expect ~50% of reads to map to the nuclear genome, while the use of INTACT purification will increase this number to > 90%.

13. Sequencing reads are checked for overall quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) or equivalent. The reads are aligned to the TAIR10 Arabidopsis thaliana genome (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release) using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). The resulting SAM file is converted to a binary BAM file, which is sorted and indexed using Samtools (http://samtools.sourceforge.net/). The quality of the resulting BAM file, including fragment size distribution, is analyzed using Picard Tools (https://broadinstitute.github.io/picard/). Alignment data is visualized using the Integrated Genome Viewer (http://software.broadinstitute.org/software/igv/). For ease of visualization, BAM files were

converted to BigWig files using DeepTools BamPECoverage tool (http://deeptools.readthedocs.io/en/latest/index.html). Downstream analyses of ATAC-seq data include calling peaks with HOMER (http://homer.salk.edu/homer/index.html), editing BED files with bedtools (http://bedtools.readthedocs.io/en/latest/) and identifying transcription factor footprints using pyDNase (http://pythonhosted.org/pyDNase/).

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Grant no. 1238243. We thank Paja Sijacic and Shannon Torres for helping to optimize the protocol for nuclei isolation and for suggestions on the manuscript.

TABLES AND FIGURES

Fig 2.1

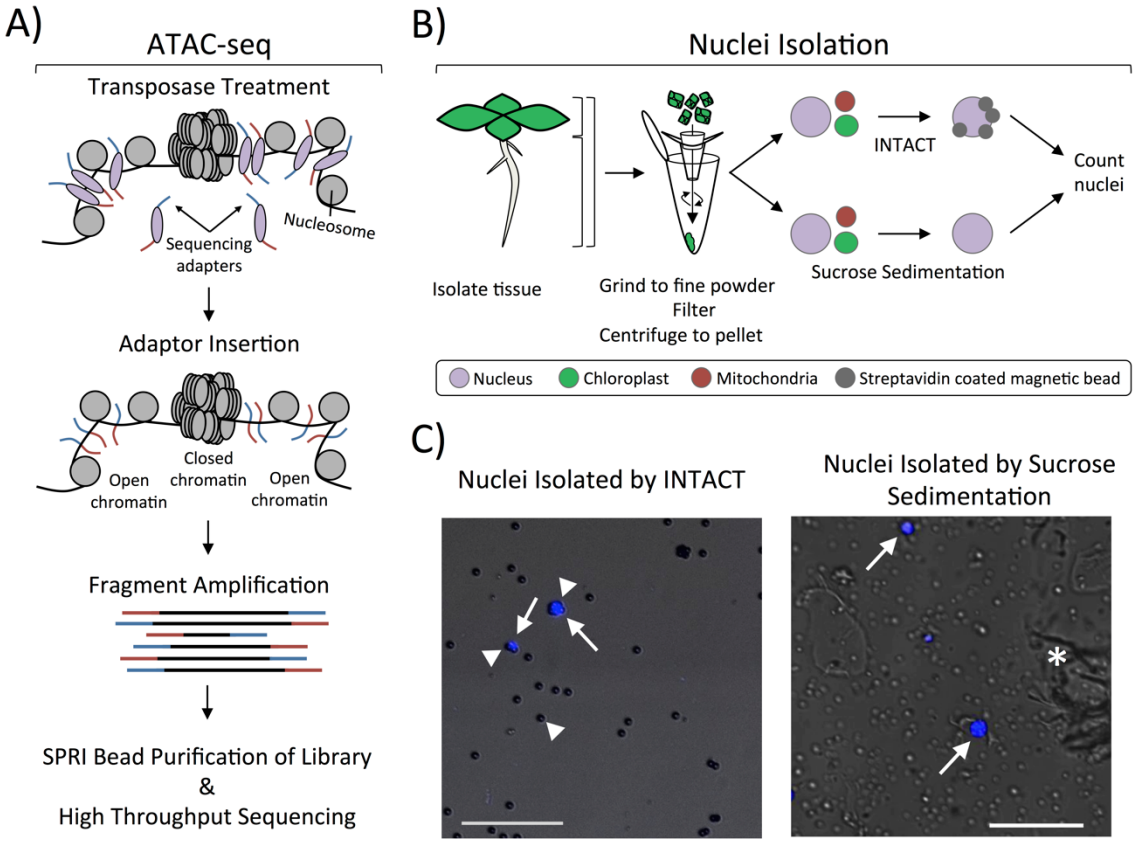


Figure 2.1 ATAC-seq profiling using nuclei isolated by INTACT or sucrose sedimentation. A) Overview of the ATC-seq procedure. Nuclei are incubated with sequencing adapter-loaded Tn5 transposase, which diffuses into the nucleus to interact with chromatin. Sequencing adapters are inserted into open chromatin regions, and the fragmented DNA is amplified wherever the sequencing adapters were inserted. This generates a library of DNA fragments in which each end represents an insertion site. The amplified libraries are purified and sequenced with next generation sequencing. B) Two different methods for purifying nuclei from Arabidopsis thaliana can be used: 1) INTACT for isolating nuclei from specific cell types, and 2) sucrose sedimentation to isolate total nuclei from input tissue. The two methods have the same initial steps: tissue is collected from a specific part of the plant (root, leaf, or the entire plant), ground to a fine powder, resuspended, filtered, and centrifuged to pellet nuclei and cellular debris. Nuclei isolation using tissue that expresses INTACT transgenes uses streptavidin coated magnetic beads

to affinity purify biotinylated nuclei out of the resuspended pellet. This allows for the isolation of nuclei from specific cell-types that express the nuclear tagging fusion (NTF) and the biotin ligase BirA, resulting in very low contamination by organellar genomes. Alternatively, total nuclei can be isolated from tissue by resuspending the nuclei/debris pellet in a buffer with Triton X-100 to lyse organelles and centrifuging through a dense sucrose layer. Nuclei isolated from both procedures are stained with DAPI and quantified using a hemocytometer. C) Fluorescent microscope images of nuclei (white arrows) stained with the DNA-binding dye DAPI (blue) isolated either through INTACT or sucrose sedimentation. INTACT isolated nuclei are identified by their DAPI-fluorescence and binding to multiple beads (white arrowhead). Beads are easily visualized by increasing transmission of white light while viewing the nuclei in the DAPI channel. Sucrose sedimentation isolated nuclei (white arrows) are DAPI-stained objects around 4-6 μm in diameter, although they can vary in size and shape depending on starting tissue. Much more cellular debris (white asterisk) is observed in sucrose sedimentation-isolated nuclei as compared to INTACT-purified nuclei, but this should not impact the procedure described here. Each picture contains a 50 μm scale bar shown at the bottom left.

Fig 2.2

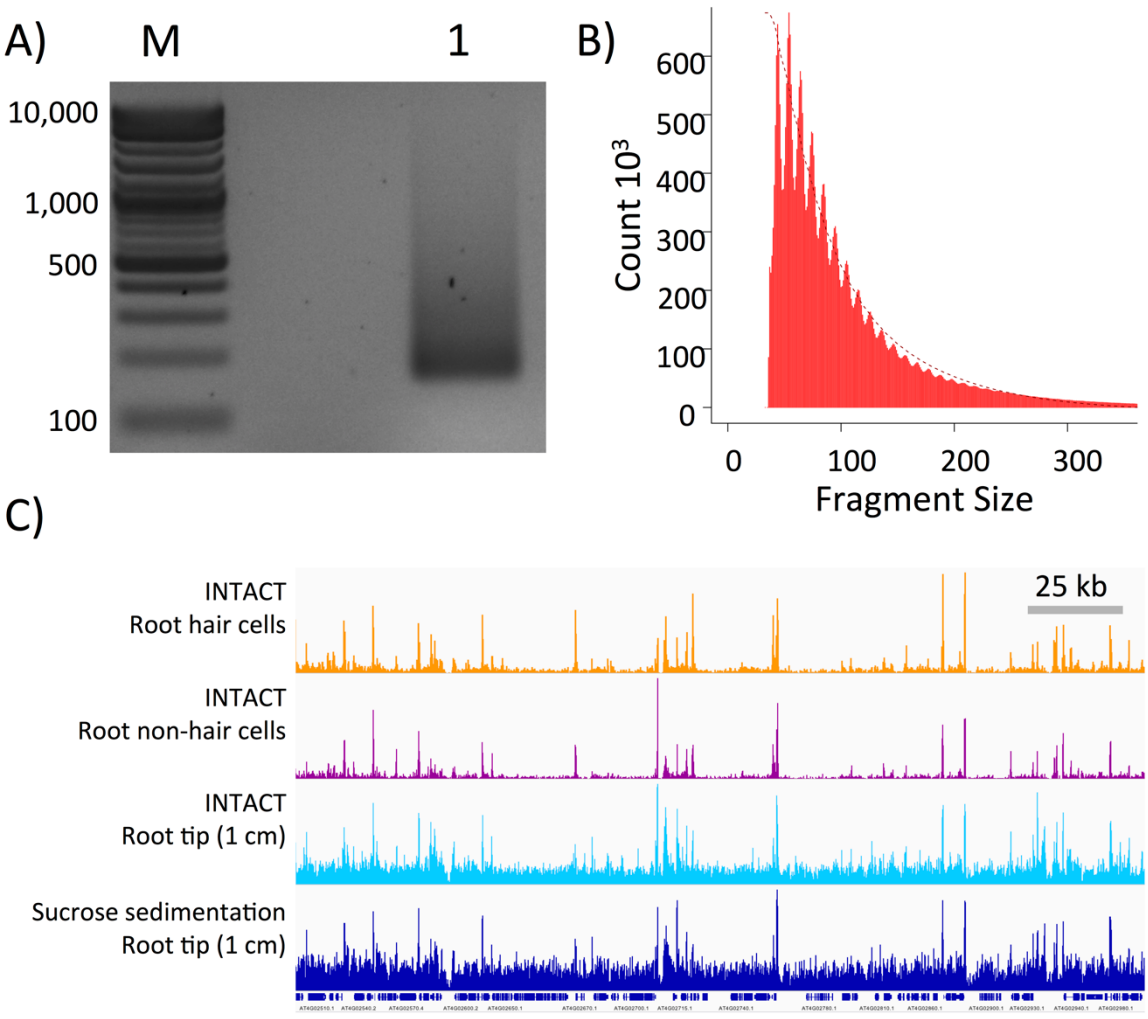


Figure 2.2 ATAC-seq library preparation and high-throughput sequencing. A) An amplified ATAC-seq library purified with Ampure XP beads (lane "1") was resolved in a 2% agarose gel stained with ethidium bromide. Lane "M" is the molecular weight marker lane. Amplified library fragments generally range in size from 180 bp to several kb in size. The size distribution of the resolved gel may vary somewhat, but the final product should be free of adapter dimers (distinct band around 125 bp) and primer dimers (distinct band around 80 bp). See Note 11. B) Insert sizes of ATAC-seq paired-end reads from 50,000 nuclei isolated by INTACT from non-hair cells calculated using the InsertSizeMetrics option from Picard Tools (Note 13). The distribution shows periodicity of helical pitch of DNA for fragments smaller than 200 bp. Fragments containing one or more nucleosomes, related to insert periodicity increasing in

150 bp, were not observed using the transposase:nuclei and bead:DNA ratios described in this protocol. C) Integrated Genome Viewer snapshot of four different libraries sequenced on the Illumina platform. The tracks shown are of ATAC sequencing reads from INTACT isolated nuclei from root hair cells (orange), root non-hair cells (purple), root tip (cyan), and sucrose sedimentation isolated nuclei from 1 cm root tip (navy). Gene tracks are shown below the ATAC-seq tracks and a 25 kb scale bar is shown.

Table 2.1

Transposition reaction mix

Component	Volume (µL)
2X TD Buffer	25
Water	22.5
TDE1 Transposase	2.5
Total	50

Table 2.2

Transposed DNA Amplification mix

Component	Volume (µL)
Transposed DNA (from Subheading 3.3 step 4)	10
Water	10
25 µM ATAC Primer 1	2.5
25 µM ATAC barcoded Primer 2*	2.5
2X NEBNext High Fidelity PCR Mix	25
Total	50

*A different barcoded Primer 2 should be used for each library that is to be pooled into a single sequencing run.

Table 2.3

Thermal Cycling Conditions for Transposed DNA Amplification

Cycle number	Temperature (°C)	Time
1	72	5 min
	98	30 sec
5 cycles	98	10 sec
	63	30 sec
	72	1 min
	4	Hold

Table 2.4

qPCR Library Amplification Mix

Component	Volume (µL)
Amplified library (from Subheading 3.4 step 2)	5
Water	0.45
25 µM ATAC Primer 1	0.5
25 µM ATAC barcoded Primer 2	0.5
20X Evagreen dye	0.75
50X ROX dye*	0.30
2X NEBNext High Fidelity PCR Mix	7.5
Total	15

*ROX concentration may vary depending on qPCR instrument. The amount described here is optimized for the ABI Step-One-Plus instrument.

Table 2.5

qPCR Cycling Conditions to Determine Additional Library Amplification Cycles

Cycle number	Temperature (°C)	Time
1	98	30 sec
20 cycles	98	10 sec
	63	30 sec
	72	1 min

Table 2.6

Final Library Amplification

Cycle number	Temperature (°C)	Time
1	98	30 sec
N cycles	98	10 sec
	63	30 sec
	72	1 min
	4	Hold

LITERATURE CITED

Li B., Carey M., Workman J.L. (2007) The role of chromatin during transcription. Cell 128:707-719

Song L., Crawford G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010: pdb prot5384. doi:10.1101/pdb.prot5384

Ken Z. (2005) Micrococcal Nuclease Analysis of Chromatin Structure. Current Protocols in Molecular Biology 21.1:1-17

Park P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680

Buenrostro J.D., Giresi P.G., Zaba L.C., Chang H.Y., Greenleaf W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10: 1213-1218

Buenrostro J.D., Wu B., Chang H.Y., Greenleaf W.J. (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 109:21.29 1-9

Gendrel A., Lippman Z., Martienssen R., Colot V. (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. Nature Methods 2: 213-218

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

Wang D., Deal R.B. (2015) Epigenome Profiling of Specific Plant Cell Types Using a StreamLined INTACT Protocol and ChIP-seq. Mehods Mol Biol 1284: 3-25

CHAPTER 3: PROFILING OF ACCESSIBLE CHROMATIN REGIONS ACROSS MULTIPLE PLANT SPECIES AND CELL TYPES REVEALS COMMON GENE REGULATORY PRINCIPLES AND NEW CONTROL MODULES

Kelsey A. Maher*, Marko Bajic*, Kaisa Kajala, Mauricio Reynoso, Germain Pauluzzi, Donnelly A. West, Kristina Zumstein, Margaret Woodhouse, Kerry Bubb, Michael W. Dorrity, Christine Queitsch, Julia Bailey-Serres, Neelima Sinha, Siobhan M. Brady, and Roger B. Deal

This work is published in *Plant Cell* (2018) 1:15-36. doi: 10.1105/tpc.17.00581.

Supplemental tables can be found in the online publication.

*These authors contributed equally to this work.

ABSTRACT

The transcriptional regulatory structure of plant genomes remains poorly defined relative to animals. It is unclear how many *cis*-regulatory elements exist, where these elements lie relative to promoters, and how these features are conserved across plant species. We employed the Assay for Transposase-Accessible Chromatin (ATAC-seq) in four plant species (*Arabidopsis thaliana*, *Medicago truncatula*, *Solanum lycopersicum*, and *Oryza sativa*) to delineate open chromatin regions and transcription factor (TF) binding sites across each genome. Despite 10-fold variation in intergenic space among species, the majority of open chromatin regions lie within 3 kb upstream of a transcription start site in all species. We find a common set of four TFs that appear to regulate conserved gene sets in the root tips of all four species, suggesting that TF-gene networks are generally conserved. Comparative ATAC-seq profiling of *Arabidopsis* root hair and non-hair cell types revealed extensive similarity as well as many cell type-specific differences. Analyzing TF binding sites in differentially accessible regions identified a MYB-driven regulatory module unique to the hair cell, which appears to control both cell fate regulators and abiotic stress responses. Our analyses

revealed common regulatory principles among species and shed light on the mechanisms producing cell type-specific transcriptomes during development.

INTRODUCTION

The transcription of protein coding genes is controlled by regulatory DNA elements, including both the core promoter and more distal enhancer elements (Lee and Young, 2000). The core promoter is a short DNA region surrounding the transcription start site (TSS), at which RNA polymerase II and general transcription factors are recruited. Enhancer elements act as platforms for recruiting both positive- and negative-acting transcription factors (TFs), and serve to integrate multiple signaling inputs in order to dictate the spatial and temporal control of transcription from the core promoter. As such, enhancer functions are critical for directing transcriptional output during cell differentiation and development, as well as coordinating transcriptional responses to environmental change (Ong and Corces, 2011). Despite their importance, only a small number of *bona fide* enhancers have been characterized in plants, and we lack a global view of their general distribution and action in plant genomes (Weber et al., 2016).

In large part, our limited knowledge of plant *cis*-regulatory elements arises from the unique difficulties in identifying these elements. While some enhancers exist near their target core promoter, others can be thousands of base pairs upstream or downstream, or even within the transcribed region of a gene body (Ong and Corces, 2011; Spitz and Furlong, 2012). Furthermore, enhancers generally do not display universal sequence conservation, aside from sharing of individual TF binding sites, which makes them very challenging to locate. By contrast, core promoters can be readily identified through mapping the 5' ends of transcripts (Morton et al., 2014; Mejia-Guerra et al., 2015). It was recently discovered that many enhancer elements in animal genomes could be identified with relatively high confidence based on a unique combination of flanking histone posttranslational modifications (PTMs), such as an enrichment for H3K27ac and H3K4me1. This characteristic histone PTM signature has led to the annotation of such elements in several animal models and specialized cell types (Heintzman et al., 2009; Bonn et al., 2012). However, the only currently known association between plant *cis*-regulatory elements and histone PTMs

appears to be a modest correlation with H3K27me3 (Zhang et al., 2012b; Zhu et al., 2015). Though encouraging, this mark is not unique to these elements, and cannot be used to identify enhancers on its own.

A long-known and general feature of sequence-specific DNA-binding proteins is their ability to displace nucleosomes upon DNA binding, leading to an increase in nuclease accessibility around the binding region (Gross and Garrard, 1988; Henikoff, 2008). In particular, DNaseI treatment of nuclei coupled with high-throughput sequencing (DNase-seq) has been used to probe chromatin accessibility. This technology has served as an important tool in identifying regulatory elements throughout animal genomes (Thurman et al., 2012) and more recently in certain plant genomes (Zhang et al., 2012b; Zhang et al., 2012a; Pajoro et al., 2014; Sullivan et al., 2014). In addition, a differential micrococcal nuclease sensitivity assay has also been used to probe functional regions of the maize genome, demonstrating the versatility of this approach (Vera et al., 2014; Rodgers-Melnick et al., 2016).

DNase-seq has been used successfully to identify open chromatin regions in different tissues of both rice and *Arabidopsis* (Zhang et al., 2012a; Pajoro et al., 2014; Zhu et al., 2015). Over a dozen of the intergenic DNase-hypersensitive sites in *Arabidopsis* were tested and shown to act as enhancer elements by activating a minimal promoter-reporter cassette, demonstrating that chromatin accessibility is an important factor in enhancer identification (Zhu et al., 2015). Collectively, these DNase-seq studies show that the majority of open chromatin sites exist outside of genes in rice and *Arabidopsis*, that differences in open chromatin sites can be identified between tissues, and that a large proportion of intergenic open chromatin sites are in fact regulatory, at least in *Arabidopsis*. Another recent significant advance came from using DNase-seq to examine the changes in *Arabidopsis* chromatin accessibility and TF occupancy that occur during development and in response to abiotic stress (Sullivan et al., 2014). This work showed that TF-to-TF regulatory network connectivity appears to be similar between *Arabidopsis*, human, and *C. elegans*, and that such networks were extensively 'rewired' in response to stress. This study also showed that many genetic variants linked to complex traits were preferentially located in accessible chromatin regions, portending the potential for harnessing natural variation in regulatory DNA for plant breeding.

We are still left with many open questions regarding the general conservation of transcriptional regulatory

landscapes across plant genomes. For example, it remains unclear how many *cis*-regulatory elements generally exist in plant genomes, where they reside in relation to their target genes, and to what extent these features are conserved across plant genomes. Furthermore, it is not clear how the *cis*-regulatory elements within a single genome confer cell type-specific transcriptional activity – and thus cell type identity – during development. In the present study, we seek to build on previous work and to address some of these outstanding questions by analyzing chromatin accessibility across multiple, diverse plant species, and between two distinct cell types.

From a methodological perspective, the DNase-seq procedure is relatively labor-intensive and requires a large number of starting nuclei for DNaseI treatment, which can be a major drawback for conducting cell type-specific profiling investigations. More recently, the Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) was developed as an alternative approach (Buenrostro et al., 2013). ATAC-seq employs treatment of isolated nuclei with an engineered transposase that simultaneously cleaves DNA and inserts sequencing adapters, such that cleaved fragments originating from open chromatin can be converted into a high-throughput sequencing library by Polymerase Chain Reaction (PCR). Sequencing of the resulting library provides readout highly similar to that of DNase-seq, but ATAC-seq requires far fewer nuclei (Buenrostro et al., 2015). The relatively simple procedure for ATAC-seq and its low nuclei input, combined with its recent application in *Arabidopsis* and rice (Wilkins et al., 2016; Bajic et al., 2017; Lu et al., 2017), has made it widely useful for assaying plant DNA regulatory regions. In this study, we first optimized ATAC-seq for use with crude nuclei and nuclei isolated by INTACT (Isolation of Nuclei TAgged in specific Cell Types) affinity purification (Deal and Henikoff, 2010). We then applied this method to INTACT-purified root tip nuclei from *Arabidopsis thaliana*, *Medicago truncatula*, *Solanum lycopersicum* (tomato), and *Oryza sativa* (rice), as well as the root hair and non-hair epidermal cell types of *Arabidopsis*. The use of diverse plant species of both dicot and monocot lineages allowed us to assay regulatory structure over a broad range of evolutionary distances. Additionally, analysis of the *Arabidopsis* root hair and non-hair cell types allowed us to identify distinctions in chromatin accessibility that occurred during the differentiation of developmentally linked cell types from a common progenitor stem cell.

In our cross-species comparisons, we discovered that the majority of open chromatin sites in all four species exist outside of transcribed regions. The open sites also tended to cluster within several kilobases upstream of the transcription start sites despite the large differences in intergenic space between the four genomes. When orthologous genes were compared across species, we found that the number and location of open chromatin regions were highly variable, suggesting that regulatory elements are not statically positioned relative to target genes over evolutionary timescales. However, we found evidence that particular gene sets remain under control by common TFs across these species. For instance, we discovered a set of four TFs that appear to be integral for root tip transcriptional regulation of common gene sets in all species. These include HY5 and MYB77, which were previously shown to impact root development in *Arabidopsis* (Oyama et al., 1997; Shin et al., 2007).

When comparing the two *Arabidopsis* root epidermal cell types, we found that their open chromatin profiles are qualitatively very similar. However, many quantitative differences between cell types were identified, and these regions often contained binding motifs for TFs that were more highly expressed in one cell type than the other. Further analysis of several such cell type-enriched TFs led to the discovery of a hair cell transcriptional regulatory module driven by ABI5 and MYB33. These factors appear to co-regulate a number of additional hair cell-enriched TFs, including MYB44 and MYB77, which in turn regulate many downstream TF genes as well as other genes impacting hair-cell fate, physiology, secondary metabolism, and stress responses.

Overall, our work suggests that the *cis*-regulatory structure of these four plant genomes is strikingly similar, and that TF-target gene modules are also generally conserved across species. Furthermore, early differential expression of high-level TFs between the *Arabidopsis* hair and non-hair cells appears to drive a TF cascade that at least partially explains distinctions between hair and non-hair cell transcriptomes. Our data also highlight the utility of comparative chromatin profiling approaches and will be widely useful for hypothesis generation and testing.

RESULTS AND DISCUSSION

Application of ATAC-seq in *Arabidopsis* root tips

The Assay for Transposase-Accessible Chromatin (ATAC-seq) method was introduced in 2013 and has since been widely adopted in many systems (Buenrostro et al., 2013; Mo et al., 2015; Scharer et al., 2016; Lu et al., 2017). This technique utilizes a hyperactive Tn5 transposase that is pre-loaded with sequencing adapters as a probe for chromatin accessibility. When purified nuclei are treated with the transposase complex, the enzyme freely enters the nuclei and cleaves accessible DNA, both around nucleosomes and at nucleosome-depleted regions arising from the binding of transcription factors (TFs) to DNA. Upon cleavage of DNA, the transposon integrates sequencing adapters, fragmenting the DNA sample in the process. Regions of higher accessibility will be cleaved by the transposase more frequently and generate more fragments – and ultimately more reads, once the sample is sequenced. Conversely, less accessible regions will have fewer fragments and reads. After PCR-amplification of the raw DNA fragments, paired-end sequencing of the ATAC-seq library can reveal nucleosome-depleted regions where TFs are bound.

In this study, we set out to apply ATAC-seq to multiple plant species as well as different cell types from a single species. As such, we first established procedures for using the method with *Arabidopsis*, starting with root tip nuclei affinity-purified by INTACT (Isolation of Nuclei TAgged in specific Cell Types). We also established a protocol to use nuclei purified by detergent lysis of organelles followed by sucrose sedimentation, with the goal of broadening the application of ATAC-seq to non-transgenic starting tissue. We began with an *Arabidopsis* INTACT transgenic line constitutively expressing both the nuclear envelope targeting fusion protein (NTF) and biotin ligase (BirA) transgenes. Co-expression of these transgenes results in all the nuclei in the plant becoming biotinylated, and thus amenable to purification with streptavidin beads (Deal and Henikoff, 2010; Sullivan et al., 2014). Transgenic INTACT plants were grown on vertically oriented nutrient agar plates to facilitate root growth, and total nuclei were isolated from the 1 cm root tip region. These nuclei were further purified either by treatment with 1% (v/v) Triton X-100 and sedimentation through a sucrose cushion (‘Crude’ purification) or affinity-purified using streptavidin-coated magnetic beads (INTACT purification). In both cases 50,000 nuclei from each purification strategy were used as the input for ATAC-seq (Figure 3.1A). Overall, both Crude and INTACT-purified nuclei

yielded very similar results (Figure 3.1B and C, Figure S3.1). One clear difference that emerged was the number of reads that map to organellar DNA between the nuclei preparation methods. While the total reads of Crude nuclei preparations mapped approximately 50% to organellar genomes and 50% to the nuclear genome, the total reads of INTACT-purified nuclei consistently mapped over 90% to the nuclear genome (Table 3.1). The issue of organellar genomes contaminating ATAC-seq reactions is a common one, resulting in a large percentage of organelle-derived reads that must be discarded before further analysis. This issue was also recently shown to be remedied by increasing the purity of nuclei prior to ATAC-seq by use of fluorescence-activated nuclei sorting (Lu et al., 2017). To compare between datasets for the Crude and INTACT preparation strategies, we analyzed the enrichment of ATAC-seq reads using Hotspot peak mapping software (John et al., 2011). Though designed for use with DNase-seq data, Hotspot can also be readily used with ATAC-seq data. The number of enriched regions found with this algorithm did not differ greatly between nuclei preparation types, nor did the SPOT score (a signal-specificity measurement representing the proportion of sequenced reads that fall into enriched regions) (Table 3.1). These results suggest that the datasets are generally comparable regardless of the nuclei purification method.

Visualization of the Crude- and INTACT-ATAC-seq datasets in a genome browser revealed that they were highly similar to one another and to DNase-seq data from whole root tissue (Figure 3.1B). Further evidence of similarity among these datasets was found by examining the normalized read count signal in all datasets (both ATAC-seq and DNase-seq) within the regions called as 'enriched' in the INTACT-ATAC-seq dataset. For this and all subsequent peak calling in this study, we used the *findpeaks* algorithm in the HOMER package (Heinz et al., 2010), which we found to be more versatile and user-friendly than Hotspot. Using this approach, we identified 23,288 enriched regions in our INTACT-ATAC-seq data. We refer to these peaks, or enriched regions, in the ATAC-seq data as transposase hypersensitive sites (THSs). We examined the signal at these regions in the whole root DNase-seq dataset and both Crude- and INTACT-ATAC-seq datasets using heatmaps and average plots. These analyses showed that THSs detected in INTACT-ATAC-seq tended to be enriched in both Crude-ATAC-seq and DNase-seq signal (Figure 3.1C). In addition, the majority of enriched regions (19,516 of 23,288) were found to overlap between the root-tip

INTACT-ATAC-seq and the whole-root DNase-seq data (Figure 3.1D) and the signal intensity over DNase-seq or ATAC-seq enriched regions was highly correlated between the datasets (Figure S3.1).

To examine the distribution of hypersensitive sites among datasets, we identified enriched regions in both types of ATAC-seq datasets and the DNase-seq dataset, and then mapped these regions to genomic features. We found that the distribution of open chromatin regions relative to gene features was nearly indistinguishable among the datasets (Figure 3.1E). In all cases, the majority of THSs (~75%) were outside of transcribed regions, with most falling within 2 kb upstream of a transcription start site (TSS) and within 1 kb downstream of a transcript termination site (TTS).

Overall, these results show that ATAC-seq can be performed effectively using either Crude or INTACT-purified nuclei, and that the data in either case are highly comparable to that of DNase-seq. While the use of crudely purified nuclei should be widely useful for assaying any tissue of choice without a need for transgenics, it comes with the drawback that ~50% of the obtained reads will be from organellar DNA. The use of INTACT-purified nuclei greatly increases the cost efficiency of the procedure and can also provide access to specific cell types, but requires pre-established transgenic lines.

Comparison of root tip open chromatin profiles among four species

Having established an efficient procedure for using ATAC-seq on INTACT affinity-purified nuclei, we used this tool to compare the open chromatin landscapes among four different plant species. In addition to the *Arabidopsis* INTACT line described above, we also generated constitutive INTACT transgenic plants of *Medicago truncatula* (*Medicago*), *Oryza sativa* (rice), and *Solanum lycopersicum* (tomato). Seedlings of each species were grown on vertically oriented nutrient plates for one week after radicle emergence, and nuclei from the 1 cm root tip regions of each seedling were isolated and purified with streptavidin beads. ATAC-seq was performed in at least two biological replicates for each species, starting with 50,000 purified nuclei in each case. Visualization of the mapped reads across each genome showed notable consistencies in the data for all four species. In all cases, the reads localize to discrete peaks that are distributed across the genome, as expected (Figure 3.2A). Examination of a syntenic region found in all four genomes

suggested at least some degree of consistency in the patterns of transposase accessibility around orthologous genes (Figure 3.2A).

To specifically identify regions of each genome that were enriched in ATAC-seq signal (THSs), we used the HOMER *findpeaks* function on each biological replicate experiment. For further analysis, we retained only THS regions that were found in at least two biological replicates of ATAC-seq in each species. These reproducible THSs were then mapped to genomic features in each species in order to examine their distributions. As seen previously for *Arabidopsis*, the majority of THSs (~70-80%) were found outside of transcribed regions in all four species (Figure 3.2B). For this analysis, we classified these extragenic THSs (THSs found anywhere outside of transcribed regions) as proximal upstream (< 2 kb upstream of the transcription start site, or TSS), proximal downstream (< 1 kb downstream of the transcript termination site, or TTS) or intergenic (> 2 kb upstream from a TSS or > 1 kb downstream from a TTS). The proportion of THSs in the proximal upstream and intergenic regions varied greatly with genome size, and thus the amount of intergenic space in the genome. For example, a full 52% of THSs in *Arabidopsis* – the organism with the smallest genome (~120 Mb) and highest gene density of the four species – were in the proximal upstream region. This percentage drops as genome size and intergenic space increase, with 37% of the THSs in the proximal upstream region in the rice genome (~400 Mb), 30% in the *Medicago* genome (~480 Mb), and a mere 11% in the tomato genome (~820 Mb). The percentage of total THSs in the proximal downstream region followed a similar pattern, marking 17% of the THSs in *Arabidopsis*, 12% in rice and *Medicago*, and 6% in tomato. Finally, the proportion of THSs classified as intergenic followed the inverse trend as expected, with 12% of the THSs in intergenic regions for *Arabidopsis*, 30% for rice and *Medicago*, and 50% for tomato (Figure 3.2B). Thus, while the overall proportion of extragenic THSs is similar among species, the distance of these sites from genes tends to increase with genome size, which is roughly proportional to the average distance between genes.

Since the majority of THSs were found upstream of the nearest gene for each species, we next classified the regions based on their distance from the nearest TSS. We binned THSs in each genome into twelve distance categories, starting with those > 10 kb upstream of the TSS, then into eleven bins of 999 bp moving

in toward the TSS, and finally a TSS-proximal bin of 100-0 bp upstream of the TSS (Figure 3.2C). Starting with this TSS-proximal bin, we find that ~17% of the upstream THSs in *Arabidopsis*, *Medicago*, and rice are within 100 bp of the TSS, whereas 2.7% of the upstream THSs in tomato are within 100 bp of the TSS. Moving away from the TSS, we find that 91% of the total upstream THSs fall within 2.9 kb of the TSS in *Arabidopsis*, while this number decreases with genome size, with 84% for rice, 73% for *Medicago*, and 65% for tomato. In the distance bin spanning 9.9 kb to 3 kb upstream, we find 7% of the total upstream THSs in *Arabidopsis*, 15% in rice, 23% in *Medicago*, and 32% in tomato. Finally, the THSs that are more than 10 kb away from the TSS accounts for 0.8% of the total upstream THSs in *Arabidopsis*, 0.9% in rice, 2.3% in *Medicago*, and 3.3% in tomato. Overall, it is clear that in all species the majority of THSs are within 3 kb upstream of a TSS, suggesting that most *cis*-regulatory elements in these genomes are likely to be proximal to the core promoter. In the species with the largest genomes and intergenic distances (*Medicago* and tomato), THSs tend to be spread over a somewhat wider range upstream of the TSS. However, even in these cases, only a few hundred THSs in total are more than 10 kb away from the nearest gene. It is worth noting that the distribution of THSs in *Medicago* is more similar to that of tomato than rice, despite the genome size being more similar to rice. This suggests that THSs tend to be further away from TSSs in *Medicago* than would be expected based on genome size alone.

As most THSs fall near genes, we next investigated from the opposite perspective – for any given gene, how many THSs were associated with it? In this regard, we find that the *Arabidopsis*, *Medicago*, and rice genomes are highly similar (Figure 3.2D). In all three genomes, of the subset of genes that have *any* upstream THSs, ~70% of these genes have a single site, ~20% have two sites, 5-7% have three sites, and 2-3% have four or more THSs. By contrast, the tomato genome has a different trend. Of the subset of tomato genes with *any* upstream THS, only 27% of the genes have a single site, and this proportion gradually decreases with increasing THS number, with 2.7% of the tomato genes in this subset having 10 or more THSs.

Overall, we have found that THSs have similar size and genomic distribution characteristics across all four species (Table S3.1). The majority of THSs in all species are found outside of genes, mainly upstream

of the TSS, and these sites tend to cluster within 3 kb of the TSS. Furthermore, most genes with an upstream THS in *Arabidopsis*, *Medicago*, and rice have only 1-2 THSs, whereas tomato genes tend to have a larger number of upstream THSs. Whether this increase in upstream THSs in tomato is reflective of an increase in the number of regulatory elements per gene based on clade-specific alterations in gene regulation, DNA copy number changes, or simply the greater abundance of transposons and other repeat elements is not entirely clear. Compared to the other species, tomato THSs are much more abundant and tend to be smaller in size than those of the other species, and the tomato ATAC-seq data generally appear to have a lower signal-to-noise ratio (Table S3.8, Figure 3.2A). While it is unclear why the data from tomato are distinct in these ways, it is clear that tomato THSs occupy mostly genic regions of the genome, as expected, and are highly reproducible between biological replicate experiments (Figure S3.2).

Collectively, these results suggest that there is a relatively small number of regulatory elements per gene in plants. These elements tend to be focused near the promoter rather than at more distal sites as has been observed in animal, particularly mammalian, genomes (Stadhouders et al., 2012). The assumptions implicit in this argument are that open chromatin sites near a TSS reflect regulatory elements that regulate that TSS and not a more distant one, and that upstream elements contribute the majority of regulatory effects. These assumptions appear to be generally validated by many reporter assays showing that an upstream fragment of several kilobases is frequently sufficient to recapitulate native transcription patterns (Medford et al., 1991; Masucci et al., 1996; Ruzicka et al., 2007; Tittarelli et al., 2009; Li et al., 2012), as well as our observation that upstream THSs are the most abundant class of open chromatin sites.

Open chromatin features are not directly conserved among orthologous genes

Given that many of the properties of open chromatin regions were shared among *Arabidopsis*, *Medicago*, rice, and tomato, we next asked whether the numbers and locations of THSs – and thus putative regulatory elements – were conserved among orthologous genes across species. For these analyses, we identified 373 syntenic orthologs (Table S3.2) that were found in all four genomes and asked whether members of each ortholog set harbored a similar number of open chromatin regions across the species. Again, using root tip

THSs present in at least two biological replicates for each species, we counted the number of THSs within 5 kb upstream of the TSS for each ortholog in each species. We then examined these data for similarities and differences in upstream THS number (Figure 3.3A). While no clear trend of strong conservation in the number of upstream THSs emerged from this analysis, there was a small subset of orthologs that did have upstream THSs in similar numbers across species. However, this was a very small proportion of the total. As seen in earlier analyses, tomato genes tended to have a larger number of upstream THSs compared to the other species, and most of the 373 orthologs in tomato did have at least one upstream THS. This was not the case in the other three species, where many of the orthologs had no detectable upstream THSs within 5 kb of the TSS. Among the four species, *Arabidopsis* and *Medicago* showed the greatest similarity in upstream THS number, but even in this case the similarity was minimal despite the relatively closer phylogenetic relationship between these two organisms.

We next examined the distribution of open chromatin regions across the upstream regions of these 373 orthologous genes relative to their expression level in *Arabidopsis*, reasoning that there could be patterns of open chromatin similarity based on THS positions, rather than numbers. For this analysis, we examined the normalized ATAC-seq signal across the upstream region of all 373 orthologous genes, from -5000 bp to +100 bp relative to the TSS of each gene (Figure 3.3B). Orthologs were then ranked within the heatmap based on the transcript level of each *Arabidopsis* ortholog in the root tip (Li et al., 2016), from highest to lowest expression. For each *Arabidopsis* ortholog we also included the upstream THS number to ascertain how this feature might correlate with transcript level for *Arabidopsis*. While there was some consistency among species in that open chromatin often overlapped with the TSS, we did not observe any clear pattern in transposase hypersensitivity within the upstream regions of these orthologs. K-means clustering of the heatmaps similarly did not reveal evidence for conservation of open chromatin patterns among orthologs (Figure S3.3A). An important caveat to this analysis is that many of these syntenic orthologs may not be functional homologs, or 'expressologs' (Patel et al., 2012), due to subfunctionalization within gene families. As such, we identified a smaller group (52) of expressologs on which to perform a similar test (Table S3.3). While these expressolog genes have both maximally high protein level similarity and expression pattern

similarity, including expression in the root, there was also no clear correspondence in upstream THS number among them (Figure S3.3B).

There does not appear to be strong conservation in the number and location of open chromatin sites at orthologous genes across species. Assuming that these genes are still under control of common TFs, this suggests that regulatory elements could be free to migrate, and perhaps split or fuse, while retaining the regulatory parameters of the target gene in question.

One interesting finding from these analyses was that the pattern of upstream THS number does not correlate with expression level, at least for *Arabidopsis* (Figure 3.3B). Thus, THSs must not simply represent activating events upstream of the TSS but may also represent binding of repressive factors. Further, we found no correlation between upstream THS number and expression entropy among all genes in the *Arabidopsis* genome, suggesting a more complex relationship between regulatory element distribution and target gene transcription (Figure S3.3C).

Evidence for co-regulation of common gene sets by multiple TFs across species

While there does not appear to be a consistent pattern in the number or placement of open chromatin regions around orthologs or expressologs, we wanted to examine whether it would be possible to find common regulators of specific gene sets among species using a deeper level of analysis. To do this, we first searched for common TF motifs in root tip THSs across the four species. Using the THSs that were found in at least two replicates for each species, we employed the MEME-ChIP motif analysis package (Machanick and Bailey, 2011; Ma et al., 2014) to identify overrepresented motifs of known TFs. We discovered 30 motifs that were both overrepresented and common among all species (Table S3.4). We narrowed our list of candidate TFs by considering a variety of factors, including the expression of each TF in the root tip, any known mutant root phenotypes involving those TFs, and whether genome-wide binding information was available for each candidate in *Arabidopsis*. Ultimately, we selected 4 TFs for further analysis: ELONGATED HYPOCOTYL 5 (HY5), ABSCISIC ACID RESPONSIVE ELEMENTS-BINDING FACTOR 3 (ABF3), C-REPEAT/DRE BINDING FACTOR 2 (CBF2), and MYB DOMAIN PROTEIN 77

(MYB77). It is worth noting that among these factors, both HY5 and MYB77 had been previously implicated in root development (Oyama et al., 1997; Zhao et al., 2014). Like HY5 and MYB77, CBF2 and ABF3 have been implicated in stress responses as well as abscisic acid (ABA) signaling (Kang et al., 2002; Knight et al., 2004). Furthermore, overexpression of ABF3 leads to increased tolerance to multiple abiotic stresses in *Arabidopsis*, rice, cotton, and alfalfa (Oh et al., 2005; Abdeen et al., 2010; Wang et al., 2016; Kerr et al., 2017). Given this evidence, we decided to focus on these factors for further study.

We first sought to define the target genes for each of these four TFs in *Arabidopsis* by combining our chromatin accessibility data with published genome-wide binding data for each factor in *Arabidopsis* (Table 3.2). Because an accessible chromatin region (a THS) represents the displacement of nucleosomes by a DNA-binding protein, we reasoned that our THS profiles for a given tissue would represent virtually all possible protein binding sites in the epigenomes of root tip cells. Similarly, by using *in vitro* genomic binding data (DAP-seq) (O'Malley et al., 2016b) or ChIP-seq data from a highly heterogeneous tissue, we could identify the spectrum of possible binding sites for that TF, such that the intersection of these datasets would represent the binding sites for that TF in the sample of interest. While there are caveats to this approach, we reasoned that it was more likely to generate false negatives than false positives and would give us a set of high confidence target genes to analyze for each TF. In this regard, ChIP-seq data may be more robust because they represent *in vivo* binding, while DAP-seq is an *in vitro* assay and may not capture binding sites that depend on chromatin properties or interactions with other TFs. On the other hand, ChIP-seq data are inherently limited by the cell types present in the sample used.

We first tested this approach in *Arabidopsis* with each of the four TFs of interest. Using THSs from the *Arabidopsis* root tip that were found in at least two biological replicates, we used the motif-identification tool FIMO (Grant et al., 2011) to identify THSs that contained a significant occurrence of the TF motif of interest. The THSs that contained a significant motif match were considered *predicted binding sites*. We then identified predicted binding sites that also overlapped with a known binding site for that TF (a DAP-seq or ChIP-seq peak), and these were considered *high confidence binding sites* for that TF in the root tip (Figure S3.4). The predicted binding sites (motif-containing THSs) were themselves very good predictors

of the true binding sites for these four TFs (Table 3.2). For example, of the 1,316 *Arabidopsis* root tip THSs with an occurrence of the ABF3 motif (Mathelier et al., 2014), 1,279 (97%) overlapped with an ABF3 ChIP-seq peak from whole 2-day-old seedlings (Song et al., 2016). Similarly, 89% of predicted CBF2 binding sites (Weirauch et al., 2014a) overlapped with a CBF2 DAP-seq peak (O'Malley et al., 2016a), 74% of predicted MYB77 binding sites (Weirauch et al., 2014a) overlapped with a MYB77 DAP-seq peak (O'Malley et al., 2016a), and 61% of predicted HY5 binding sites (Mathelier et al., 2014) overlapped with a HY5 DAP-seq peak (O'Malley et al., 2016a). In each case, the high confidence binding sites (motif-containing THSs that overlap with a ChIP- or DAP-seq peak) were assigned to their nearest TSS in order to identify the putative target genes for each TF (Figure S3.4).

With these lists of target genes for each TF in the *Arabidopsis* root tip, we looked for gene sets that were regulated by more than one factor, as means of identifying co-regulatory associations between these four TFs. We found extensive co-targeting among these four TFs, with gene sets being targeted by one, two, three, or all four of these TFs to a degree that was far higher than what would be expected by chance (Figure 3.3C). For example, of the 1,271 ABF3 target genes, 297 (23%) are also targeted by HY5 (hypergeometric $p = 2.1 \times 10^{-56}$). Among these 297 genes, 46 are targeted by ABF3, HY5, and CBF2, and seven are targeted by all four TFs. We also asked where the binding sites driving this pattern were located relative to the target genes. To do this we considered only binding sites within the 5 kb upstream region of a TSS, and repeated the target gene assignment and analysis of target gene overlaps between TFs. This subsetting reduced the total number of target genes for each factor by ~20%, but did not substantially alter the percentages of target gene overlap among the four TFs (Figure S3.5A). These results collectively suggest that these four TFs have important roles in root tip gene regulation both individually and in combination, and that the majority of their binding sites (~80%) fall within the 5 kb region upstream of the TSS for target genes. In addition, we find that the binding sites for multiple TFs often occur in the same THS (Figure S3.5B).

We next sought to examine the target genes and proportions of target gene overlaps between the four species to address the conservation of co-regulatory relationships among these four TFs. Given that no TF binding data is available for the other three species and knowing that the majority of our predicted binding

sites in *Arabidopsis* corresponded to known binding sites (Table 3.2; 61-97%), we opted to also use the predicted binding sites for each of the four TFs in *Medicago*, tomato, and rice, with the knowledge that these sets may contain some false positives. For these analyses we used the *Arabidopsis* TF motifs – since these have not been directly defined for the other species – with the caveat that the DNA binding specificity of these factors may not be identical among species.

We again used FIMO to identify significant occurrences of each TF motif within the root tip THSs found in at least two biological replicates for each of our four species. We then mapped the predicted binding sites of each TF to the nearest TSS to define target genes for each TF in each species (Table S3.5). We then analyzed the overlap of TFs at target genes in each species using 4-way Venn diagrams, similar to Figure 3.3C. To compare regulatory associations across species, we considered each of the 15 categories in every species-specific 4-way Venn diagram as a regulatory category. For example, one regulatory category consists of the genes targeted only by ABF3 alone, another would be those targeted only by HY5 and ABF3 at the exclusion of the other two TFs, and so on. For each regulatory category in each species, we calculated the percentage of the total target genes in that category (number of genes in the regulatory category/total number of genes targeted by *any* of the 4 TFs), and then compared these percentages between species (Figure 3.3D). We found remarkably consistent proportions of the target genes in nearly all regulatory categories across all four species. However, notable deviations from this consistency among species were seen in the proportion of rice genes targeted by MYB77 alone and rice genes targeted CBF2 and HY5 together. In most cases, the proportions of target genes in different regulatory categories were most similar between *Arabidopsis* and *Medicago*, and these were generally more similar to tomato than to rice, consistent with the evolutionary distances between the species (Vanneste et al., 2014). Commonly overrepresented Gene Ontology (GO) terms among gene sets in particular regulatory categories across species further support the notion of regulatory conservation (Figure S3.5C), although these analyses are limited by the depth of GO annotation in some of these species.

These findings suggest that while neither syntenic orthologous gene sets nor expressolog gene sets tend to share open chromatin patterns, the genes under control of specific TFs or specific combinations of TFs

appear to be relatively stable over evolutionary time, at least for the four TFs we examined. One simple explanation for this phenomenon is that the locations of transcriptional regulatory elements are somewhat malleable over time as long as proper transcriptional control is maintained. In this model, these elements would be free to relocate in either direction, and potentially even merge or split. This would maintain proper control over the target gene, but give each ortholog or expressolog a unique chromatin accessibility profile depending on the exact morphology and distribution of the functionally conserved regulatory elements. This idea of modularity is consistent with previous observations that the *Drosophila* even-skipped stripe 2 enhancer can be rearranged and still retain functionality (Ludwig et al., 2000; Ludwig et al., 2005).

The results also shed light on the interconnectedness of specific TFs in root tip cells and indicate durability of these co-regulatory relationships over time. They also generate readily testable hypotheses regarding the how HY5, ABF3, MYB77, and CBF2 operate during root development. For example, given that HY5 appears to regulate over 1,000 genes in the *Arabidopsis* root tip (Figure 3.3C), and that hundreds of these are annotated with GO terms including *biological regulation* and *response to stimulus*, we predict that *hy5* mutants would have defects in root tip morphology and growth. Indeed, HY5 was previously shown to be involved in the regulation of lateral root growth initiation and gravitropism (Oyama et al., 1997), and we observe that the primary root tips in *hy5* mutants also frequently show a bulging and malformed appearance, as well as severe gravitropism defects (Figure S3.6).

Commonalities and distinctions in the open chromatin landscapes of *Arabidopsis* root epidermal cell types

Having examined questions of regulatory conservation between species, we then explored regulatory elements and TFs relationships between cell types within a single species. In this case, we chose to focus on the root epidermal hair and non-hair cell types in *Arabidopsis*. Since these two cell types are derived from a common progenitor, they are prime candidates to offer insight into the epigenomic alterations that occur during – and likely drive – cell differentiation. Specifically, we investigated to what extent the open chromatin landscapes would differ between cell types and whether differences in THSs could pinpoint the

sites of differential transcriptional regulation. Furthermore, we wanted to understand whether we could use this information to examine the TF-to-TF regulatory connections that underlie the transcriptomic and physiological differences between these cell types.

We used two previously described INTACT transgenic lines as starting material for these experiments: one having biotin-labeled nuclei exclusively in the root hair (H) cells, and another with labeled nuclei only in the root epidermal non-hair (NH) cells (Deal and Henikoff, 2010). Nuclei were purified from each fully differentiated cell type by INTACT, and 50,000 nuclei of each type were subjected to ATAC-seq. Visualization of these cell type-specific datasets in a genome browser, along with the *Arabidopsis* whole 1 cm root tip ATAC-seq data, showed a high overall degree of similarity among the three datasets (Figure 3.4A). Comparison of the ATAC-seq signal intensity at common THS regions genome-wide revealed that these two cell types have open chromatin patterns that are highly similar to one another, but distinct from that of the whole root tip (Figure S3.7).

To identify regions of differential accessibility between the cell types and the whole root tip, we considered THS regions that were found in at least two biological replicates of each cell type or tissue. The total number of these reproducible THSs was 32,942 in the whole root tip, 35,552 for the H cells, and 28,912 for the NH cells. The majority of these sites (18,742) were common (overlapping) in all three sample types (Figure 3.4B) and thus likely represent regulatory sites that are utilized in multiple *Arabidopsis* root cell types. We also found 6,562 THSs that were common to both root epidermal cell types but were not found in the whole root tip, suggesting that these may represent epidermal-specific regulatory elements. In a search for unique THSs in each of the three sample types (those not overlapping with a THS in any other sample), we found 10,455 THSs that were unique to the whole root tip, 7,537 unique to the H cells, and 2,574 that were unique to the NH cells. We refer to these regions as differential THSs (dTHSs). The dTHSs identified only in the H or NH cell type were of further interest because they may represent regulatory elements that drive the transcriptomic differences between these two epidermal cell types.

To examine the extent of chromatin accessibility differences at these dTHSs, we visualized the accessibility signals from each cell type at both H cell dTHSs and NH cell dTHSs. First, using the 7,537

regions identified as H cell dTHSs, we used heatmaps and average plots to examine the normalized ATAC-seq read count across these regions in each cell type (Figure 3.4C, left panel). We then repeated this analysis using the 2,574 NH cell dTHSs (Figure 3.4C, right panel). In each case, it was clear that the regions we identified as dTHSs showed significant differences in chromatin accessibility between the two cell types. However, the differences in chromatin accessibility between cell types were quantitative (varying intensity) rather than qualitative (all-or-nothing). This indicates that, at large, the dTHSs represent sites that are highly accessible in one cell type and less so in the other, rather than being strictly present in one and absent in the other. Therefore, we refer to these sites from this point on as cell type-enriched dTHSs to convey the notion of quantitative differences between cell types.

To identify the genes that might be impacted by cell type-enriched dTHSs, we mapped each dTHS to its nearest TSS and considered that to be the target gene. We found that the 7,537 H-enriched dTHSs mapped to 6,008 genes, while the 2,574 NH-enriched dTHSs mapped to 2,295 genes. Thus, the majority of genes that are associated with a dTHS are only associated with one such site. This is consistent with our previous findings that most *Arabidopsis* genes are associated with a single upstream THS (Figure 3.2D).

We then asked how the set of genes associated with dTHSs overlapped with those whose transcripts that show differential abundance between the two cell types. Using data from a recent comprehensive RNA-seq analysis of flow sorted *Arabidopsis* root cell types (Li et al., 2016a), we identified sets of transcripts that were more highly expressed in H versus NH cell types. To be considered a *cell type-enriched gene*, we required a gene to have a transcript level with two-fold or greater difference in abundance between H and NH cell types, as well as at least five reads per kilobase per million mapped reads (RPKM) in the cell type with a higher transcript level. Using this relatively conservative approach, we derived a list of 3,282 H cell-enriched genes and 2,731 NH cell-enriched genes. We then asked whether the genes associated with cell type-enriched dTHSs were also cell type-enriched genes (Figure 3.4D). Of the 3,282 H cell-enriched genes, 743 were associated with a H cell-enriched dTHS, 258 were associated with a NH cell-enriched dTHS, and 108 genes were associated with a dTHS in both cell types. Among the 2,731 NH cell-enriched genes, 156 were associated with a NH cell-enriched dTHS, 516 were associated with a H cell-enriched dTHS, and 52

genes showed dTHSs in both cell types. These results suggest that cell type-enriched expression of a gene is frequently associated with a dTHS in the cell type where the gene is highly expressed, but is also often associated with a dTHS in the cell type where that gene is repressed. This highlights the importance of transcriptional activating events in the former case and repressive events in the latter. Interestingly, for a smaller set of cell type-enriched genes we observed dTHSs at a given gene in both cell types, indicating regulatory activity at the gene in both cell types.

We next asked what proportion of the transcriptome differences between H and NH cells might be explained based on differential chromatin accessibility. Of the 3,282 H cell-enriched genes, 1,109 have a dTHS in one or both of the cell types, and among the 2,731 NH cell-specific genes, 724 have a dTHS in one or both cell types. Assuming that each dTHS represents a regulatory event contributing to the differential expression of its identified target gene, we could explain differential expression of 33% of the H cell-enriched genes and 27% of the NH cell-enriched genes. The remaining ~70% of the identified cell type-enriched genes without clear chromatin accessibility differences may be explained in numerous ways. These genes may not require a change in chromatin accessibility, changes in chromatin accessibility may fall below our limit of detection, or these transcripts may be primarily regulated at the post-transcriptional level rather than at the chromatin-accessibility level that we measured.

Another key question relates to the significance of the cell-type-enriched dTHSs that do not map to differentially expressed genes. These could be explained by an inability to detect all differentially expressed genes, perhaps simply due to the stringency of our definition of cell type-enriched genes. An important biological possibility to consider is that many of these regulatory regions do not in fact regulate the closest gene, but rather act over a distance such that they are orphaned from their true target genes in our analysis. Another possibility is that many of the differential protein binding events represented by these dTHSs are unrelated to transcriptional regulation.

Overall, the accessible chromatin landscapes of the root epidermal H and NH cells appear to be nearly identical in a qualitative sense, but differ significantly at several thousand sites in each cell type. The reasons for the quantitative, rather than all-or-nothing, nature of this phenomenon are not entirely clear. Are the

accessibility differences between cell types reflective of unique protein assemblages at the same element in different cell types, or do they instead reflect differences in abundance of the same proteins at an element in different cell types? While these questions certainly warrant further investigation and experimentation, we can gain further insight into the regulatory differences between cell types through deeper examination of the differentially accessible chromatin regions in each.

TF motifs in cell type-specific THSs identify regulators and their target genes

As a means of identifying specific transcription factors (TFs) that might be important in specifying the H and NH cell fates, we sought to identify overrepresented motifs in the differentially accessible regions of each cell type. We used each set of cell type-enriched dTHSs as input for MEME-ChIP analyses (Machanick and Bailey, 2011) and examined the resulting lists of overrepresented motifs. We initially found 219 motifs that were significantly overrepresented relative to genomic background only in H cell-enriched dTHSs and 12 that were significantly overrepresented only in NH cell-enriched dTHSs (Table S3.6). In order to narrow our list of candidate TFs to pursue, we vetted these lists of potential cell type-enriched TFs by considering their transcript levels in each cell type as well as the availability of genome-wide binding data. Based on the available data, we narrowed our search to five transcription factors of interest: four H cell-enriched TF genes (MYB33, ABI5, NAC083, and At5g04390) and one NH-enriched TF gene (WRKY27) (Table 3.3).

We next attempted to directly identify the binding sites for each TF by differential ATAC-seq footprinting between the cell types. The logic behind this approach is the same as that for DNase-seq footprinting – that the regions around a TF binding site are hypersensitive to the nuclease or transposase due to nucleosome displacement, but the sites of physical contact between the TF and DNA will be protected from transposon insertion/cutting, and thus leave behind a characteristic "footprint" of reduced accessibility on a background of high accessibility (Hesselberth et al., 2009; Vierstra and Stamatoyannopoulos, 2016). We reasoned that we could identify binding sites for each of these cell type-enriched TFs by comparing the footprint signal at each predicted binding site (a motif occurrence within a THS) between H and NH cells.

For this analysis, we examined the transposase integration patterns around the motifs of each TF in both cell types as well as in purified genomic DNA subjected to ATAC-seq, to control for transposase sequence bias. It was recently reported in *Arabidopsis* that many TF motifs exhibit conspicuous transposase integration bias on naked DNA (Lu et al., 2017), and our results were in line with these findings for all five TFs of interest here (Figure S3.8). While we observed footprint-like patterns in the motif-containing THSs in our ATAC-seq data, these patterns in each case were also evident on purified genomic DNA. As such, it was not possible to distinguish true binding sites from these data, as any footprint signal arising from TF binding was already obscured by the transposase integration bias. For unknown reasons, many TF motif DNA sequences seem to inherently evoke hyper- and/or hypo-integration by the transposase, and this automatically obscures any potentially informative footprint signal that could be obtained by integration during ATAC-seq on nuclei. Similar technical concerns have also been raised for DNaseI footprinting (Sung et al., 2016). These results suggest that the ATAC-seq footprinting approach may be useful for certain TFs, but these will likely need to be examined on a case-by-case basis. Given this issue and the resulting lack of evidence for footprints of our TFs of interest, we decided to take the approach of defining TF target sites as we did for our studies of root tip TFs.

As described earlier, we defined high confidence binding sites for the 5 TFs of interest as TF motif-containing THSs in the cell type of interest (predicted binding sites) that *also* overlapped with an enriched region for the TF in publicly available DAP-seq data (O'Malley et al., 2016a) or ChIP-seq data (Figure S3.4). Assigning these high confidence binding sites to their nearest TSS allowed us to define thousands of target genes for these factors in the root epidermal cell types (Table 3.3 and Table S3.7). Compared to our analysis of root tip TFs, our capability to predict target sites based on motif occurrences in THSs was much reduced for the four H cell-enriched and one NH cell-enriched TFs examined here. For further analyses, we decide to focus on three of the TFs that were more highly expressed in the H cell type and had the largest number of high confidence target genes: ABI5, MYB33, and NAC083.

We first asked how many of the high confidence target genes for these TFs were also preferentially expressed in one cell type or the other. We found that for all three TFs, a large percentage of the total target

genes are H cell-enriched in their expression (17-21%), while many others are NH cell-enriched (6-9%) (Figure 3.5A). These results are intriguing as they suggest that the activities of these TFs may be generally context-dependent. At the same time however, the majority of the target genes for each TF were not more highly expressed in one cell type compared to the other.

Each of these H cell-enriched TFs could activate other H cell-enriched genes, but what are their functions at regulatory elements near genes that are expressed at low levels in the H cell and high levels in the NH cell? One possibility is that these factors are activators of transcription in the context of H cell-enriched genes but act as repressors or are neutral toward the target genes that are NH cell-enriched in their expression. This may reflect context-dependency in the sense that the effect on transcription of a target gene may depend on the local milieu of other factors.

We next examined whether ABI5, MYB33, and NAC083 target any of the same genes. Similar to the root tip TFs examined previously, we found that these three TFs also appear to have extensive co-regulatory relationships (Figure 3.5B). For example, 207 target genes were shared between ABI5 and NAC083, 238 were shared between ABI5 and MYB33, and 50 target genes were shared by all 3 factors. We further analyzed the genes that were co-targeted by ABI5 and MYB33, finding that 57 of the co-targeted genes were H-cell enriched. As such, we performed Gene Ontology (GO) analysis on the H cell-enriched targets as well as the full set of target genes to gain insight into the functions of this co-regulatory relationship (Figure 3.5C). Many of the ABI5/MYB33 target genes were annotated as being involved in responses to ABA as well as water, salt, and cold stress. This is consistent with the known roles of these proteins in ABA signaling (Finkelstein and Lynch, 2000; Reyes and Chua, 2007). Interesting, seven of the 57 ABI5/MYB33 target genes that were H cell-enriched were also annotated with the term *regulation of transcription*, suggesting that ABI5 and MYB33 may be at the apex of a transcriptional regulatory cascade in the H cell type.

Identification of a new regulatory module in the root hair cell type

Based on our findings that ABI5 and MYB33 co-target seven H cell-enriched TFs, we decided to investigate this potential pathway further. Among the seven TFs putatively co-regulated by ABI5 and MYB33 and having H cell-enriched transcript expression were DEAR5, ERF11, At3g49930, SCL8, NAC087, and two additional MYB factors: MYB44 and MYB77. Aside from MYB77, none of these TFs had been previously reported to produce root-specific phenotypes when mutated. MYB77 was previously shown to interact with Auxin Response Factors (ARFs) (Shin et al., 2007) and to be involved in lateral root development through promotion of auxin-responsive gene expression (Shin et al., 2007). Interestingly, the ABA receptor, PYL8, was shown to physically interact with both MYB77 and MYB44, and to promote auxin-responsive transcription by MYB77 (Zhao et al., 2014). MYB44 has also been implicated in ABA signaling through direct interaction with an additional ABA receptor, PYL9 (Li et al., 2014), as well as repression of jasmonic acid (JA)-responsive transcription (Jung et al., 2010). These factors have additionally been implicated in salicylic acid (SA) and ethylene signaling (Yanhui et al., 2006; Shim et al., 2013). Given that MYB44 and MYB77 are paralogs (Dubos et al., 2010) that appear to integrate multiple hormone response pathways in a partly redundant manner (Jaradat et al., 2013), we decided to identify high confidence target genes (Figure S3.4) for each of them for further study.

We again defined high confidence binding sites as THSs in H cells that contain a significant motif occurrence for the factor and also overlap with a DAP-seq or ChIP-seq enriched region for that factor. Using this approach, we found that MYB44 and MYB77 each target over 1,000 genes individually and co-target 483 genes (Figure 3.6A). In addition, MYB44 and MYB77 appear to regulate one another, while MYB77 also appears to target itself. This feature of self-reinforcing co-regulation could serve as an amplifying and sustaining mechanism to maintain the activity of this module once activated by ABI5, MYB33, and potentially other upstream factors.

To gain a deeper understanding of the impact of MYB44 and MYB77 on downstream processes, we performed Gene Ontology (GO) analysis of the target genes for each factor. First considering all target genes, regardless of their expression in the H cell type, we found a variety of overrepresented GO terms for each that were consistent with the known roles of these factors in hormone signaling (Figure 3.6B). For

example, both factors targeted a large number of genes annotated with the terms *response to ABA stimulus*, *response to ethylene stimulus*, and *response to SA stimulus*. Additionally, MYB44 alone targeted many genes with the annotation *response to JA stimulus*, consistent with its previously reported role as a negative regulator of JA signaling (Jung et al., 2010). Interestingly, the largest overrepresented gene functional category for both factors was *transcription factor activity* (102 genes for MYB77 and 183 genes for MYB44). This indeed further suggests that these factors initiate a cascade of transcriptional effects. The next-largest overrepresented term was *plasmodesma*, indicating that production and/or regulation of cell-cell connecting structures are likely controlled by these factors. Plasmodesmata are important for numerous epidermal functions including cell-to-cell movement of TFs such as CPC and TRY (Schellmann et al., 2002; Wada et al., 2002) and transport of other macromolecules and metabolites (Lucas and Lee, 2004).

We also analyzed overrepresented ontology terms in the MYB77 and MYB44 targets that were classified as H cell-enriched genes. Among the MYB77 target genes in this category were known regulators of H cell fate, while numerous H cell-enriched MYB44 target genes were annotated as being involved in response to water and phosphate starvation (Figure 3.6C). The ontology category that was overrepresented in both target lists was *negative regulation of transcription* (6 MYB77 targets and 7 MYB44 targets), suggesting that these factors exert additional specific effects on the H cell transcriptome by regulating a subset of potentially repressive TFs.

The fact that MYB77 and MYB44 target a large number of genes that show H cell-enriched expression suggests that these factors serve as activators of transcription at these targets, and this is supported by published accounts of transcriptional control by these factors (Persak and Pitzschke, 2014). However, both factors also target NH cell-enriched genes as well as genes without preferential expression between the cell types. This phenomenon was also observed for the H-enriched TFs ABI5, MYB33, and NAC083 (Figure 3.5), suggesting that certain TFs may generally serve as activators but may also have context-dependent repressive functions. Such a functional switch could occur through direct mechanisms such as structural alteration by alternative splicing or post-translational modification, functional alteration by partnering with a specific TF or chromatin-modifying complex, or perhaps indirectly by binding to a target site to occlude

the binding of other factors necessary for transcriptional activation. The numerous reports of dual function transcription factors in animals and plants support the notion that this may be a general phenomenon (Ikeda et al., 2009; Boyle and Despres, 2010; Li et al., 2016b).

Collectively these results suggest that the MYB44/MYB77 module in the H cell specifies a cascade of downstream transcriptional regulation, some of which is positive and some of which is negative. This module likely represents an important hub in controlling H cell fate as well as a variety of physiological functions and environmental responses in this cell type. The fact that MYB77 was also discovered in our analyses of root tip TFs suggests that this factor likely has a broader role in other cell types during early root development, in addition to a role in specification of the H cell versus the NH cell fate. An important next step will be to perform genetic manipulations of these factors (knockout and inducible overexpression, for example), in order to test and elaborate on the specific predictions made by our model.

SUMMARY AND CONCLUSIONS

In this study, we used ATAC-seq profiling of accessible chromatin to investigate questions regarding the transcriptional regulatory landscape of plant genomes and its conservation across species. We also investigated the similarities and differences in open chromatin landscapes in two root cell types that arise from a common progenitor, allowing us to identify and analyze TFs that act specifically in one cell type versus the other. Overall, we are able to gain several new insights from this work.

In optimization of our ATAC-seq procedures, we found that the assay can be performed effectively on crudely purified nuclei but that this approach is limited by the large proportion of reads arising from organelle genomes (Table 3.1). This issue is ameliorated by the use of the INTACT system to affinity-purify nuclei for ATAC-seq, which also provides access to individual cell types. Consistent with previous reports, we found that the data derived from ATAC-seq are highly similar to those from DNase-seq (Figure 3.1). In comparing our root tip ATAC-seq data to DNase-seq data from whole roots, we found that some hypersensitive regions were detected in one assay but not the other. This discrepancy is most likely attributable to differences in starting tissue and laboratory conditions, rather than biological differences in

the chromatin regions sensitive to DNaseI versus the hyperactive Tn5 transposase. This interpretation would fit with the large number of differences also observed in THS overlap between *Arabidopsis* root tip and epidermal cell types.

In a comparison of open chromatin among the root tip epigenomes of *Arabidopsis*, *Medicago*, tomato, and rice, we found the genomic distribution of THSs in each were highly similar. About 75% of THSs lie outside of transcribed regions, and the majority of these THSs are found within 3 kb upstream of the TSS in all species (Figure 3.2). Thus, the distance of upstream THSs from the TSS is relatively consistent among species and is not directly proportional to genome size or intergenic space for these representative plant species. Among genes with an upstream THS, 70% of these genes in *Arabidopsis*, *Medicago*, and rice have a single such feature, 20% have two upstream THSs, and less than 10% have three or more. In contrast, only 27% of tomato genes with an upstream THS have a single THS, 20% have two, and the proportion with 4-10 THSs is 2-7 times higher than that for any other species examined. This increase in THS number in tomato could be reflective of an increase in the number of regulatory elements per gene, but is perhaps more likely a result of the greater number of long-terminal repeat retrotransposons near genes in this species (Xu and Du, 2014). In either case, our investigation revealed that open chromatin sites – and by extension transcriptional regulatory elements – in all four species are focused in the TSS-proximal upstream regions and are relatively few in number per gene. This suggests that transcriptional regulatory elements in plants are generally fewer in number and are closer to the genes they regulate than those of animal genomes. For example, the median distance from an enhancer to its target TSSs in *Drosophila* was found to be 10 kb, and it was estimated that each gene had an average of four enhancers (Kvon et al., 2014). It was also recently reported that in human T cells, the median distance between enhancers and promoters was 130 kb, far greater than the distances we have observed here across plant species (Mumbach et al., 2017).

Analysis of over-represented TF motifs in THSs across species suggested that many of the same TFs are at play in early root development in all species. Perhaps more surprisingly, co-regulation of specific gene sets by multiple TFs seems to be frequently maintained across species (Figure 3.3). Taken together with the lack of shared open chromatin profiles among orthologous genes and expressologs, these findings suggest

that transcriptional regulatory elements may relocate over evolutionary time within a window of several kilobases upstream of the TSS, but regulatory control by specific TFs is relatively stable.

Our comparison of the two *Arabidopsis* root epidermal cell types, the hair (H) and non-hair (NH) cells, revealed that open chromatin profiles were highly similar between cell types. By examining THSs that were exclusive to one cell type, we were able to find several thousand THSs that were quantitatively more accessible in each cell type compared to the other (Figure 3.4). Mapping of these differential THSs (dTHSs) to their nearest genes revealed that in each cell type there were many dTHSs that were near genes expressed more abundantly in that cell type, as well as many near genes with the opposite expression pattern. This suggests that some dTHSs represented transcriptional activating events whereas others were repressive in nature.

Analysis of TF motifs at these dTHSs between cell types identified a suite of TFs that were more highly expressed in H cells and whose motifs were significantly overrepresented in H cell-enriched dTHSs. Analysis of three of these TFs – ABI5, MYB33, and NAC083 – revealed that each factor targets a large number of H cell-enriched genes as well as a smaller number of NH cell-enriched genes (Figure 3.5). These factors also have many overlapping target genes among them, and ABI5 and MYB33 both target seven additional H cell-enriched TFs. Among these seven H-enriched TFs are two additional MYB factors: MYB77 and MYB44 (Figure 3.6). Examination of the high confidence target genes of MYB77 and MYB44 revealed that these paralogous factors appeared to regulate each other as well as many other common target genes, including large numbers of other TF genes. Hundreds of the MYB77 and MYB44 target genes were also more highly expressed in the H cell relative to the NH cell, suggesting that these factors set off a broad transcriptional cascade in the H cell type. In addition, they appear to directly regulate many H cell-enriched genes involved in cell fate specification and water and phosphate acquisition. This type of cooperative action by pairs of MYB paralogs has also been documented recently in *Arabidopsis* and other species (Millar and Gubler, 2005; Matus et al., 2017; Wang et al., 2017), and the fact that many target genes for each MYB factor are not regulated by the other may reflect a degree of subfunctionalization between the paralogs.

An important question arising from our results is whether classifying a TF as strictly an activator or repressor is generally accurate in most cases. For example, the H cell-enriched TFs that we examined all have apparent target genes that are highly expressed in the H cell type as well as targets that are expressed at very low levels, if at all, in the H cell type. In fact, these latter genes are often much more highly expressed in the NH cell type. Given that a number of these TFs have been shown to activate transcription in specific cases, this suggests that they promote the transcription of H cell-enriched targets and either repress or have no effect on NH cell-enriched target genes. One explanation for this phenomenon is that these TFs have "dual functionality" as activators and repressors, depending on the context (Bauer et al., 2010). However, it is equally possible that these factors do not play a direct role in gene repression. For example, the binding of an activator near a repressed gene may be functionally irrelevant to the regulation of that gene, or it may be the case that other gene-specific repressors may also be bound nearby and override the activity of the activator. This phenomenon will be worth exploring as it may deepen our understanding of the intricacies of transcriptional control.

In this study, we outline a widely applicable approach for combining chromatin accessibility profiling with available genome-wide binding data to construct models of TF regulatory networks. The putative TF regulatory pathways we have illuminated through our comparison across species and cell types provide important hypotheses regarding the evolution of gene regulatory mechanisms in plants and the mechanisms of cell fate specification, that are now open to experimental analysis.

METHODS

Plant materials and growth conditions

Plants used in this study were of the *Arabidopsis thaliana* Col-0 ecotype, the A17 ecotype of *Medicago truncatula*, the M82 LA3475 cultivar of tomato (*Solanum lycopersicum*), and the Nipponbare cultivar of rice (*Oryza sativa*). Transgenic plants of each species for INTACT were produced by transformation with a binary vector carrying both a constitutively expressed biotin ligase and constitutively expressed nuclear

tagging fusion protein (NTF) containing a nuclear outer membrane association domain (Ron et al., 2014). The binary vector used for *Medicago* was identical to the tomato vector (Ron et al., 2014), but was constructed in a pB7WG vector containing phosphinothricin resistance gene for plant selection and it retains the original *AtACT2p* promoter. The binary vector used for rice is described in Reynoso *et al.* (submitted). Transformation of rice was carried out at UC Riverside and tomato transformation was carried out at the UC Davis plant transformation facility. *Arabidopsis* plants were transformed by the floral dip method (Clough and Bent, 1998) and composite transgenic *Medicago* plants were produced according to established procedures (Limpens et al., 2004).

For root tip chromatin studies, constitutive INTACT transgenic plant seeds were surface sterilized and sown on ½-strength Murashige and Skoog (MS) media (Murashige and Skoog, 1962) with 1% (w/v) sucrose in 150 mm diameter Petri plates, except for tomato and rice where full-strength MS with 1% (w/v) sucrose and without vitamins was used. Seedlings were grown on vertically oriented plates in controlled growth chambers for 7 days after germination, at which point the 1 cm root tips were harvested and frozen immediately in liquid N_2 for subsequent nuclei isolation. The growth temperature and light intensity was 20°C and 200 μmol/m^2/sec for *Arabidopsis* and *Medicago*, 23°C and 80 μmol/m^2/sec for tomato, and 28°C/25°C day/night and 110 μmol/m^2/sec for rice. Light cycles were 16 h light/8 h dark for all species.

For studies of the *Arabidopsis* root hair and non-hair cell types, previously described INTACT transgenic lines were used (Deal and Henikoff, 2010). These lines are in the Col-0 background and carry a constitutively expressed biotin ligase gene (*ACT2p:BirA*) and a transgene conferring cell type-specific expression of the NTF gene (from the *GLABRA2* promoter in non-hair cells or the *ACTIN DEPOLYMERIZING FACTOR8* promoter in root hair cells). Plants were grown vertically on plates as described above for 7 days, at which point 1.25 cm segments from within the fully differentiated cell zone were harvested and flash frozen in liquid N_2. This segment of the root contains only fully differentiated cells and excludes the root tip below and any lateral roots above.

Nuclei isolation

For comparison of ATAC-seq using crude and INTACT-purified *Arabidopsis* nuclei, a constitutive INTACT line was used (*ACT2p:BirA*/*UBQ10p:NTF*) (Sullivan et al., 2014) and nuclei were isolated as described previously (Bajic et al., 2017). In short, after growth and harvesting as described above, 1-3 g of root tips were ground to a powder in liquid N_2 in a mortar and pestle and then resuspended in 10 ml of NPB (20 mM MOPS [pH 7], 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine, 0.2 mM spermine, 1× Roche Complete protease inhibitors) with further grinding. This suspension was then filtered through a 70 μM cell strainer and centrifuged at 1,200 x g for 10 min at 4° C. After decanting, the nuclei pellet was resuspended in 1 ml of NPB and split into two 0.5 ml fractions in new tubes. Nuclei from one fraction were purified by INTACT using streptavidin-coated magnetic beads as previously described (Bajic et al., 2017) and kept on ice prior to counting and subsequent transposase integration reaction. Nuclei from the other fraction were purified by non-ionic detergent lysis of organelles and sucrose sedimentation, as previously described (Bajic et al., 2017). Briefly, these nuclei in 0.5 ml of NPB were pelleted at 1,200 x g for 10 min at 4° C, decanted, and resuspended thoroughly in 1 ml of cold EB2 (0.25 M sucrose, 10 mM Tris [pH 8], 10 mM $MgCl_2$, 1% Triton X-100, and 1× Roche Complete protease inhibitors). Nuclei were then pelleted at 1,200 x g for 10 min at 4° C, decanted, and resuspended in 300 μl of EB3 (1.7 M sucrose, 10 mM Tris [pH 8], 2 mM $MgCl_2$, 0.15% Triton X-100, and 1× Roche Complete protease inhibitors). This suspension was then layered gently on top of 300 μl of fresh EB3 in a 1.5 ml tube and centrifuged at 16,000 x g for 10 minutes at 4° C. Pelleted nuclei were then resuspended in 1 ml of cold NPB and kept on ice prior to counting and transposase integration.

For INTACT purification of total nuclei from root tips of *Medicago*, tomato and rice, as well as purification of *Arabidopsis* root hair and non-hair cell nuclei, 1-3 g of starting tissue was used. In all cases, nuclei were purified by INTACT and nuclei yields were quantified as described previously (Bajic et al., 2017).

Assay for transposase-accessible chromatin with sequencing (ATAC-seq)

Freshly purified nuclei to be used for ATAC-seq were kept on ice prior to the transposase integration reaction and never frozen. Transposase integration reactions and sequencing library preparations were then carried out as previously described (Bajic et al., 2017). In brief, 50,000 purified nuclei or 50 ng of *Arabidopsis* leaf genomic DNA were used in each 50 µl transposase integration reaction for 30 min at 37° C using Nextera reagents (Illumina, FC-121-1030). DNA fragments were purified using the Minelute PCR purification kit (Qiagen), eluted in 11 µl of elution buffer, and the entirety of each sample was then amplified using High Fidelity PCR Mix (NEB) and custom barcoded primers for 9-12 total PCR cycles. These amplified ATAC-seq libraries were purified using AMPure XP beads (Beckman Coulter), quantified by qPCR with the NEBNext Library Quantification Kit (NEB), and analyzed on a Bioanalyzer High Sensitivity DNA Chip (Agilent) prior to pooling and sequencing.

High throughput sequencing

Sequencing was carried out using the Illumina NextSeq 500 or HiSeq2000 instrument at the Georgia Genomics Facility at the University of Georgia. Sequencing reads were either single-end 50 nt or paired-end 36 nt and all libraries that were to be directly compared were pooled and sequenced on the same flow cell.

Sequence read mapping, processing, and visualization

Sequencing reads were mapped to their corresponding genome of origin using Bowtie2 software (Langmead and Salzberg, 2012) with default parameters. Genome builds used in this study were *Arabidopsis* version TAIR10, *Medicago* version Mt4.0, Tomato version SL2.4, and Rice version IRGSP 1.0.30. Mapped reads in *.sam* format were converted to *.bam* format and sorted using Samtools 0.1.19 (Li et al., 2009). Mapped reads were then filtered using Samtools to retain only those reads with a mapping quality score of 2 or higher (Samtools "*view*" command with option "*-q 2*" to set mapping quality cutoff). *Arabidopsis* ATAC-seq reads were further filtered with Samtools to remove those mapping to either the

chloroplast or mitochondrial genomes, and root hair and non-hair cell datasets were also subsampled such that the experiments within a biological replicate had the same number of mapped reads prior to further analysis. For normalization and visualization, the filtered, sorted *.bam* files were converted to bigwig format using the "*bamcoverage*" script in deepTools 2.0 (Ramirez et al., 2016) with a bin size of 1 bp and RPKM normalization. Use of the term *normalization* in this paper refers to this process. Heatmaps and average plots displaying ATAC-seq data were also generated using the "*computeMatrix*" and "*plotHeatmap*" functions in the deepTools package. Genome browser images were made using the Integrative Genomics Viewer (IGV) 2.3.68 (Thorvaldsdottir et al., 2013) with bigwig files processed as described above.

Identification of orthologous genes among species

Orthologous genes among species were selected exclusively from syntenic regions of the four genomes. Syntenic orthologs were identified using a combination of CoGe SynFind (https://genomevolution.org/CoGe/SynFind.pl) with default parameters, and CoGe SynMap (https://genomevolution.org/coge/SynMap.pl) with QuotaAlign feature selected and a minimum of six aligned pairs required (Lyons and Freeling, 2008; Lyons et al., 2008).

Peak calling to detect transposase hypersensitive sites (THSs)

Peak calling on ATAC-seq data was performed using the "*Findpeaks*" function of the HOMER package (Heinz et al., 2010). The parameters "*-region*" and "*-minDist 150*" were used to allow identification of variable length peaks and to set a minimum distance of 150 bp between peaks before they are merged into a single peak, respectively. We refer to the peaks called in this way as "transposase hypersensitive sites", or THSs.

Genomic distribution of THSs

For each genome, the distribution of THSs relative to genomic features was assessed using the PAVIS web tool (Huang et al., 2013) with "upstream" regions set as the 2,000 bp upstream of the annotated transcription start site and "downstream" regions set as 1,000 bp downstream of the transcript termination site.

Transcription factor motif analyses

ATAC-seq transposase hypersensitive sites (THSs) that were found in two replicates of each sample were used for motif analysis. The regions were adjusted to the same size (500 bp for root tip THSs or 300 bp for cell type-specific dTHSs). The MEME-ChIP pipeline (Machanick and Bailey, 2011) was run on the repeat-masked fasta files representing each THS set to identify overrepresented motifs, using default parameters. For further analysis, we used the motifs derived from the DREME, MEME, and CentriMo programs that were significant matches (E value < 0.05) to known motifs. Known motifs from both Cis-BP (Weirauch et al., 2014b) and the DAP-seq database (O'Malley et al., 2016) were used in all motif searches.

Assignment of THSs to genes

For each ATAC-seq data set the THSs were assigned to genes using the "*TSS*" function of the PeakAnnotator 1.4 program (Salmon-Divon et al., 2010). This program assigns each peak/THS to the closest transcription start site (TSS), whether upstream or downstream, and reports the distance from the peak center to the TSS based on the genome annotations described above.

ATAC-seq footprinting

To examine motif-centered footprints for TFs of interest we used the "*dnase_average_profile.py"* script in the pyDNase package (Piper et al., 2013). The script was used in ATAC-seq mode ["-*A*" parameter] with otherwise default parameters.

Publicly available DNase-seq, DAP-seq, ChIP-seq, and RNA-seq data

For comparison to our ATAC-seq data from root tips, we used a published DNase-seq dataset from 7-day-old whole *Arabidopsis* roots (SRX391990), which was generated from the same INTACT transgenic line used in our experiments (Sullivan et al., 2014).

Publicly available ChIP-seq and DAP-seq datasets were also used to identify genomic binding sites for transcription factors of interest. These include ABF3 (AT4G34000; SRX1720080) and MYB44 (AT5G67300; SRX1720040)(Song et al., 2016), HY5 (AT5G11260; SRX1412757), CBF2 (AT4G25470; SRX1412036), MYB77 (AT3G50060; SRX1412453), ABI5 (AT2G36270; SRX670505), MYB33 (AT5G06100; SRX1412418), NAC083 (AT5G13180; SRX1412546), MYB77 (AT3G50060; SRX1412453), WRKY27 (AT5G52830; SRX1412681), and At5g04390 (SRX1412214) (O'Malley et al., 2016). Raw reads from these files were mapped and processed as described above for ATAC-seq data, including peak calling with the HOMER package.

Published RNA-seq data from *Arabidopsis* root hair and non-hair cells (Li et al., 2016a) were used to define transcripts that were specifically enriched in the root hair cell relative to the non-hair cell (hair cell enriched genes), and vice versa (non-hair enriched genes). We defined cell type-enriched genes as those whose transcripts were at least two-fold more abundant in one cell type than the other and had an abundance of at least five RPKM in the cell type with higher expression.

Defining high confidence target sites for transcription factors

We used FIMO (Grant et al., 2011) to identify motif occurrences for TFs of interest, and significant motif occurrences were considered to be those with a p-value < 0.0001. Genome-wide high confidence binding sites for a given transcription factor were defined as transposase hypersensitive sites in a given cell type or tissue that also contain a significant motif occurrence for the factor and also overlap with a known enriched region for that factor from DAP-seq or ChIP-seq data (see also Figure S3.2 for a schematic diagram of this process).

Gene ontology analysis

Gene Ontology (GO) analyses using only *Arabidopsis* genes were carried out using the GeneCodis 3.0 program (Nogales-Cadenas et al., 2009; Tabas-Madrid et al., 2012). Hypergeometric tests were used with p-value correction using the false discovery rate (FDR) method. AgriGO was used for comparative GO analysis of gene lists among species, using default parameters (Du et al., 2010; Tian et al., 2017).

Accession Numbers

The raw and processed ATAC-seq data described here have been deposited to the NCBI Gene Expression Omnibus (GEO) database under record number GSE101482. The characteristics of each dataset (individual accession number, read numbers and mapping characteristics, and THS statistics) are included in Table S3.8.

ACKNOWLEDGEMENTS

We thank Paja Sijacic and Shannon Torres for constructive criticism of the manuscript. This work was supported by funding from the National Science Foundation (Plant Genome Research Program grant #IOS-123843) to J.B-S., N.S., S.M.B., and R.B.D.; D.A.W. was supported in part by funding from the Elise Taylor Stocking Memorial Fellowship, and K.K. was supported in part by the Finnish Cultural Foundation.

AUTHOR CONTRIBUTIONS

R.B.D., S.M.B., N.S., J.B-S., K.A.M., M. B., K.K, and M.R, G.P., and D.A.W. designed the research project. K.A.M. performed all experiments on *Arabidopsis* root tips as well as hair and non-hair cells. M.B. performed all experiments on *Medicago* root tips, K.K., D.A.W, and K.Z. performed all experiments on tomato root tips, and M.R. and G.P. performed all experiments on rice root tips. M.W. performed all analyses of syntenic regions and identification of orthologous genes among species. K.B., M.D., and C.Q. analyzed ATAC-seq data sets with Hotspot software and also contributed expertise in other analyses. R.B.D, K.A.M, and M.B. analyzed the data, and R.B.D. drafted the manuscript with subsequent input and editing from all authors.

FIGURES

Fig 3.1

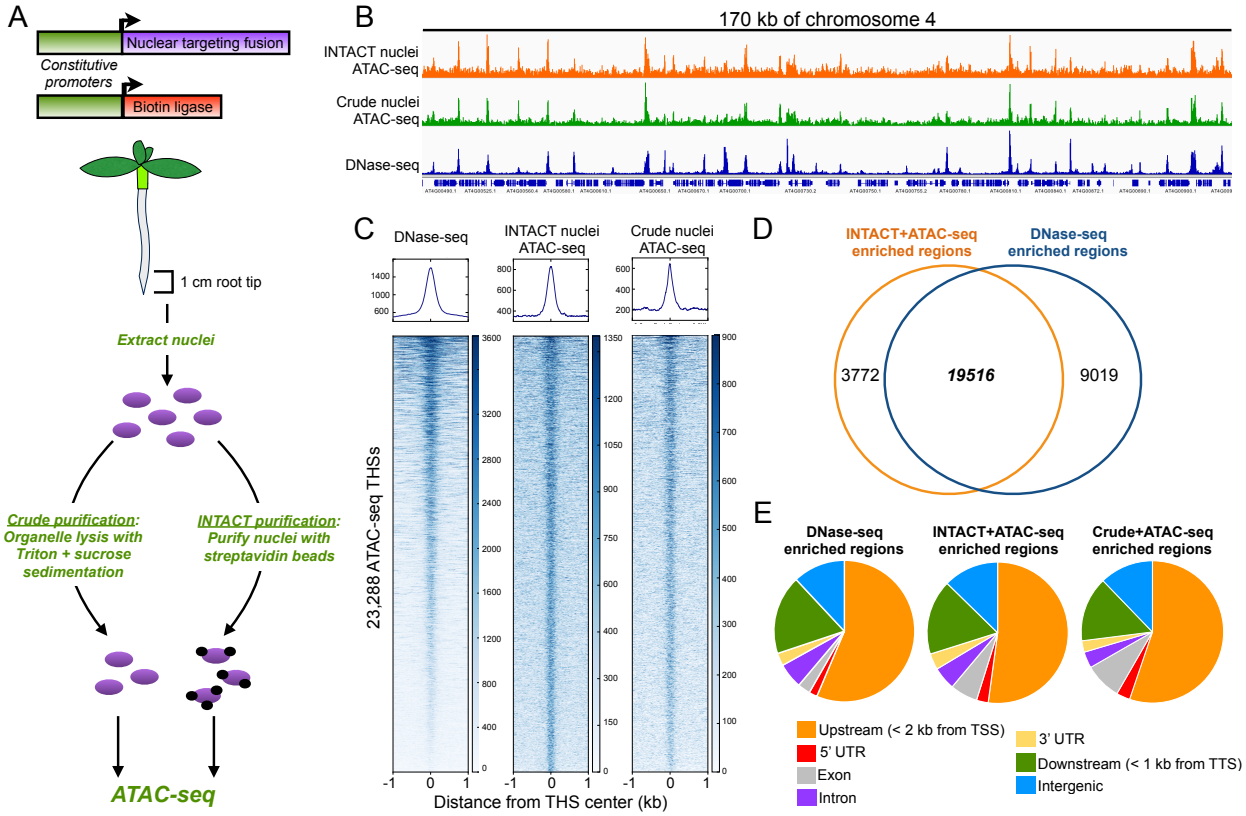


Figure 3.1 Application of ATAC-seq to *Arabidopsis* and comparison with DNase-seq data. (A) Schematic of the INTACT system and strategy for testing ATAC-seq on nuclei with different levels of purity. Upper panel shows the two transgenes used in the INTACT system: the nuclear targeting fusion (NTF) and biotin ligase. Driving expression of both transgenes using constitutive promoters generates biotinylated nuclei in all cell types. Below is a diagram of a constitutive INTACT transgenic plant, showing the 1 cm root tip section used for all nuclei purifications. Root tip nuclei were isolated from transgenic plants and either purified by detergent lysis of organelles followed by sucrose sedimentation (Crude) or purified using streptavidin beads (INTACT). In each case 50,000 purified nuclei were used as input for ATAC-seq. **(B)** Genome browser shot of ATAC-seq data along a 170 kb stretch of chromosome 4 from INTACT-purified and Crude nuclei, as well as DNase-seq data from whole root tissue. Gene models are displayed on the bottom track. **(C)** Average plots and heatmaps of DNase-seq and ATAC-seq signals at the

23,288 ATAC-seq transposase hypersensitive sites (THSs) in the INTACT-ATAC-seq dataset. The regions in the heatmaps are ranked from highest DNase-seq signal (top) to lowest (bottom) **(D)** Venn diagram showing the overlap of enriched regions identified in root tip INTACT-ATAC-seq and whole root DNase-seq datasets. **(E)** Genomic distributions of enriched regions identified in DNase-seq, INTACT-ATAC-seq, and Crude-ATAC-seq datasets.

Fig 3.2

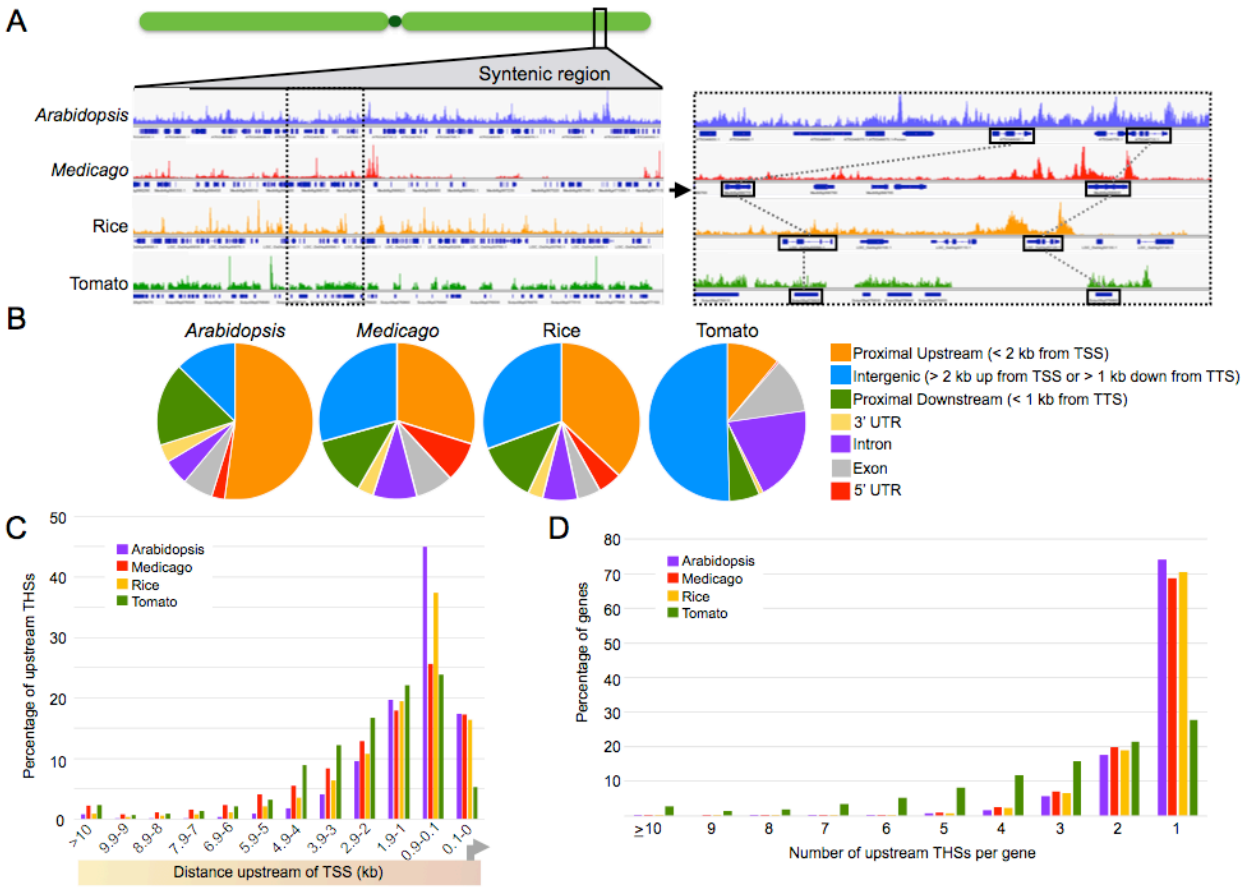


Figure 3.2 ATAC-seq profiling of *Arabidopsis*, *Medicago*, tomato, and rice. (A) Comparison of ATAC-seq data along syntenic regions across the species. The left panel shows a genome browser shot of ATAC-seq data across a syntenic region of all four genomes. ATAC-seq data tracks are shown above the corresponding gene track for each species. The right panel is an enlargement of the region surrounded by a dotted box in the left panel. Orthologous genes are surrounded by black boxes connected by dotted lines between species. Note the apparent similarity in transposase hypersensitivity upstream and downstream of the rightmost orthologs. **(B)** Distribution of ATAC-seq transposase hypersensitive sites (THSs) relative to genomic features in each species. **(C)** Distribution of upstream THSs relative to genes in each species. THSs are binned by distance upstream of the transcription start site (TSS). The number of peaks in each bin is expressed as a percentage of the total upstream THS number in that species. **(D)** Number of upstream

THSs per gene in each species. Graph shows the percentage of all genes with a given number of upstream THSs.

Fig 3.3

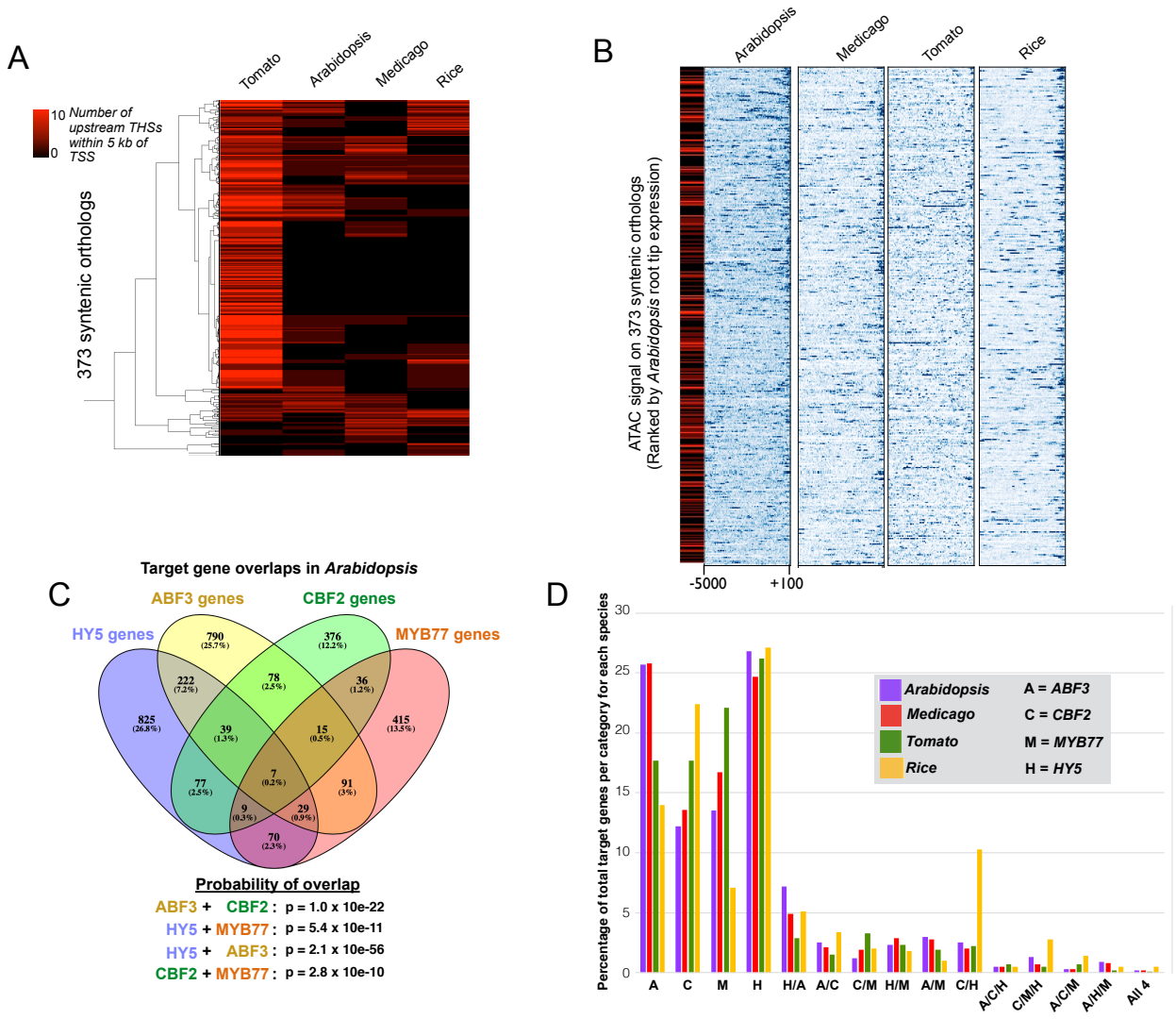


Figure 3.3 Characterization of open chromatin regions and regulatory elements in *Arabidopsis*, *Medicago*, tomato, and rice. (A) Heatmap showing the number of upstream THSs at each of 373 syntenic orthologs in each species. Each row of the heatmap represents a syntenic ortholog, and the number of THSs within 5 kb upstream of the TSS is indicated with a black-to-red color scale for each ortholog in each species. Hierarchical clustering was performed on orthologs using uncentered correlation and average linkage. **(B)** Normalized ATAC-seq signals upstream of orthologous genes. Each row of the heatmaps represents the upstream region of one of the 373 syntenic orthologs in each species. ATAC-seq signal is shown across each ortholog from +100 to -5000 bp relative to the TSS, where blue is high signal and white

is no signal. Heatmaps are ordered by transcript level of each *Arabidopsis* ortholog in the root tip, from highest (top) to lowest (bottom). The leftmost heatmap in black-to-red scale indicates the number of upstream THSs from -100 to -5000 bp associated with each of the *Arabidopsis* orthologs, on the same scale as in (A). **(C)** Overlap of predicted target genes for HY5, ABF3, CBF2, and MYB77 in the *Arabidopsis* root tip. Predicted binding sites for each factor are those THSs that also contain a significant motif occurrence for that factor. Venn diagram shows the numbers of genes with predicted binding sites for each factor alone and in combination with other factors. Significance of target gene set overlap between each TF pair was calculated using a hypergeometric test with a population including all *Arabidopsis* genes reproducibly associated with an ATAC-seq peak in the root tip (13,714 total genes). For each overlap, we considered all genes co-targeted by the two factors. **(D**) Conveying data similar to that in (C), the clustered bar graph shows the percentage of total target genes that fall into a given regulatory category (targeted by a single TF or combination of TFs) in each species.

Fig 3.4

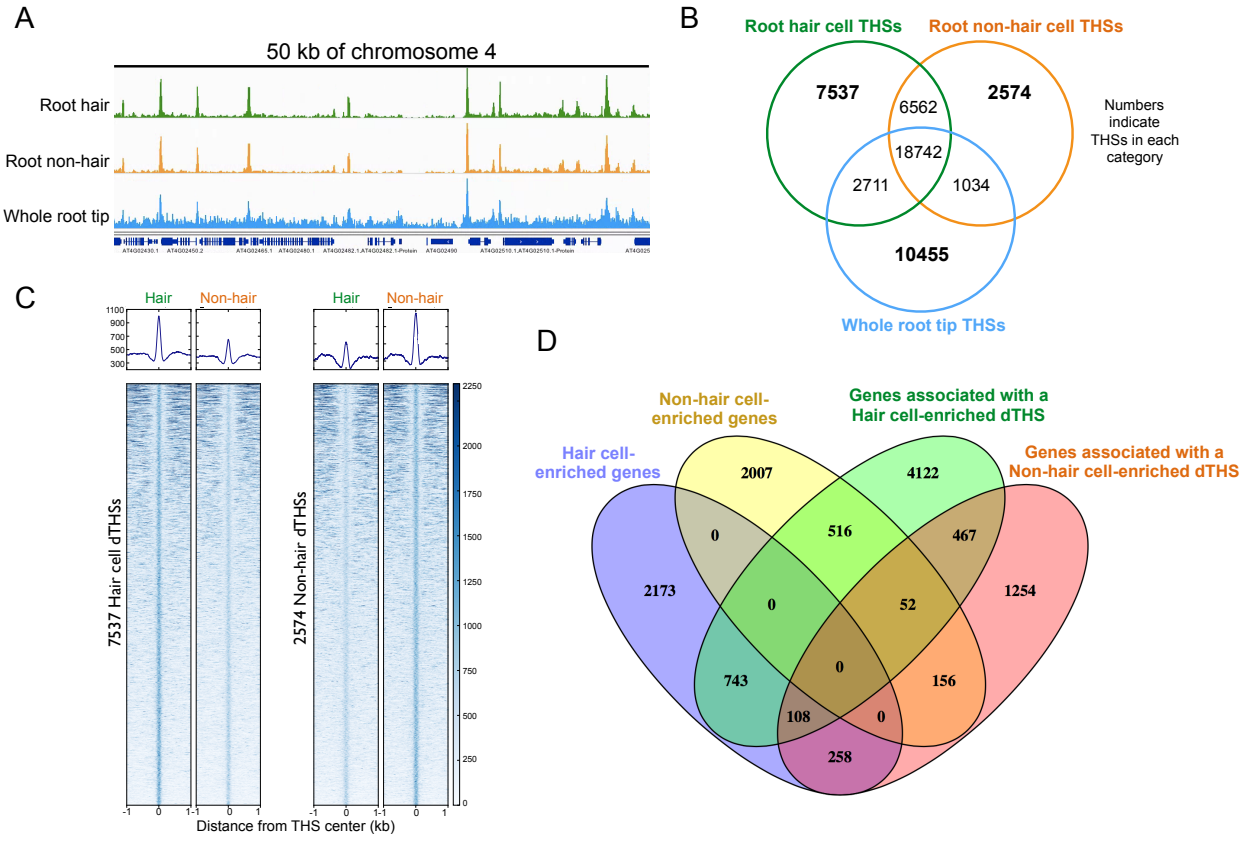


Figure 3.4 Characterization of open chromatin regions in the *Arabidopsis* root hair and non-hair cell types. (A) Genome browser shot of ATAC-seq data from root hair cell, non-hair cell, and whole root tip representing 50 kb of Chromosome 4. **(B)** Overlap of THSs found in two biological replicates of each cell type or tissue. Numbers in bold indicate THSs that are only found in a given cell type or tissue (differential THSs, or dTHSs). **(C)** Average plots and heatmaps showing normalized ATAC-seq signals over 7,537 root hair cell dTHSs (left panels) and 2,574 non-hair cell-enriched dTHSs (right panels). Heatmaps are ranked in decreasing order of total ATAC-seq signal in the hair cell panel in each comparison. Data from one biological replicate is shown here and both replicate experiments showed very similar results. **(D)** Venn diagram of overlaps between cell type-enriched gene sets and genes associated with cell type-enriched dTHSs. Transcriptome data from hair (purple) and non-hair cells (yellow) are from Li et al. (2016) *Developmental Cell*. Genes were considered cell type-enriched if they had a 2-fold or higher difference between cell types and a read count of 5 RPKM or greater in the cell type with higher expression.

Fig 3.5

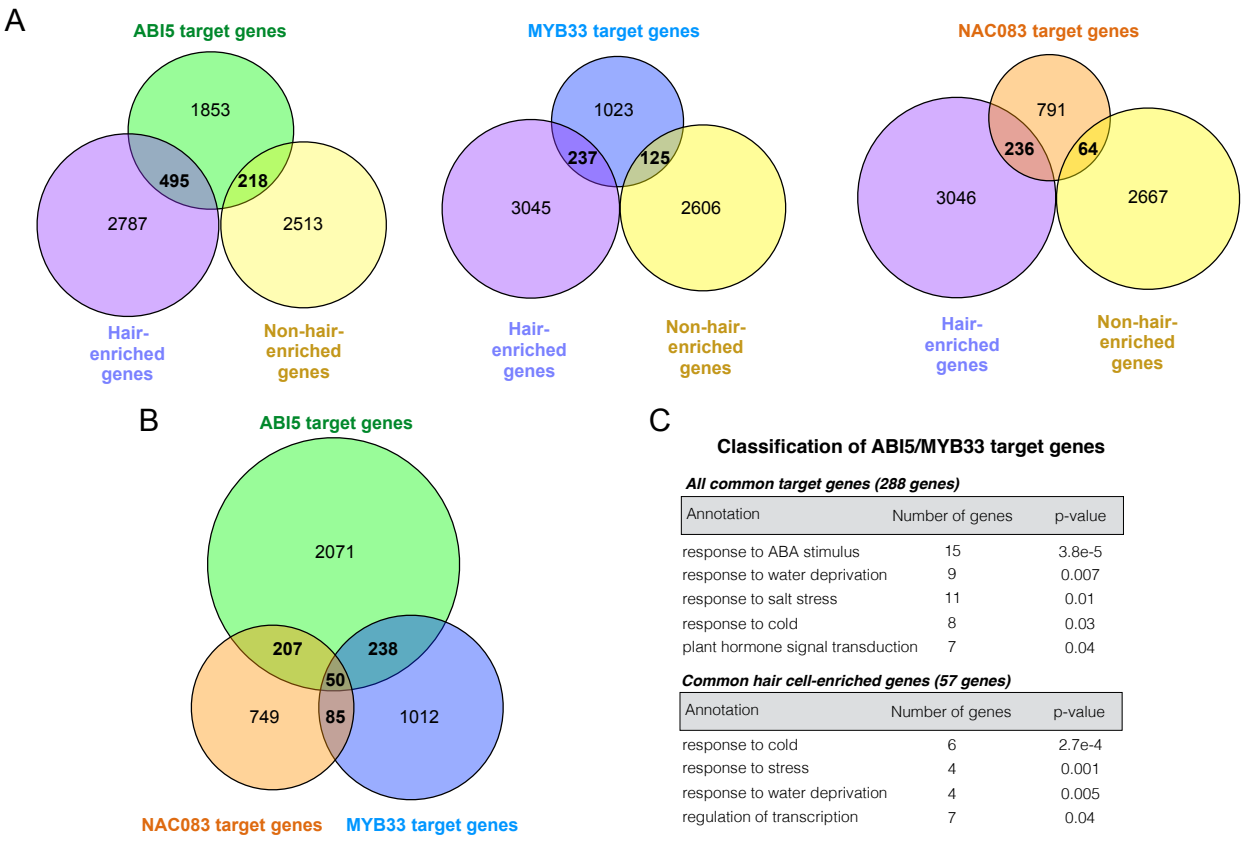
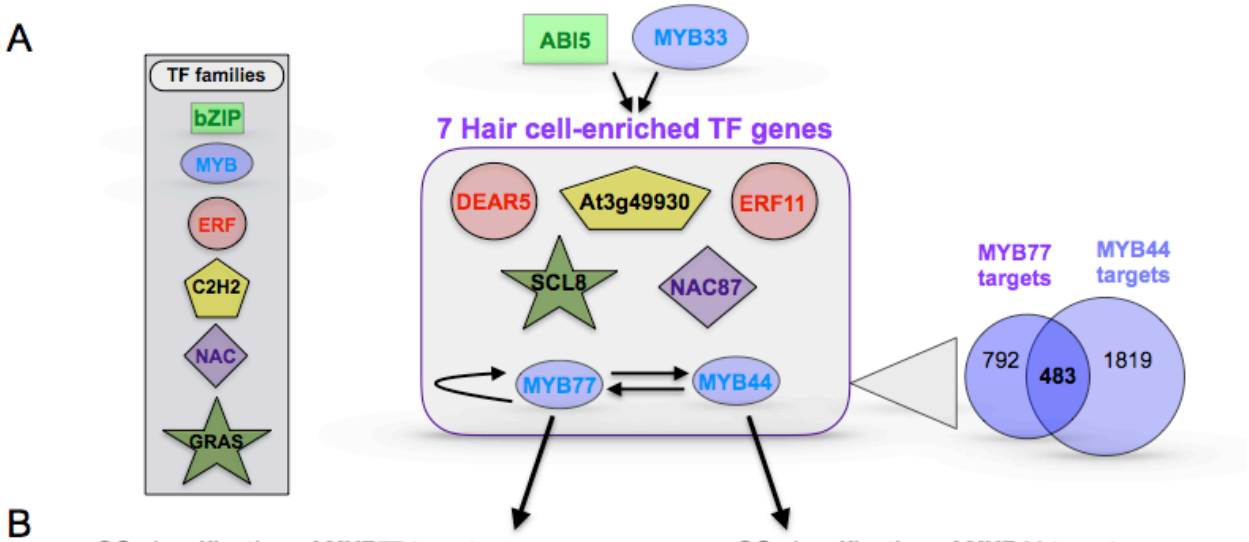


Figure 3.5 Targeting of cell type-enriched genes by H cell-enriched TFs, and co-regulatory associations among H-cell enriched TFs. Genome-wide high confidence binding sites for each TF were defined as open chromatin regions in the hair cell that contain a significant motif occurrence for the factor and also overlap with a known enriched region for that factor from DAP-seq or ChIP-seq data. Target genes were defined by assigning each high confidence binding site to the nearest TSS. **(A)** Venn diagrams showing high confidence target genes for ABI5, MYB33, and NAC083 and their overlap with cell type-enriched genes. **(B)** Overlap of ABI5, MYB33, and NAC083 high confidence target genes. **(C)** Gene Ontology (GO) analysis was performed to illuminate biological functions of genes co-targeted by ABI5 and MYB33. The upper panel shows significantly enriched GO terms for all 288 genes targeted by both ABI5 and MYB33. For each enriched annotation term, the number of genes in the set with that term is shown, followed by the FDR-corrected p-value. The lower panel lists significantly enriched GO-terms for the 57 hair cell-enriched genes co-targeted by ABI5 and MYB33. The seven hair cell-enriched genes

associated with the term *regulation of transcription* were chosen for further analysis. All annotation terms in the lists are at the Biological Process level except for the KEGG pathway term ‘plant hormone signal transduction’.

Fig 3.6



B

GO classification of MYB77 target genes

All target genes (1277 genes)

Annotation	Number of genes	p-value
Plasmodesma (CC)	66	4.5e-5
Response to salt stress (BP)	36	2.6e-4
Response to ethylene stimulus (BP)	17	2.7e-4
Transcription factor activity (MF)	102	6.1e-4
Response to SA stimulus (BP)	16	4.7e-3
Response to ABA stimulus (BP)	26	0.04

GO classification of MYB44 target genes

All target genes (2302 genes)

Annotation	Number of genes	p-value
Plasmodesma (CC)	99	5.1e-5
Response to salt stress (BP)	51	6.6e-4
Response to ethylene stimulus (BP)	23	2.6e-4
Transcription factor activity (MF)	183	7.9e-7
Response to SA stimulus (BP)	23	1.5e-3
Response to ABA stimulus (BP)	44	1.2e-3
Response to JA stimulus (BP)	25	1.0e-3
Response to Auxin stimulus (BP)	35	0.014
Biosynthesis of secondary metabolites (KEGG)	74	5.3e-3

C

Hair cell-enriched target genes (219 genes)

Annotation	Number of genes	p-value
Negative regulation of transcription (BP)	6	5.3e-3
Root hair cell development (BP)	3	3.7e-3
Root hair cell differentiation (BP)	2	0.043

Hair cell-enriched target genes (425 genes)

Annotation	Number of genes	p-value
Negative regulation of transcription (BP)	7	0.01
Response to water deprivation (BP)	14	7.8e-5
Response to phosphate starvation (BP)	6	0.02

Figure 3.6 A transcriptional regulatory module in the root hair cell type. (A) Diagram of the proposed regulatory module under control of ABI5 and MYB33. As referenced in **Figure 3.5C**, ABI5 and MYB33 co-target seven TFs that are preferentially expressed in the hair cell relative to the non-hair cell type. The family classification of each of the seven TFs is denoted in the figure key. Among the seven hair cell-specific target TFs are two MYB family members, MYB77 and MYB44. High confidence binding sites for these two MYB factors were again defined as open chromatin regions in the hair cell that contain a significant motif occurrence for the factor and also overlap with a known enriched region for that factor

from DAP-seq or ChIP-seq data. Each high confidence binding site was then assigned to the nearest TSS to define the target gene for that site. This analysis revealed that MYB44 and MYB77 target each other, and MYB77 targets itself. Both factors target thousands of additional genes, 483 of which are in common (Venn diagram on the lower right of the schematic. Arrows coming down from MYB77 and MYB44 point to GO analyses of that factor's target genes. **(B)** The upper tables represent enriched annotation terms for all target genes of the factor, regardless of differential expression between H and NH cells, while the lower tables **(C)** represent enrichment of terms within target genes that are preferentially expressed in the hair cell relative to the non-hair cell. Annotation term levels are indicated as Cellular Component (CC), Biological Process (BP), Molecular Function (MF) or KEGG pathway (KEGG). For each annotation, the number of target genes associated with that term is shown to the right of the term, followed by the FDR-corrected p-value for the term enrichment in the rightmost column. Groups of terms boxed in gray are those that differ between MYB44 and MYB77. The structure of the module suggests that ABI5 and MYB33 drive a cascade of TFs including MYB77 and MYB44, which act to amplify this signal and also further regulate many additional TFs. Additional target genes of MYB77 and MYB44 include hair cell differentiation factors, hormone response genes, secondary metabolic genes, and genes encoding components of important cellular structures such as plasmodesmata.

LITERATURE CITED

Abdeen, A., Schnell, J., and Miki, B. (2010). Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor ABF3. BMC Genomics **11,** 69.

Bajic, M., Maher, K.A., and Deal, R.B. (2017). Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. Methods in Molecular Biology **1675,** 183-201.

Bauer, D.C., Buske, F.A., and Bailey, T.L. (2010). Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. BMC Bioinformatics **11,** 366.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczynski, B., Riddell, A., and Furlong, E.E. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet **44,** 148-156.

Boyle, P., and Despres, C. (2010). Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. Plant Signal Behav **5,** 629-634.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol **109,** 21 29 21-29.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods **10,** 1213-1218.

Clough, S.J., and Bent, A.F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J **16,** 735-743.

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

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res **38,** W64-70.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. Trends in plant science **15,** 573-581.

Finkelstein, R.R., and Lynch, T.J. (2000). The Arabidopsis abscisic acid response gene ABI5 encodes a basic leucine zipper transcription factor. Plant Cell **12,** 599-609.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics **27,** 1017-1018.

Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu Rev Biochem **57,** 159-197.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antosiewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature **459,** 108-112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell **38,** 576-589.

Henikoff, S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet **9,** 15-26.

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

Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., and Li, L. (2013). PAVIS: a tool for Peak Annotation and Visualization. Bioinformatics **29,** 3097-3099.

Ikeda, M., Mitsuda, N., and Ohme-Takagi, M. (2009). Arabidopsis WUSCHEL is a bifunctional transcription factor that acts as a repressor in stem cell regulation and as an activator in floral patterning. Plant Cell **21,** 3493-3505.

Jaradat, M.R., Feurtado, J.A., Huang, D., Lu, Y., and Cutler, A.J. (2013). Multiple roles of the transcription factor AtMYBR1/AtMYB44 in ABA signaling, stress responses, and leaf senescence. BMC Plant Biol **13,** 192.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet **43,** 264-268.

Jung, C., Shim, J.S., Seo, J.S., Lee, H.Y., Kim, C.H., Choi, Y.D., and Cheong, J.J. (2010). Non-specific phytohormonal induction of AtMYB44 and suppression of jasmonate-responsive gene activation in Arabidopsis thaliana. Mol Cells **29,** 71-76.

Kang, J.Y., Choi, H.I., Im, M.Y., and Kim, S.Y. (2002). Arabidopsis basic leucine zipper proteins that mediate stress-responsive abscisic acid signaling. Plant Cell **14,** 343-357.

Kerr, T.C., Abdel-Mageed, H., Aleman, L., Lee, J., Payton, P., Cryer, D., and Allen, R.D. (2017). Ectopic expression of two AREB/ABF orthologs increases drought tolerance in cotton (Gossypium hirsutum). Plant Cell Environ.

Knight, H., Zarka, D.G., Okamoto, H., Thomashow, M.F., and Knight, M.R. (2004). Abscisic acid induces CBF gene transcription and subsequent induction of cold-regulated genes via the CRT promoter element. Plant Physiol **135,** 1710-1717.

Kvon, E.Z., Kazmar, T., Stampfel, G., Yanez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature **512,** 91-95.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods **9,** 357-359.

Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet **34,** 77-137.

Li, D., Li, Y., Zhang, L., Wang, X., Zhao, Z., Tao, Z., Wang, J., Wang, J., Lin, M., Li, X., and Yang, Y. (2014). Arabidopsis ABA Receptor RCAR1/PYL9 Interacts with an R2R3-Type MYB Transcription Factor, AtMYB44. International Journal of Molecular Sciences **15,** 8473-8490.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25,** 2078-2079.

Li, J., Farmer, A.D., Lindquist, I.E., Dukowic-Schulze, S., Mudge, J., Li, T., Retzel, E.F., and Chen, C. (2012). Characterization of a set of novel meiotically-active promoters in Arabidopsis. BMC Plant Biol **12,** 104.

Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P.N. (2016a). High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation. Developmental Cell **39,** 508-522.

Li, T., Wu, X.Y., Li, H., Song, J.H., and Liu, J.Y. (2016b). A Dual-Function Transcription Factor, AtYY1, Is a Novel Negative Regulator of the Arabidopsis ABA Response Network. Mol Plant **9,** 650-661.

Limpens, E., Ramos, J., Franken, C., Raz, V., Compaan, B., Franssen, H., Bisseling, T., and Geurts, R. (2004). RNA interference in Agrobacterium rhizogenes-transformed roots of Arabidopsis and Medicago truncatula. J Exp Bot **55,** 983-992.

Lu, Z., Hofmeister, B.T., Vollmers, C., DuBois, R.M., and Schmitz, R.J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Res **45,** e41.

Lucas, W.J., and Lee, J.Y. (2004). Plasmodesmata as a supracellular control network in plants. Nat Rev Mol Cell Biol **5,** 712-726.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403,** 564-567.

Ludwig, M.Z., Palsson, A., Alekseeva, E., Bergman, C.M., Nathan, J., and Kreitman, M. (2005). Functional evolution of a cis-regulatory module. PLoS Biol **3,** e93.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J **53,** 661-673.

Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008). The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Tropical Plant Biology **1,** 181-190.

Ma, W., Noble, W.S., and Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nature protocols **9,** 1428-1450.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics **27,** 1696-1697.

Masucci, J.D., Rerie, W.G., Foreman, D.R., Zhang, M., Galway, M.E., Marks, M.D., and Schiefelbein, J.W. (1996). The homeobox gene GLABRA2 is required for position-dependent cell differentiation in the root epidermis of Arabidopsis thaliana. Development **122,** 1253-1260.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W.W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res **42,** D142-147.

Matus, J.T., Cavallini, E., Loyola, R., Holl, J., Finezzo, L., Dal Santo, S., Vialet, S., Commisso, M., Roman, F., Schubert, A., Alcalde, J.A., Bogs, J., Ageorges, A., Tornielli, G.B., and Arce-Johnson, P. (2017). A Group of Grapevine MYBA Transcription Factors Located in Chromosome 14 Control Anthocyanin Synthesis in Vegetative Organs with Different Specificities Compared to the Berry Color Locus. Plant J.

Medford, J.I., Elmer, J.S., and Klee, H.J. (1991). Molecular cloning and characterization of genes expressed in shoot apical meristems. Plant Cell **3,** 359-370.

Mejia-Guerra, M.K., Li, W., Galeano, N.F., Vidal, M., Gray, J., Doseff, A.I., and Grotewold, E. (2015). Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. Plant Cell **27,** 3309-3320.

Millar, A.A., and Gubler, F. (2005). The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. The Plant cell **17,** 705-721.

Mo, A., Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Sejnowski, T.J., Lister, R., Eddy, S.R., Ecker, J.R., and Nathans, J. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron **86,** 1369-1384.

Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U., and Megraw, M. (2014). Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. Plant Cell **26,** 2746-2760.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., Wei, Y., Nguyen, T., Greenside, P.G., Corces, M.R., Tycko, J., Simeonov, D.R., Suliman, N., Li, R., Xu, J., Flynn, R.A., Kundaje, A., Khavari, P.A., Marson, A., Corn, J.E., Quertermous, T., Greenleaf, W.J., and Chang, H.Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet **49,** 1602-1612.

Murashige, T., and Skoog, F. (1962). A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. Physiol Plantarum **15,** 473-497.

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M., and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res **37,** W317-322.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **166,** 1598.

Oh, S.J., Song, S.I., Kim, Y.S., Jang, H.J., Kim, S.Y., Kim, M., Kim, Y.K., Nahm, B.H., and Kim, J.K. (2005). Arabidopsis CBF3/DREB1A and ABF3 in transgenic rice increased tolerance to abiotic stress without stunting growth. Plant Physiol **138,** 341-351.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet **12,** 283-293.

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 & Development **11,** 2983-2995.

Pajoro, A., Madrigal, P., Muino, J.M., Matus, J.T., Jin, J., Mecchia, M.A., Debernardi, J.M., Palatnik, J.F., Balazadeh, S., Arif, M., O'Maoileidigh, D.S., Wellmer, F., Krajewski, P., Riechmann, J.L., Angenent, G.C., and Kaufmann, K. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol **15,** R41.

Patel, R.V., Nahal, H.K., Breit, R., and Provart, N.J. (2012). BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. The Plant Journal **71,** 1038-1050.

Persak, H., and Pitzschke, A. (2014). Dominant repression by Arabidopsis transcription factor MYB44 causes oxidative damage and hypersensitivity to abiotic stress. International Journal of Molecular Sciences **15,** 2517-2537.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Research **41,** e201-e201.

Piper, J., Assi, S.A., Cauchy, P., Ladroue, C., Cockerill, P.N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. BMC genomics **16,** 1000.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res **44,** W160-165.

Reyes, J.L., and Chua, N.H. (2007). ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. Plant J **49,** 592-606.

Rodgers-Melnick, E., Vera, D.L., Bass, H.W., and Buckler, E.S. (2016). Open chromatin reveals the functional maize genome. Proceedings of the National Academy of Sciences of the United States of America **113,** E3177-E3184.

Ron, M., Kajala, K., Pauluzzi, G., Wang, D., Reynoso, M.A., Zumstein, K., Garcha, J., Winte, S., Masson, H., Inagaki, S., Federici, F., Sinha, N., Deal, R.B., Bailey-Serres, J., and Brady, S.M. (2014). Hairy root transformation using Agrobacterium rhizogenes as a tool for exploring cell type-specific gene expression and function using tomato as a model. Plant Physiol **166,** 455-469.

Ruzicka, D.R., Kandasamy, M.K., McKinney, E.C., Burgos-Rivera, B., and Meagher, R.B. (2007). The ancient subclasses of Arabidopsis Actin Depolymerizing Factor genes exhibit novel and differential expression. Plant J **52,** 460-472.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics **11,** 415.

Scharer, C.D., Blalock, E.L., Barwick, B.G., Haines, R.R., Wei, C., Sanz, I., and Boss, J.M. (2016). ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. Nature Publishing Group **6,** 27030.

Schellmann, S., Schnittger, A., Kirik, V., Wada, T., Okada, K., Beermann, A., Thumfahrt, J., Jurgens, G., and Hulskamp, M. (2002). TRIPTYCHON and CAPRICE mediate lateral inhibition during trichome and root hair patterning in Arabidopsis. EMBO J **21,** 5036-5046.

Shim, J.S., Jung, C., Lee, S., Min, K., Lee, Y.W., Choi, Y., Lee, J.S., Song, J.T., Kim, J.K., and Choi, Y.D. (2013). AtMYB44 regulates WRKY70 expression and modulates antagonistic interaction between salicylic acid and jasmonic acid signaling. Plant J **73,** 483-495.

Shin, R., Burch, A.Y., Huppert, K.A., Tiwari, S.B., Murphy, A.S., Guilfoyle, T.J., and Schachtman, D.P. (2007). The Arabidopsis transcription factor MYB77 modulates auxin signal transduction. The Plant cell **19,** 2440-2453.

Song, L., Huang, S.s.C., Wise, A., Castanon, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science (New York, NY) **354,** aag1550-aag1550.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet **13,** 613-626.

Stadhouders, R., van den Heuvel, A., Kolovos, P., Jorna, R., Leslie, K., Grosveld, F., and Soler, E. (2012). Transcription regulation by distal enhancers: who's in the loop? Transcription **3,** 181-186.

Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P., Stergachis, A.B., Vernot, B., Johnson, A.K., Haugen, E., Sullivan, S.T., Thompson, A., Neri III, F.V., Weaver, M., Diegel, M., Mnaimneh, S., Yang, A., Hughes, T.R., Nemhauser, J.L., Queitsch, C., and Stamatoyannopoulos, J.A. (2014). Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in Arabidopsis thaliana. CellReports **8,** 2015-2030.

Sung, M.H., Baek, S., and Hager, G.L. (2016). Genome-wide footprinting: ready for prime time? Nat Methods **13,** 222-228.

Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res **40,** W478-483.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14,** 178-192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., and Stamatoyannopoulos, J.A. (2012). The accessible chromatin landscape of the human genome. Nature **489,** 75-82.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res.

Tittarelli, A., Santiago, M., Morales, A., Meisel, L.A., and Silva, H. (2009). Isolation and functional characterization of cold-regulated promoters, by digitally identifying peach fruit cold-induced genes from a large EST dataset. BMC Plant Biol **9,** 121.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res **24,** 1334-1347.

Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W. (2014). Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. Plant Cell **26,** 3883-3893.

Vierstra, J., and Stamatoyannopoulos, J.A. (2016). Genomic footprinting. Nature Methods **13,** 213-221.

Wada, T., Kurata, T., Tominaga, R., Koshino-Kimura, Y., Tachibana, T., Goto, K., Marks, M.D., Shimura, Y., and Okada, K. (2002). Role of a positive regulator of root hair development, CAPRICE, in Arabidopsis root epidermal cell differentiation. Development **129,** 5409-5419.

Wang, N., Xu, H., Jiang, S., Zhang, Z., Lu, N., Qiu, H., Qu, C., Wang, Y., Wu, S., and Chen, X. (2017). MYB12 and MYB22 play essential roles in proanthocyanidin and flavonol synthesis in red-fleshed apple (Malus sieversii f. niedzwetzkyana). Plant J **90,** 276-292.

Wang, Z., Su, G., Li, M., Ke, Q., Kim, S.Y., Li, H., Huang, J., Xu, B., Deng, X.P., and Kwak, S.S. (2016). Overexpressing Arabidopsis ABF3 increases tolerance to multiple abiotic stresses and reduces leaf size in alfalfa. Plant Physiol Biochem **109,** 199-208.

Weber, B., Zicola, J., Oka, R., and Stam, M. (2016). Plant Enhancers: A Call for Discovery. Trends in plant science **21,** 974-987.

Weirauch, M.T., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.C., Galli, M., Lewsey, M., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F.Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., and Hughes, T.R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158,** 1431-1443.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.M., Pham, G.M., Nicotra, A.B., Gregorio, G.B., Jagadish, S.V., Septiningsih, E.M., Bonneau, R., and Purugganan, M. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. Plant Cell **28,** 2365-2384.

Xu, Y., and Du, J. (2014). Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (Solanum lycopersicum) plants. Plant J **80,** 582-591.

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., Zhiqiang, L., Yunfei, Z., Xiaoxiao, W., Xiaoming, Q., Yunping, S., Li, Z., Xiaohui, D., Jingchu, L., Xing-Wang, D., Zhangliang, C., Hongya, G., and Li-Jia, Q. (2006). The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. Plant Mol Biol **60,** 107-124.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012a). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. Plant Cell **24,** 2719-2731.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012b). High-resolution mapping of open chromatin in the rice genome. Genome Research **22,** 151-162.

Zhao, Y., Xing, L., Wang, X., Hou, Y.J., Gao, J., Wang, P., Duan, C.G., Zhu, X., and Zhu, J.K. (2014). The ABA receptor PYL8 promotes lateral root growth by enhancing MYB77-dependent transcription of auxin-responsive genes. Sci Signal **7,** ra53.

Zhu, B., Zhang, W., Zhang, T., Liu, B., and Jiang, J. (2015). Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures. Plant Cell **27,** 2415-2426.

CHAPTER 4: CHROMATIN ACCESSIBILITY CHANGES BETWEEN *ARABIDOPSIS* STEM CELLS AND MESOPHYLL CELLS ILLUMINATE CELL TYPE-SPECIFIC TRANSCRIPTION FACTOR NETWORKS

Paja Sijacic*, Marko Bajic*, Elizabeth C. McKinney, Richard B. Meagher, Roger B. Deal

This work is published in *Plant Journal* (2018) 2:215-231. doi: 10.1111/tpj.13882.

Supplemental tables can be found in the online publication.

*These authors contributed equally to this work.

SUMMARY

Cell differentiation is driven by changes in transcription factor (TF) activity and subsequent alterations in transcription. To study this process, differences in TF binding between cell types can be deduced by probing chromatin accessibility. We used cell type-specific nuclei purification followed by the Assay for Transposase Accessible Chromatin (ATAC-seq) to delineate differences in chromatin accessibility and TF regulatory networks between stem cells of the shoot apical meristem (SAM) and differentiated leaf mesophyll cells in *Arabidopsis thaliana*. Chromatin accessibility profiles of SAM stem cells and leaf mesophyll cells were highly similar at a qualitative level, yet thousands of regions of quantitatively different chromatin accessibility were also identified. Analysis of the genomic regions preferentially accessible in each cell type identified hundreds of overrepresented TF binding motifs, highlighting sets of TFs that are likely important for each cell type. Within these sets, we found evidence for extensive co-regulation of target genes by multiple TFs that are preferentially expressed in each cell type. Interestingly, the TFs within each of these cell type-enriched sets also show evidence of extensively co-regulating each other. We further found that chromatin regions preferentially accessible in mesophyll cells tended to also be substantially accessible in the stem cells, whereas the converse was not true. This observation suggests that the generally higher accessibility of regulatory elements in stem cells might contribute to their developmental plasticity.

This work demonstrates the utility of cell type-specific chromatin accessibility profiling in quickly developing testable models of regulatory control differences between cell types.

Significance Statement

Differences in transcription factor (TF) activity between different plant cell types remains largely unexplored. We used cell type-specific nuclei purification followed by chromatin accessibility assays to delineate TF networks in stem cells of the shoot apical meristem and differentiated leaf mesophyll cells in *Arabidopsis thaliana*, providing insight into how the transcriptomes of these cell types are established and maintained.

INTRODUCTION

In higher plants, all above ground tissues are continuously produced due to the activities of self-renewing, pluripotent stem cells located in the central zone of the shoot apical meristem (SAM). Upon stem cell division, a subset of daughter cells is gradually displaced to the peripheral zones of the SAM where these cells continue to divide and differentiate. During this process, differentiating cells undergo transcriptional reprogramming as they acquire specialized fates within developing leaf primordia at the flanks of the SAM (Barton, 2010; Besnard et al., 2011).

Chromatin compaction within the nucleus often restricts the access of transcription factors (TFs) to *cis*-regulatory elements, such as promoters and enhancers (Spitz and Furlong, 2012). During differentiation, cells employ various mechanisms to induce local changes in chromatin properties, thereby modifying the accessibility of regulatory chromatin regions to the transcriptional machinery (Spitz and Furlong, 2012; Burton and Torres-Padilla, 2014). This ultimately leads to the establishment of lineage-specific TF regulatory modules and the resulting transcriptional output characteristic of a given cell type. To date, a limited number of such cell type-specific TF regulatory modules have been identified in plants. One well-studied example is the network of TFs that controls specification of the root non-hair cell type in the *Arabidopsis* root epidermis. In this system, the interactions of multiple TFs dictate expression of the non-

hair fate master regulator, GLABRA2 (GL2), which subsequently determines cell fate (Schiefelbein et al., 2014; Balcerowicz et al., 2015). This complex of TFs that regulate the expression of *GL2* was delineated through extensive genetic studies in numerous laboratories and now represents one of the best understood fate specification pathways in plants. To expedite mechanistic studies of cell fate specification in many other cell types, it will be important to be able to identify cell-type specific *cis*-regulatory regions and the transcription factors that act on them.

To measure DNA accessibility and TF binding, genome-wide analysis methods such as DNase I treatment of nuclei coupled with high-throughput sequencing (DNase-seq) have been used (John et al., 2013; He et al., 2014; Zhong et al., 2016). Mapping of DNase I hypersensitive sites (DHSs) allows for the identification of *cis*-regulatory elements because DHSs represent open chromatin regions where protein binding to DNA has displaced nucleosomes, generating a nuclease sensitive zone (Sheffield et al., 2013). Large-scale DNase-seq studies have been instrumental in identifying cell type-specific *cis*-regulatory elements, most notably including a study involving more than 100 human cell types (Thurman et al., 2012). One substantial drawback of this powerful technique, however, is the requirement for large quantities of nuclei as starting material. Recently, the simple and sensitive Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) has been described (Buenrostro et al., 2013; Buenrostro et al., 2015), which requires much smaller amount of input material (~500 to 50,000 nuclei) (Lu et al., 2016). In this method, a hyperactive Tn5 transposase loaded with sequencing adapters acts to simultaneously fragment and tag a genome with these adapters. Mapping of the transposase hypersensitive sites allows for detection of highly accessible chromatin regions and subsequent identification of TF binding sites within these regions (Lu et al., 2016).

One of the main limitations to successfully identifying cell type-specific *cis*-regulatory regions and studying the networks of transcription factors that bind to these elements is the difficulty in isolating specific cell types. The INTACT (Isolation of Nuclei TAgged in specific Cell Types) technique is one solution to this problem that is highly amenable to chromatin studies (Deal and Henikoff, 2010, 2011). This system utilizes transgenic plants carrying two transgenes. The first encodes the nuclear targeting fusion

(NTF) protein, which is comprised of a nuclear envelope-targeting domain, green fluorescent protein (GFP), and biotin ligase recognition peptide (BLRP). The second transgene is the *E. coli* biotin ligase (BirA) which specifically biotinylates the NTF protein. The *BirA* transgene is expressed from a constitutively active promoter, while the expression of NTF is driven by a cell type-specific promoter. The co-expression of these transgenes results in the biotinylation of nuclei in a specific cell type, which can then be affinity purified with streptavidin-coated magnetic beads.

In this study, we employed INTACT and ATAC-seq methods, collectively called INTACT-ATAC-seq, to identify and compare accessible chromatin regions between two distinct plant cell types: pluripotent stem cells in the central zone of the SAM, and highly specialized, fully-differentiated leaf mesophyll cells that originate from the stem cells of the SAM. The comparison of these two cell types offers a unique insight into chromatin dynamics and transcriptional control at both the starting and ending points of the differentiation process. Our results show that while most Transposase Hypersensitive Sites (THSs) are shared between both cell types, thousands of regions could be identified that were quantitatively more accessible in one cell type compared to the other. Furthermore, we identified transcription factor (TF) binding motifs within these THSs and used this information, in combination with publicly available expression and protein interaction data, to build cell-specific TF-to-TF regulatory networks, and to predict the downstream target genes of these TF networks. Our results suggest that distinct classes of TFs collaborate to produce cell type-specific transcriptomes in the stem cell and mesophyll cell types. We also demonstrate that INTACT-ATAC-seq is a powerful technique to quickly develop testable hypotheses regarding TF regulatory networks and their roles in cell fate specification.

RESULTS

Validation of cell type-specific INTACT lines and INTACT-ATAC-seq data

The *CLAVATA3* (*CLV3)* gene, a known stem cell marker (Schoof et al., 2000), is exclusively expressed in the meristematic stem cells found in the central zone of the SAM (Yadav et al., 2009). We used the upstream

and downstream regulatory sequences of *CLV3* to drive the expression of the nuclear targeting fusion (NTF) transgene selectively in the SAM stem cells. Expression of the *CLV3p::NTF* construct in *CLV3p::NTF;ACT2p::BirA* T_2 transgenic plants was confirmed using confocal microscopy by visualizing the Green Fluorescent Protein (GFP), which is a part of the NTF, specifically in the central zone of the SAM (Figure 4.1A). Similarly, the promoter of the *Rubisco small subunit 2B* (*RBC*) gene, active only in the mesophyll cells (Sawchuk et al., 2008), was used to drive the expression of the NTF in leaf mesophyll cells. The expression of this construct was visualized by confocal microscopy in the leaves of the *RBCp::NTF;ACT2p::BirA* T_2 transgenic plants. GFP expression was observed in the inner cell layers of the sectioned leaf, and is excluded from the leaf epidermis (Figure 4.1A).

The INTACT protocol for cell type-specific nuclei purification from *CLV3p::NTF;ACT2p::BirA* and *RBCp::NTF;ACT2p::BirA* T_2 transgenic plants was performed as previously described (Bajic et al., 2018). A total of 25,000 freshly isolated nuclei were used for ATAC-seq, and three biological replicates were performed per cell type. In parallel, we performed ATAC-seq on genomic DNA isolated from leaf tissue as a control for sequence-specific Tn5 transposase incorporation bias. More than 84 million reads were obtained for each biological replicate through paired-end sequencing (Figure S4.1A). After aligning the ATAC-seq reads to the *Arabidopsis thaliana* TAIR10 genome, we found that, on average, 46% of all reads from the stem cells and 21% from the mesophyll cells were successfully mapped to the nuclear genome, with the remainder of reads mapping to organelle genomes (Figure S4.1A). This level of organelle DNA carry over was unexpected based on our previous INTACT purifications from root tissue, and is likely attributable to the sheer abundance of chloroplasts in shoot tissue. All reads that aligned to the organellar genomes were subsequently omitted from downstream analyses. More than 15 million reads per replicate passed the quality filtering stage of analysis (Figure S4.1A), which is more than sufficient to successfully identify accessible chromatin regions in *Arabidopsis*, as has been recently demonstrated (Lu et al., 2016). The processed, alignment files were compared using principal component analysis (PCA) (Ramirez et al., 2016). The six libraries segregated by cell type, with low variation between replicates, indicating the high level of reproducibility in our datasets (Figure S4.1B). For each library, we analyzed the fragment size

distribution of the aligned reads to determine the number of nucleosome-containing (>150 bp) and nucleosome-free reads (<150 bp). Nucleosome-free reads are regions of accessible chromatin where a transcription factor is likely bound. Conversely, nucleosome-containing reads are less accessible to transcription factor binding and are therefore less relevant to the scope of this study. In the stem cell and mesophyll ATAC-seq datasets, we saw a fragment size distribution primarily of 100 bp fragments and smaller, indicating that our libraries were composed of primarily nucleosome-free reads (Figure S4.1C). Additionally, the periodic dips in the size distribution graphs demonstrate a clear pattern of the helical pitch of DNA, further confirming that our transposase treatment was of sufficient coverage. The fragment size distribution for genomic DNA ATAC-seq library was smaller, primarily 50 bp in size, and lacked a clear representation of the helical pitch of DNA (Figure S4.1C). In summary, INTACT-ATAC-seq is a very effective method for obtaining a large quantity of highly reproducible accessible chromatin reads in *Arabidopsis* mesophyll and stem cells.

Identification and genomic distribution of cell type-specific accessible chromatin regions

Since the ATAC-seq data among all replicates were highly reproducible (Figure S4.1B), we focused our analysis on the two biological replicates with the highest number of aligned reads for each cell type (Figure S4.1A, replicates 2 and 3 for each cell type). To keep our analysis consistent across samples, we first scaled the reads from each cell type to the same sequencing depth (15,288,699 reads, Figure S4.1A) and then used the peak calling function of the HOMER package (Heinz et al., 2010) to identify open chromatin regions. From this set of transposase hypersensitive sites (THSs) identified by HOMER, we examined only the THS regions that were identified in both replicates of each cell type, which we refer to as reproducible THSs (Table S4.1). The majority of these reproducible THSs (22,961 of 30,459 sites) were common to both cell types, while 5,283 and 2,215 THSs were reproducibly called only in one cell type (stem cells and mesophyll cells, respectively) (Figure 4.1B and C). The genomic distribution of reproducible THSs is very similar between the two cell types, with 53% of THSs located within 2 kb upstream of the gene transcription start sites (TSSs), 18% located within the gene body, 16% located within 1 kb downstream of transcription

termination sites (TTSs), and 10% located in the intergenic region (Figure 4.1D). This genomic distribution of reproducible THSs suggests that the majority of *cis*-regulatory regions in *Arabidopsis* genome are located in the vicinity of gene core promoters, as previously observed in other *Arabidopsis* cell types (Maher et al., 2017)

Since the majority of identified THSs were common to both cell types (Figure 4.1C) we hypothesized that there may still be quantitative differences between cell types at the shared THSs that would not be identified by our all-or-nothing peak calling approach. To examine quantitative differences in accessible chromatin regions between the two cell types, we calculated the normalized total read counts at each THS in the merged set of reproducible THSs for both cell types (*i.e.* all THSs shown in Figure 4.1C). The calculated read counts were then evaluated using DESeq2 to identify reproducible quantitative differences in accessibility between cell types (Love et al., 2014) (Table S4.2). Only those THSs that had a log fold change of 1 or more in a specific cell type were categorized as THSs enriched in that cell type (see Experimental Procedures).

With this approach, we identified a total of 7,394 THSs that are more accessible in stem cells and 5,895 THSs that are more accessible in mesophyll cells (Figure 4.2A). This analysis captured the majority of THSs originally identified as cell type-unique by peak calling, and added several thousand differential THSs to each cell type that were previously classified as being present in both cell types by peak calling alone. We now refer to these collections of THSs that are quantitatively significantly different between cell types as cell type-enriched THSs.

Each set of cell type-enriched THSs had a similar genomic distribution which matched the trend of the overall THS distribution, with more than 75% of these THSs mapping within 2 kb upstream of the gene TSSs and 1 kb downstream of TTSs (Figure S4.2A). Heatmaps and average plots of the ATAC-seq signal from mesophyll, stem cell, and genomic DNA datasets were visualized over stem cell-enriched and mesophyll-enriched THSs to examine the chromatin accessibility differences between cell types at these sites (Figure 4.2B). At stem cell-enriched THSs, the stem cell ATAC-seq signal is strongest in the centers of these regions (with an average maximum of approximately 1500 RPKM) and drops off sharply to either

side. In contrast, the mesophyll cells show far less accessibility in these regions, but are nonetheless accessible to transposase integration to some degree (Figure 4.2B, left panels). At mesophyll-enriched THSs, the mesophyll cells show the highest accessibility in these regions with an average maximum of 4,400 RPKM (Figure 4.2B, right panels). It is worth noting that this read signal is much higher than that seen in stem cells at stem cell-enriched THSs. Interestingly, stem cells also show relatively high accessibility at mesophyll-enriched THSs, with an average maximum in the same range as that seen at stem cell-enriched THSs. These results strongly indicate that mesophyll-enriched THSs are already highly accessible in stem cells, but not vice versa.

On the other hand, ATAC-seq reads from genomic DNA were present at negligible levels at both the stem cell-enriched and mesophyll-enriched THSs (Figure 4.2B). In fact, we identified only 35 THSs in genomic DNA by peak calling, with approximately 75% of them located in the intergenic regions of the genome (Figure S4.3A and B). These results suggest a very low level of Tn5 integration bias at this scale. Taken together, we successfully used INTACT-ATAC-seq to identify cell type-enriched THSs, which reflect the reproducible differences in the chromatin accessibility between the stem cells and mesophyll cells.

Gene ontology analysis of the genes associated with cell type-enriched THSs

THSs represent accessible, nucleosome-free, chromatin regions and are likely to contain *cis*-regulatory elements that control the expression of nearby genes. To identify the genes associated with the cell type-enriched THSs we used the PeakAnnotator program (Salmon-Divon et al., 2010) to assign each THS to the nearest gene TSS, regardless of whether the TSS is upstream or downstream. We will hereafter refer to the genes associated with the stem cell-enriched THSs as the stem cell THS-proximal genes, and the genes associated with the mesophyll-enriched THSs as the mesophyll THS-proximal genes. The 7,394 stem cell-enriched THSs mapped to the 5,490 stem cell THS-proximal genes, while the 5,895 mesophyll-enriched THSs mapped to the 4,513 mesophyll THS-proximal genes (Figure 4.2C and Figure S4.2B). These results indicate that in each cell type a single gene sometimes has more than one cell type-enriched THS associated

with it, while the majority of genes that have a nearby cell type-enriched THS are associated with a single such site. As shown in Figure S4.2B, a greater number of ATAC-seq reads were observed across the gene bodies of these THS-proximal gene sets for the cell type they were originally identified in. In other words, the stem cell THS-proximal genes showed more ATAC-seq reads across their gene bodies in the stem cell dataset compared to the mesophyll dataset, and vice versa (Figure S4.2B). These results suggest that the proximal genes of enriched THSs have more accessible chromatin across their gene body, and therefore are more likely to be highly transcribed in the cell type where the THS is enriched. It was also found that the majority of ATAC-seq reads relative to these genes are localized proximally upstream of the transcription start sites (TSSs) and downstream of the transcript end site (TES) (Figure S4.2B). Minimal transposase bias was found in these analyses, and was primarily confined to gene bodies in both sets of genes. Importantly, however, such bias was not observed at the specific sites where the majority of our enriched THSs were located (Figure 4.2B, S2B).

We next used AgriGO (Du et al., 2010; Tian et al., 2017) to identify overrepresented Gene Ontology (GO) terms within the THS-proximal gene sets for each cell type (Table S4.3). We focused our analysis only on the GO terms that had a false discovery rate (FDR) of less than 0.05. This analysis revealed that many of the genes associated with the stem cell-enriched THSs are involved in the regulation of cell differentiation and shoot development, while the genes proximal to the mesophyll-enriched THSs were predominantly involved in response to biotic and abiotic stimuli, which is consistent with the known functions of these two cell types (Figure 4.2D).

Enriched motif analysis and identification of cell type-specific transcriptional regulatory networks

As described above, THSs represent more accessible chromatin regions, which likely contain TF binding sites that can recruit TFs to regulate the expression of nearby genes. To identify specific transcription factors that may play important roles in establishing and maintaining the stem cell and mesophyll cell fates during development, we first identified sequence motifs that were overrepresented in cell type-enriched THSs. This was achieved by performing a MEME-ChIP analysis on the repeat-masked sequences within these

THS regions (Machanick and Bailey, 2011). We discovered a total of 364 overrepresented motifs within the stem cell-enriched THSs and 291 motifs within mesophyll-enriched THSs (Figure 4.3A and Table S4.4). We identified 244 motifs that were present in both cell-type-enriched THSs while 120 and 47 motifs were uniquely found within the stem cell-enriched and mesophyll-enriched THSs, respectively (Figure S4). Although the identification of cell type-unique sequence motifs points to the TFs that are likely important for biology of the two cell types, the more relevant question is whether these TFs are preferentially expressed in the corresponding cell type. Therefore, to determine which TFs show differential expression in one cell type or other, we ranked the TFs that bind all identified motifs based on their expression level in each cell type using publicly available RNA-seq and microarray data (Endo et al., 2014; You et al., 2017). This was done by first calculating the relative expression rank by percentile for each gene in these datasets (see Experimental Procedures). Then, the difference in expression rank for each TF with an overrepresented motif in each cell type was measured between the two cell types (Table S4.5). Importantly, we also confirmed that the highest and lowest expressed genes in each cell type showed very high and very low chromatin accessibility, respectively, as expected (Figure S4.5). In total, we identified 23 stem cell-enriched and 129 mesophyll-enriched TFs that have at least a two fold difference in their relative expression ranking between cell types (Figure 4.3, Figure S4.4, and Table S4.5). We then used these TF sets as input for the STRING database, which combines both known protein-protein interactions and functional interactions among genes (*e.g.* co-expression, text mining association, interactions in orthologs from other species, etc.) to infer and predict functional connections between a set of input genes (Szklarczyk et al., 2017).

Using this approach, we discovered a putative stem cell-specific functional network of 5 interconnected TFs that belong to two distinct families: INDETERMINATE DOMAIN C2H2 zinc finger protein family (IDD) and GATA factor zinc finger transcription factor protein family (Figure 4.3B and Figure S4.6A). The *Arabidopsis* IDD family of TFs has 16 members, which are involved in promoting gibberellin signaling, auxin biosynthesis and transport, and lateral organ differentiation, but are best known for their control of tissue formation during root development (Cui et al., 2013; Yoshida et al., 2014; Moreno-Risueno et al., 2015). The GATA TF family is comprised of approximately 30 members, which can be

divided into four subfamilies (Behringer and Schwechheimer, 2015). Of these subfamilies, the best studied TFs are the members of the B-GATA subfamily, including GNC and its paralog GNL, which are involved in the control of greening and regulation of plant development downstream of the hormones gibberellin, cytokinin, and auxin (De Rybel et al., 2010; Richter et al., 2013; Ranftl et al., 2016).

We carried out FIMO analysis (Grant et al., 2011) using all the stem cell THS sequences to identify motif occurrences and thus predicted binding sites for each of the four TFs that showed evidence of connectivity in the STRING-derived regulatory network: INDETERMINATE DOMAIN 2 (IDD2), INDETERMINATE DOMAIN 7 (IDD7), GATA TRANSCRIPTION FACTOR 1 (GATA1), and GATA TRANSCRIPTION FACTOR 15 (GATA15). The fifth TF from the STRING network, JKD, was not included in this analysis because it is also a member of the IDD family and we simplified our analysis by keeping only the two members of each family. The identified predicted binding sites of these four TFs were then used to locate the nearest TSS to identify the set of predicted target genes for each of the four TFs (Figure 4.4A and B). Using this approach, we discovered 3,218 predicted target genes for GATA15, 5,946 for IDD2, 3,603 for GATA1, and 5,322 for IDD7. Out of the 9,962 target genes for these TFs, 569 genes are predicted targets for all four TFs. We performed GO analysis on this group of genes using AgriGO (Figure 4.4C). The GO terms overrepresented in this analysis revealed that many of the target genes predicted to be regulated by IDD and GATA TFs are involved in control of auxin-mediated signaling, regulation of transcription, and shoot development (Figure 4.4C). A STRING network of interactions among these target genes, under high stringency (a minimum interaction score of 0.700), is shown in Figure S4.7. Notably, we found that the known stem cell regulator CLV3 is a target of this IDD/GATA gene regulatory network.

The 129 mesophyll-enriched TFs whose motifs were overrepresented in the mesophyll-enriched THSs (Figure 4.3A) had a high density of functional interconnections when analyzed with the STRING database (Figure S4.6B). We identified three major mesophyll-specific sub networks of TFs. The largest sub network was comprised of 41 extensively interconnected TFs, including 10 members of WRKY and 11 members of ERF family of TFs, which are known to regulate various biotic and abiotic stress responses

(Yang et al., 2005; Chen et al., 2010; Son et al., 2012; Birkenbihl et al., 2017; Bolt et al., 2017; Chen et al., 2017; Scarpeci et al., 2017). Seven out of the eight TFs in the second sub network belong to the TEOSINTE BRANCHED 1, CYCLOIDEA, PCF1 (TCP) family, which is known to control plant growth and organ development, including leaf development (Koyama et al., 2007; Li, 2015; Alvarez et al., 2016; Danisman, 2016). The third sub network included eight well-connected TFs. Among these are three members of the PIF family: PHYTOCHROME INTERACTING FACTOR 3-LIKE 5 (PIL5), PHYTOCHROME INTERACTING FACTOR 3-LIKE 6 (PIL6), and PHYTOCHROME-INTERACTING FACTOR 7 (PIF7). Two additional TFs found in this subnetwork, BES1-INTERACTING MYC-LIKE 1 (BIM1) and BRASSINAZOLE-RESISTANT 1 (BZR1), are involved in the brassinosteroid (BR) hormone signaling pathway. PIFs belong to the bHLH family of TFs and are best known as negative regulators of chlorophyll biosynthesis and photomorphogenesis (Stephenson et al., 2009; Leivar and Quail, 2011; Zhang et al., 2013; Pfeiffer et al., 2014). BRs are important regulators of many aspects of plant growth and developmental processes including cell elongation, responses to biotic and abiotic stresses, and photomorphogenesis (Saini et al., 2015; Singh and Savaldi-Goldstein, 2015). We decided to explore this PIF/BR regulatory sub network in more detail since both PIFs and BRs have been implicated in the regulation of chloroplast biogenesis (Stephenson et al., 2009; Leivar and Quail, 2011; Yu et al., 2011; Zhang et al., 2013; Pfeiffer et al., 2014), and therefore may directly affect the physiology and development of mesophyll cells.

As with the IDD/GATA regulatory network in the stem cells, our next goal was to use 4 mesophyll-enriched TFs that belong to two different families of TFs (two TFs representing each family) to identify their putative target genes. In this case, we included two additional TFs, out of the 129 mesophyll-enriched TFs, that belong to the bZIP family: bZIP16, and bZIP53. We chose to include bZIP TFs because it has been recently shown that PIF and bZIP TFs HY5 and HYH interact with each other and form heterodimers to antagonistically regulate chlorophyll biosynthesis (Chen et al., 2013; Toledo-Ortiz et al., 2014) and the production of Reactive Oxygen Species (ROS) during deetiolation (Chen et al., 2013).

Using all the mesophyll THS sequences we performed FIMO analysis (Grant et al., 2011) to identify predicted binding sites for each of the four TFs of interest: PIL5, PIL6, bZIP16, and bZIP53. We

then identified putative target genes by assigning each predicted binding site to its nearest TSS. As seen for the stem cell TFs, all four of the mesophyll-enriched TFs also showed extensive co-regulation of common target genes. We then performed GO analysis on the set of 487 target genes putatively regulated by all four TFs (Figure 4.4B). Many of the GO terms identified describe known functions of mesophyll cells including response to abiotic stimulus and light stimulus, photosynthesis-light reaction, and carbohydrate biosynthetic process (Figure 4.4C and Table S4.6). These results suggest that the PIL5, PIL6, bZIP16, and bZIP53 TFs likely play important roles in regulating mesophyll physiological functions.

We discovered that many of the putative target genes of the IDD and GATA TFs in stem cells, and PIFs and bZIPs in mesophyll cells, are TFs themselves. This finding that TFs may regulate the expression of other TFs has been recently reported in plants (Heyndrickx et al., 2014; Pfeiffer et al., 2014; Sullivan et al., 2014). In our case, these results allude to the presence of cell type-specific transcription factor cascades, which may regulate important biological processes in stem cells and mesophyll cells (GO term: regulation of transcription, Figure 4.4C). To explore TF-to-TF connections in greater detail, we explored putative regulatory connections among the stem cell and mesophyll TFs to illuminate how they might regulate each other. Previous studies have used similar models with great success in order to build de-novo TF regulatory networks for 41 human cell types (Neph et al., 2012) and Arabidopsis seedlings under different environmental conditions (Sullivan et al., 2014). The model presented in Figure 4.5A describes the logic of this analysis, in which each TF can bind to its recognition motif found within its own regulatory regions and/or within the regulatory regions of other TF genes. For instance, the proximal regulatory region of a hypothetical transcription factor gene (TF5) contains DNA-binding motifs of four other TFs: TF1, TF2, TF3, and TF4. The DNA-binding motif of the TF5 is, on the other hand, found in the upstream region of TF4, which also has its recognition motif present in the regulatory regions of TF1 and TF2 (Figure 4.5A). Thus, an extensive co-regulatory network of multiple TFs can be mapped in this manner.

Using the strategy described in Figure 4.5A for the predicted target genes for each TF, we derived more comprehensive stem cell-specific and mesophyll-specific putative regulatory circuitries of TFs, further uncovering complex combinatorial interactions among TFs within these networks (Figure 4.5B).

For example, in the stem cell-specific TF network, IDD7 appears to regulate itself and three other TFs: IDD2, GATA1, and GATA15, but not JKD or MYB13 (Figure 4.5B). On the other hand, JKD may regulate the expression of IDD2, IDD7, and GATA1, but not that of GATA15 and MYB13. GATA1 and MYB13 seem to regulate each other, while GATA15 appears to be most downstream component in this TF hierarchy since it does not regulate any other TF in this network.

Similarly, in the mesophyll-specific TF regulatory network, PIL6 and BZR1 seem to regulate both themselves and each other. bZIP16 appears to be at the apex of this regulatory module since it regulates three different TFs, while all others regulate the expression of two or fewer different TFs. Importantly, this model predicts that PIL6 and bZIP16, as well as PIL5 and bZIP53, co-regulate the expression of RVE1, which resembles the coordinated TF interaction previously described for another pair of bHLH/bZIP TFs: PIFs and HY5/HYH (Chen et al., 2013; Toledo-Ortiz et al., 2014).

DISCUSSION

Cell type-specific THSs contain *cis*-regulatory elements relevant to the physiology of stem cells and mesophyll cells

Since fully differentiated mesophyll cells in leaves are at the very end of the differentiation process from stem cells in the meristem, it was perhaps surprising to find that more than 91% of the reproducible mesophyll THSs identified by peak calling alone were also present in stem cells (Figure 4.1C). These results indicated that the accessible chromatin regions of these two cell types are not as different as we originally anticipated. There are several possibilities that can potentially explain why 91% of mesophyll THSs are also detected in stem cells. For instance, it has been shown that the gene expression within the CLV3 domain is not homogeneous and that there are important transcriptional differences between the L1 layer and inner SAM layers (Yadav et al., 2014). Similar expression heterogeneity may exist between the two different types of mesophyll cells: spongy and palisade, within the leaf tissue. This technical limitation of our approach could be overcome in the future with a single-cell ATAC-seq approach. Another explanation

could be that other regulatory proteins or remodelers are already bound at these THS sites in stem cells keeping these areas open and poised for later TF binding during mesophyll differentiation. In addition, the contamination of stem cell nuclei with differentiated cell nuclei during purification could potentially account for the observed THSs similarities between the two cell types. However contamination is unlikely for several reasons: a) previous studies have demonstrated very high specificity of INTACT (over 90% of purity of isolated nuclei, Deal and Henikoff 2010, 2011), b) clear NTF signal observed only in stem cells (Figure 4.1A), and c) low yield of INTACT-isolated nuclei from CLV3 line is consistent with the expected yields given the starting amount of tissue, suggesting that majority of nuclei are from stem cells (Table S4.7). In spite of all this, we were able to identify several thousand cell type-unique THSs (Figure 4.1C) that were only detected in one cell type or the other. Since the majority of THSs were shared between the two cell types, we performed a quantitative analysis to identify THSs that were differentially accessible between stem cells and mesophyll cells. This analysis led to the identification of several thousand more THSs in each cell type than were identified by the absolute, all-or-nothing, peak calling strategy. We assigned each of these cell type-enriched THSs to their nearest TSS as the putative target gene regulated by the differential accessibility event. Gene Ontology (GO) analysis of the stem cell THS-proximal genes identified overrepresented GO terms that describe known functions of the SAM stem cells in regulating cell differentiation and organ development (Figure 4.2D). Similarly, we identified GO terms for the mesophyll THS-proximal genes that are consistent with established roles of mesophyll cells in mediating the responses of various biotic and abiotic stresses (Figure 4.2D).

Overall, these results indicate that INTACT-ATAC-seq allows us not only to successfully identify cell type-enriched THSs, but also that these THSs likely contain regulatory elements that are highly relevant for the biology of these two cell types.

Mesophyll-enriched THSs are already accessible in stem cells

In comparing open chromatin regions between cell types, we discovered that the stem cell-enriched THSs tend to be much more highly accessible in stem cells relative to mesophyll, but that these regions still

showed a low level of accessibility in the mesophyll cell type (Figure 4.3B). This is consistent with our previous observation that the root epidermal hair and non-hair cell types show mainly quantitative, rather than qualitative differences in chromatin accessibility (Maher et al., 2017) These results also suggest that, at least in the *Arabidopsis* cell types examined, a given regulatory region is never completely inaccessible in any cell type, and this may simply reflect the proportion of cells in the population in which a TF binding event is occurring at that location.

When we examined chromatin accessibility at mesophyll-enriched THSs, we found that while accessibility was far higher in mesophyll cells, the stem cells also showed significant accessibility at these sites (Figure 4.3B). Thus, while stem cell-enriched THSs represent regions that are highly accessible in stem cell and far less so in mesophyll, the mesophyll-enriched THSs tend to already be highly accessible in the progenitor stem cells. This suggests that even mesophyll cell-enriched THSs are available for TF binding in stem cells, and this phenomenon could underlie the developmental flexibility of stem cells. Whether this observation is a unique characteristic of the SAM stem cell chromatin or more universal feature of any plant stem cell chromatin in comparison to differentiated cells remains to be tested. Regardless, we can hypothesize that one of the reasons why pluripotent stem cells in the SAM would maintain more accessible regulatory elements is to allow them flexibility to change their transcriptome in response to stimuli. In other words, by being more open, the stem cell chromatin is more "primed" for transcriptional reprogramming, thereby endowing stem cells with the plasticity to efficiently respond to differentiation cues.

Identified cell type-specific transcriptional modules are likely important for the establishment of lineage-specific regulatory programs in each cell type

In a search for TFs that should be relevant to the biology of each cell type, we analyzed the differentially enriched THS regions to identify putative cell type-specific *cis*-regulatory motifs as well as the TFs that bind them (Figure 4.3A). Using publicly available expression data for these two cell types, we found TFs that were differentially expressed in each cell type and whose motifs were also overrepresented in THSs

enriched in that cell type. We analyzed these cell type-enriched TFs using the STRING database to identify modules of TFs that might act coordinately in each cell type (Figure 4.3B).

Following the logic that the identified TF motifs likely represent true TF binding events when they occur within an open chromatin region of the corresponding cell type, we were able to predict the target genes of TFs of interest (Figure 4.4A). We found that in each cell type, cell type-enriched TFs showing connections in the STRING database also tended to co-regulate many genes (Figure 4.4C). In each case, a relatively large gene set appeared to be co-regulated by all four TFs, and GO analysis of these gene sets revealed functions consistent with the biology of each cell type. Thus, while using predicted target genes, rather than direct measurement of TF binding by ChIP-seq, may lead to the inclusion of false positive binding events, there is strong evidence that many true positives exist among the putative target genes.

The TF modules we identified in each cell type by expression and STRING analysis were then used to define regulatory interactions between cell type-specific TFs (Figure 4.5B). The predicted combinatorial interactions among TFs within these regulatory networks were extensive and likely play important roles in establishing and/or maintaining cell type-specific transcriptional programs during differentiation. These new hypotheses can now be experimentally tested. For instance, the STRING-derived stem cell TF network was comprised of the members of IDD and GATA TF families. While the functions of individual members of IDD and GATA families of TFs are more or less known, to our knowledge, the functional interactions between GATA and IDD TFs have never been proposed or studied. In addition, IDDs are known to regulate lineage identity, patterning, and formative divisions throughout *Arabidopsis* root growth (Moreno-Risueno et al., 2015) but have not been implicated in a similar role during vegetative meristem development and differentiation, which our data now suggest. Interestingly, *CLV3* itself was identified as a target gene of the IDD/GATA regulatory network (Figure S4.7), further supporting our hypothesis that this regulatory circuitry may play an important role in stem cell homeostasis. One way to test these hypotheses is by manipulating the expression of the IDD/GATA TFs specifically in the SAM stem cells. For instance, the *CLV3* regulatory sequences can be used to drive the expression of RNAi or artificial microRNA constructs to specifically knock down the IDD/GATA TFs in the stem cell population. In addition, an inducible

overexpression system may be utilized to overproduce IDD/GATA TFs specifically in the SAM in order to monitor chromatin and transcriptional changes. The results from these experiments will address whether the IDD/GATA TFs indeed act as important regulators of SAM function, and further characterize these regulatory connections.

It has been previously demonstrated that the PIF and bZIP transcription factors HY5 and HYH antagonistically regulate chlorophyll biosynthesis during seedling development (Chen et al., 2013; Toledo-Ortiz et al., 2014). Our results now indicate that other members of the PIF and bZIP families of TFs may cooperate in regulating chlorophyll biosynthesis and light responses in mesophyll cells. This intricate cooperation between two TFs with potentially opposite functions may allow mesophyll cells to fine-tune their transcriptional programs in response to various hormonal and environmental stimuli during leaf development and/or throughout daily light/dark cycles. Experimental manipulations to modulate the expression of mesophyll PIFs and bZIPs by overexpressing or suppressing these TFs specifically in mesophyll cells can now be performed to test these new hypotheses.

INTACT-ATAC-seq as a powerful technique for predicting cell type-specific transcriptional regulatory networks

In this study, we combined the INTACT method with ATAC-seq to successfully isolate nuclei from two specific cell types and locate differentially accessible chromatin regions containing important *cis*-regulatory motifs. This allowed us to identify the TFs that likely bind at these regulatory elements and to construct cell type-specific TF regulatory networks. Our data provide new hypotheses and will serve as a valuable resource that can be used to derive further de-novo models of transcriptional regulatory networks relevant to cell fate specification during differentiation. These hypotheses can be experimentally tested and the results from these experiments used to further build upon and expand our current understanding of the regulatory mechanisms controlling cell fate and function during plant development.

METHODS

Plant growth conditions and transformation

Arabidopsis thaliana plants of the Columbia (Col-0) ecotype were grown in soil or on half-strength Murashige and Skoog (MS) media (Murashige and Skoog, 1962) agar plates in growth chambers under fluorescent lights, with 16 hour light-8 hour dark cycle at 20°C. All seeds, either on agar plates or in the soil, were stratified for three days at 4°C prior to moving them to the growth chambers. Plasmid constructs were introduced into *Agrobacterium tumefaciens GV3101* strain by electroporation. Plant transformation was performed using the floral dip method with corresponding *Agrobacterium* clones (Clough and Bent, 1998). Primary transformant seedlings (T_1) were first selected on half-strength MS media agar plates containing 35mg/L hygromycin, 25mg/L glufosinate ammonium (BASTA), and 100mg/L timentin, and then transferred to soil.

Plasmid DNA constructs

We used the promoters of *CLAVATA3* (*CLV3)* and *Rubisco small subunit 2B* (*RBC)* genes, known to be exclusively transcribed in stem cells and mesophyll cells, respectively (Schoof et al., 2000; Sawchuk et al., 2008), to drive cell-type specific expression of the *NTF* gene. To construct the *CLV3p::NTF* plasmid, the *NTF* coding sequence (Deal and Henikoff, 2010) was first PCR amplified using the forward primer 5'-catctgcagatgaatcattcagcgaaaacc-3', introducing a *PstI* restriction site (underlined), and the reverse primer 5'-catggatcctcaagatccaccagtatcctc-3', introducing a *BamHI* restriction site (underlined). The PCR product was then digested with *PstI/BamHI* enzymes and ligated into *PstI/BamHI* sites of the *pBU14* plasmid containing the *CLV3* promoter and terminator sequences (Brand et al., 2002). The *ACT2p::BirA* plasmid has been previously described (Zilberman et al., 2008). The *RBCp::NTF* construct was produced by removing the *ADF8* promoter from the previously described *ADF8p::NTF* plasmid (Deal and Henikoff, 2010) via *XmaI* and *NheI*, and replacing it with a 1.5 kb upstream fragment of *RBC*, including the start codon.

Microscopy

The shoot apical meristems of six days old *T_3 CLV3p::NTF;ACT2p::BirA* seedlings and leaves 5 and 6 of three weeks old *T_3 RBCp::NTF;ACT2p::BirA* plants were observed using a Leica SP8 confocal laser scanning microscope with 42x immersion objectives to confirm proper expression and localization of the NTF protein. GFP fluorescence was visualized by excitation at 488 nm. The shoot apical meristems were visualized by manually removing the surrounding leaf tissue and imaging the shoot apical meristem from the side. Mesophyll cells were visualized by dissecting the leaf with a scalpel and imaging the cross section. For each sample, the tissue was immersed in perfluoroperhydrophenanthrene, covered with a cover slip, and imaged.

Nuclei isolation by INTACT

Purification of nuclei from specific cell types using the Isolation of Nuclei TAgged in specific Cell Types (INTACT) method was performed as described previously (Bajic et al., 2018) with following modifications: 0.5 grams of freshly harvested plant tissue was used for nuclei isolation; *CLV3p::NTF;ACT2p::BirA* transgenic seedlings were collected at 6 days of age and were processed by grinding the tissue to a fine powder using liquid nitrogen. Leaves 5 and 6 from three week old plants with the *RBCp::NTF;ACT2p::BirA* transgenics were collected and finely chopped with a razor blade in Nuclei Purification Buffer (NPB) on ice. For both preparations, a volume of 10 µl of Streptavidin M280 magnetic beads was used to capture biotinylated nuclei.

Compared to previous INTACT-ATAC-seq experiments using root tissue, we observed a much higher level of organelle contamination in purified nuclei from shoot tissue of both the *CLV3p::NTF;ACT2p::BirA* and *RBCp::NTF;ACT2p::BirA* transgenic lines, as revealed by the large percentage of organelle-derived reads in our datasets. During mesophyll nuclei purification in particular, we observed many large clusters of nuclei associated with magnetic beads, suggesting that chloroplasts and mitochondria may become trapped within these clusters. Further optimization of the INTACT procedure on green tissue will likely eliminate this issue. For instance, it may be necessary to use even smaller amounts of starting tissue, to further decrease the amount of streptavidin beads used in order to decrease bead

clustering, and to add additional washing steps or use higher non-ionic detergent concentrations during purification.

Assay for transposase accessible chromatin (ATAC) and library preparation

It is important to note that all INTACT-purified nuclei were isolated and used fresh, and were never frozen prior to the transposase integration reaction. Transposase tagmentation and sequencing library preparations were then carried out as previously described (Bajic et al., 2018). Briefly, 25,000 purified nuclei were resuspended in a 50 µl transposase integration reaction and incubated at 37° C for 30 min using Nextera reagents (Illumina, FC-121-1030). Tagmented DNA was purified using the MiniElute PCR purification kit (Qiagen), eluted in 11 µl of elution buffer, and then the sample was amplified using 2X high fidelity PCR mix (NEB) with custom barcoded primers for 10-12 total PCR cycles. The amplified ATAC-seq libraries were purified using AMPure XP beads (Beckman Coulter) and then quantified by qPCR using the NEBNext library quant kit (NEB). The quantified libraries were analyzed using a Bioanalyzer high sensitivity DNA chip (Agilent) before pooling and next-generation sequencing.

High throughput sequencing

Next-generation sequencing was done using the NextSeq 500 instrument (Illumina) at the Georgia Genomics Facility at the University of Georgia. All libraries were pooled and sequenced in the same flow cell using paired-end 36 nt reads.

Sequence read mapping, processing, and visualization

Sequencing reads were mapped to the *Arabidopsis thaliana* genome (version TAIR10) using Bowtie2 software (Langmead and Salzberg, 2012) with default parameters. Mapped reads in *.sam* format were converted to *.bam* format and sorted using Samtools 0.1.19 (Li et al., 2009). Mapped reads were filtered using Samtools to retain only those reads that had a mapping quality score of 2 or higher (Samtools "*view*" command with option "*-q 2*" to set mapping quality cutoff). These reads were further filtered with Samtools

to keep only the reads that mapped to nuclear chromosomes, thereby removing reads that mapped to either the chloroplast or mitochondrial genomes. Finally, the stem cell and mesophyll cell datasets were also processed such that the experiments within a biological replicate had the same number of mapped reads prior to further analysis (Samtools "*view*" command with option "-*c*" to count the number of aligned reads in each dataset and "-*S*" to scale down by the numerical fraction the number of aligned reads to be kept). For visualization, the filtered, sorted, and scaled *.bam* files were converted to the bigwig format using the "*bamcoverage*" script in deepTools 2.0 (Ramirez et al., 2016) with a bin size of 1 bp and RPKM normalization. Heatmaps and average plots displaying ATAC-seq data were also generated using the "*computeMatrix*" "*plotHeatmap*" and "*plotProfile*" functions in the deepTools package. Genome browser images were made using the Integrative Genomics Viewer (IGV) 2.3.68 (Thorvaldsdottir et al., 2013) with bigwig files processed as described above.

Peak calling to detect transposase hypersensitive sites (THSs)

Peak calling on ATAC-seq data was performed using the "*Findpeaks*" function of the HOMER package (Heinz et al., 2010) with the parameters "-*minDist 150*" and "-*region*". These parameters set a minimum distance of 150 bp between peaks before they are merged into a single peak and to allow identification of variable length peaks, respectively. We refer to the peaks called in this way as "transposase hypersensitive sites," or THSs. To deepen our analysis and increase the resolution and number of THSs called in the two cell types we utilized an additional parameter when comparing the degree of accessibility between the two cell types. The additional parameter "-*regionRes 1*" separated larger THSs into several smaller THSs without affecting the way in which THSs that were several hundred base pairs in size or smaller are called. For calling peaks in genomic DNA we similarly employed the "*Findpeaks*" function using the parameters "-*minDist 150*" and "-*region*".

Genomic distribution of THSs

The distribution of THSs relative to genomic features was identified using the PAVIS web tool (Huang et al., 2013) with "upstream" regions set as the 2,000 bp upstream of the annotated transcription start site, and "downstream" regions set as the 1,000 bp downstream of the transcript end site.

THSs enriched in a specific cell type

The number of reads (counts) present in each cell type at all the THSs called in stem cell and mesophyll ATAC-seq data was obtained using HTSeq's *htseq-count* script (Anders et al., 2015). Two replicates of each cell type were counted and the counts were processed using DESeq2 (Love et al., 2014). THSs that had an adjusted p-value ≤ 0.05 and log fold change of 1 or more for a specific cell type were identified as THSs enriched in that cell type.

Transcription factor motif analysis

ATAC-seq THSs that were enriched in one cell type or the other were used for motif analysis. The cell type-enriched THSs from each cell type were first adjusted to the same size (300 bp). The sequences present in these scaled regions were isolated using the Regulatory Sequence Analysis Tools (RSAT), which also masks any repeat sequences (Medina-Rivera et al., 2015). The masked sequences were run through MEME-ChIP with default parameters to identify motifs that were present in higher proportions than expected by chance (i.e. overrepresented motifs) (Machanick and Bailey, 2011). The DREME, MEME, and CentriMo programs were used to identify overrepresented motifs, and Tomtom matched these motifs to previously reported TF binding motifs. Motifs from both Cis-BP (Weirauch et al., 2014) and DAP-seq (O'Malley et al., 2016) databases were used in all motif searches, and only those that had an E-value ≤ 0.05 were considered significant.

Assignment of THSs to nearby genes

For each ATAC-seq dataset, the THSs were assigned to putative target genes using the "*TSS*" function of the PeakAnnotator 1.4 program (Salmon-Divon et al., 2010). This program assigns each THS to the closest

transcription start site regardless of whether it is upstream or downstream from the THS, and reports the distance from the peak center to the TSS based on the genome annotations described above.

Publicly available RNA-seq and microarray data

Published RNA-seq data from the CLV3-expressing cell population of shoot meristems (You et al., 2017), isolated from 21 day old plants, and microarray data from mesophyll cells isolated at ZT04 from 10 days old cotyledons grown under a long day (LD) cycle (GSM1219271, (Endo et al., 2014)), were used to define TFs that were differentially expressed in the stem cells relative to the mesophyll cells, and vice versa.

Calculating the relative expression ranks of TFs in RNA-seq and microarray data sets

Within the stem cell RNA-seq dataset, the genes were considered expressed if the FPKM value was ≥1, and 17,811 genes satisfied this criterion. Within the mesophyll microarray data, there were 28,583 expressed genes. For each data set, we first arranged the genes based on the level of their expression from highest to lowest. Next, for each TF of interest we calculated the percentile of its expression relative to total number of expressed genes in each data set and then measured the difference in its expression rank between the two cell types (Table S4.5). For instance, GATA15 TF ranked as the 2085th most highly expressed gene in the stem cell RNA-seq data set, which equals 11.7% (2,085/17,811) in expression ranking. The same TF in the mesophyll expression data set ranked 24,755th most highly expressed, which is 86% (24,755/28,583) in expression ranking. We then calculated the difference in expression ranking of GATA15 between the two cell types by dividing the relative expression ranks in percentages (11.7/86). TFs that have at least a two fold difference in their relative expression ranking between cell types were considered as more highly expressed in one cell type or the other.

Protein interaction analysis using STRING

Gene lists were analyzed using the STRING database to identify groups of TFs that have predicted interactions based on the co-expression analysis, publication co-occurrences, colocalization, gene

orthology, and experimental information such as yeast-2-hybrid interactions (Szklarczyk et al., 2017). The network connections between the submitted TFs were visualized by their confidence score, where a thicker line indicates a higher interaction score. Furthermore, the network was subdivided into differentially colored nodes by the Markov Cluster Algorithm score set to 3.0. This allows for the detection of genes with some evidence for interactions, but whose association does not pass the interaction threshold required to have a bona fide connection. The scale of interaction scores in STRING is as follows: 0.15=low confidence, 0.4=medium confidence, 0.7=high confidence, and 0.9=highest confidence. The minimum interaction threshold used in this study was set to at least 0.400 or 0.700. The inputs used for the STRING database were the *Arabidopsis* gene IDs.

Defining predicted binding sites for transcription factors

We used FIMO (Grant et al., 2011) to identify TF motif occurrences within the repeat-masked sequence of the *Arabidopsis* genome. Significant motif occurrences were those with a p-value < 0.0001. Predicted binding sites for a given TF were defined as motif occurrences that were present within THSs of a given cell type (see Figure 4.4A for a schematic diagram of this process).

Gene ontology analysis

Gene ontology (GO) analysis was carried out on gene lists using the AgriGO GO Analysis Toolkit, with default parameters (Du et al., 2010; Tian et al., 2017). GO terms that had a false discovery rate (FDR) of 0.05 or less were considered significant.

Accession numbers

The raw and processed ATAC-seq data described in this work has been deposited to the NCBI Gene Expression Omnibus (GEO) database under the record number GSE101940.

ACKNOWLEDGEMENTS

We would like to thank Kelsey Maher for constructive criticism of the manuscript. Funding for this work was provided by the National Science Foundation (Grant #IOS-123843) and Emory University.

Conflicts of Interest

The authors have no conflicts of interest to declare.

FIGURES

Fig 4.1

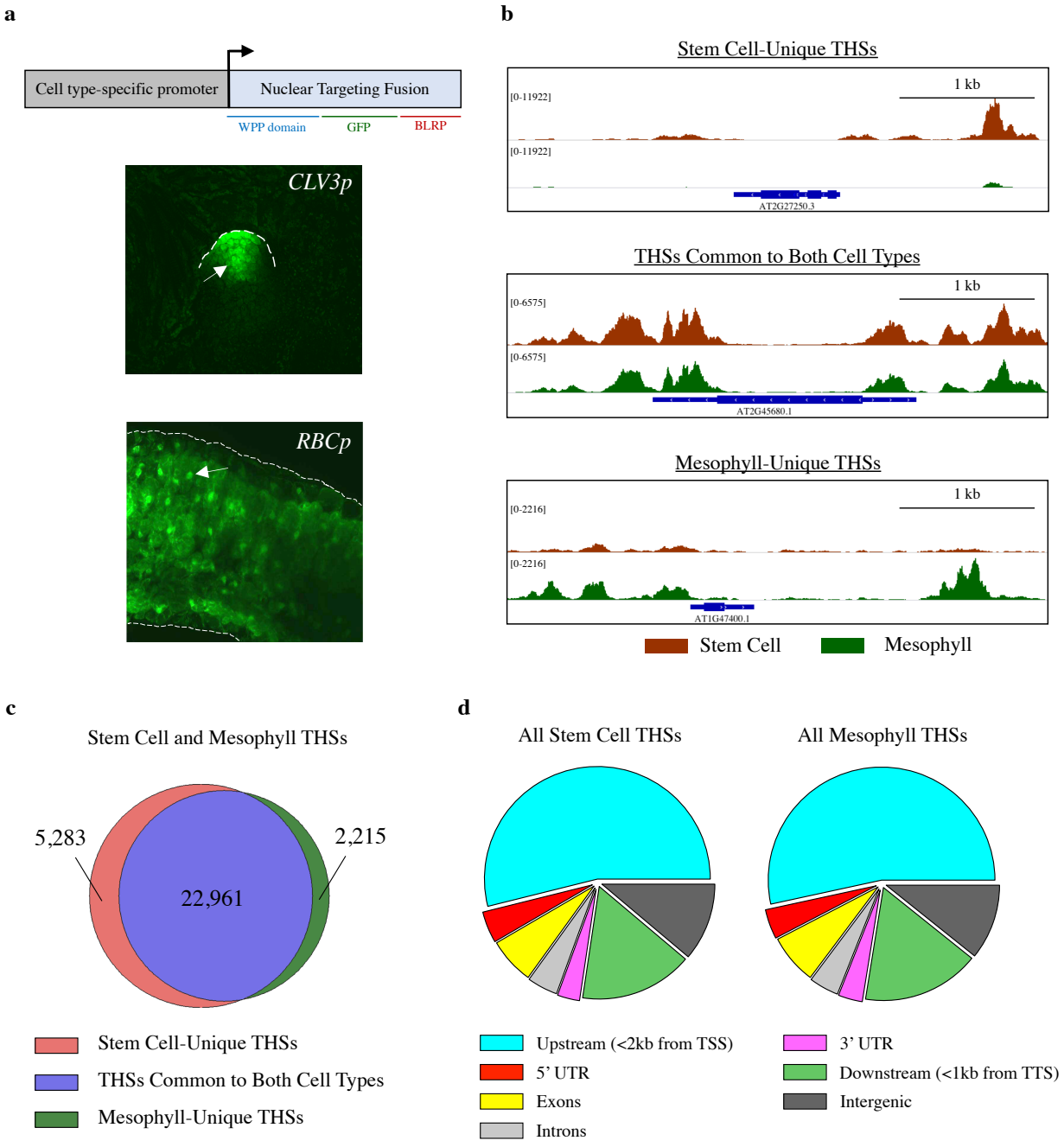
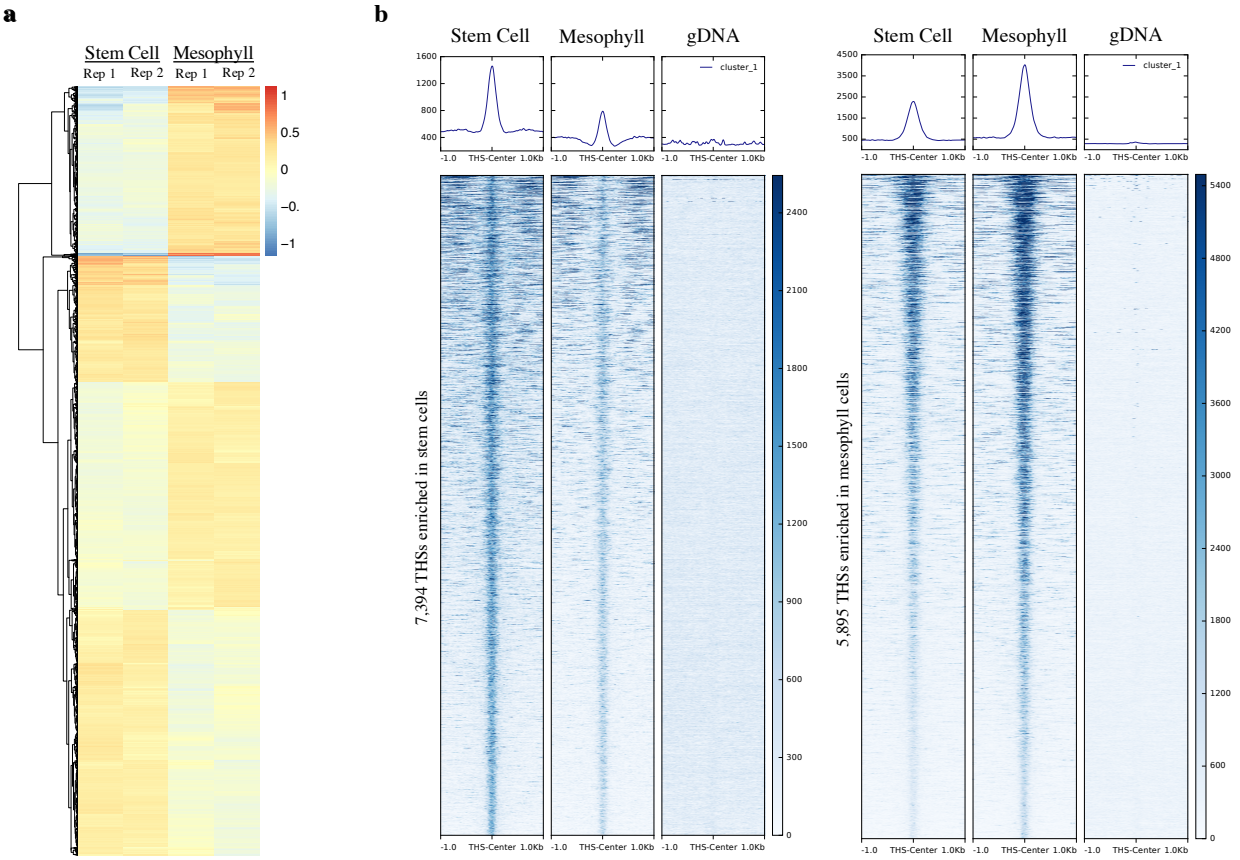
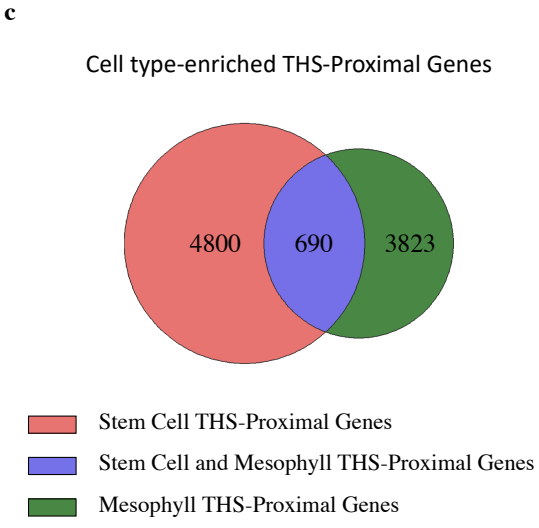


Figure 4.1 Characterization of INTACT transgenic lines and overview of ATAC-seq data from each cell type. (a) The upper panel is a schematic representation of the Isolation of Nuclei TAgged in specific Cell Types (INTACT) system for isolating nuclei from specific cell types. The Nuclear Targeting Fusion

(NTF) contains a WPP nuclear envelope-binding domain, Green Fluorescent Protein (GFP) for visualization, and a biotin ligase recognition peptide (BLRP), which can be biotinylated by the BirA biotin ligase. BirA is expressed constitutively while NTF is driven from a cell type specific promoter. When these transgenes are coexpressed in a cell the nucleus becomes biotinylated, allowing all nuclei of that cell type to be selectively purified with streptavidin beads. Below the gene diagram are confocal images of GFP expression in the *CLV3p:NTF;ACT2p:BirA* line (upper) and *RBCp:NTF;ACT2p:BirA* line (lower), showing NTF expression in the shoot apical meristem and mesophyll cells, respectively. Fluorescent nuclei are labeled with arrowheads. **(b)** Three Integrated Genome Viewer (IGV) snapshots of normalized ATAC-seq reads from shoot apical stem cell (red) and mesophyll (green) nuclei. Different categories of Transposase Hypersensitive Sites (THSs) are observed: Top panel) Stem cell-unique: THSs identified only in stem cells; Middle panel) Common to both cell types: THSs that were identified in both stem cells and mesophyll cells; and Bottom panel) Mesophyll-unique: THSs that were identified only in mesophyll cells. **(c)** Overlap of stem cell and mesophyll ATAC-seq THSs identified by peak calling in at least two biological replicates of that cell type. **(d)** Genomic distribution, generated using the software tool PAVIS, of all the THSs identified in two replicates for either stem cell or mesophyll ATAC-seq.

Fig 4.2





d

Stem Cell THS-Proximal Genes GO	FDR	GO Term Associated Genes
Multicellular organismal development	1.4e-10	416
Developmental process	1.1e-08	450
Organ development	1.5e-08	199
Anatomical structure development	2.3e-07	345
Response to gibberellin stimulus	0.00017	51
Shoot development	0.00066	87
Shoot system development	0.00083	87
Tissue development	0.0061	81
Gibberellic acid mediated signaling pathway	0.01	18
Cell differentiation	0.032	77
Mesophyll THS-Proximal Genes GO	**FDR**	**GO Term Associated Genes**
Response to stress	2e-14	421
Translational elongation	4.5e-11	168
Defense response	1.2e-08	161
Response to other organism	8e-05	117
Innate immune response	0.00013	347
Translation	0.00013	235
Response to wounding	0.00019	51
Response to water	0.00045	57
Response to water deprivation	0.00088	54
Protein metabolic process	0.0027	550

Figure 4.2 Chromatin accessibility differences between stem cells and mesophyll cells. (a) Heatmap showing the log ratio of normalized read count of the top 13,289 THSs that are statistically different

between stem cell and mesophyll ATAC-seq samples. Each line on the heatmap represents a single THS, and the values at that region are given for each of two replicates in each cell type. Increased chromatin accessibility between the four samples is colored red and decreased chromatin accessibility is colored blue, compared to an average value set to 0. **(b)** Normalized read signal in stem cell, mesophyll, and genomic DNA ATAC-seq samples over cell type-enriched THS regions. The left set of panels show ATAC-seq signal over the 7,394 stem cell-enriched THSs, while the right set of panels shows ATAC-seq signal over the 5,895 THSs enriched in mesophyll cells. **(c)** Each cell type-enriched THS was assigned to its nearest TSS as the putatively regulated target gene. Venn diagram shows the overlap of cell type-enriched THS-proximal genes. **(d)** Examples of 10 GO terms that were found only among the lists of genes that have a nearby cell type-enriched THS in a given cell type (i.e. from the non-overlapping portions of the diagram in (c)). FDR = False Discovery Rate, GO = Gene Ontology.

Fig 4.3

a

316 Motif-binding TFs Identified from Stem cell-enriched THSs

↓

$\frac{\text{TF expression rank in stem cell RNA-seq}}{\text{TF expression rank in mesophyll microarray}} \geq 2$

↓

23 TFs more highly expressed in stem cells

211 Motif-binding TFs Identified from Mesophyll-enriched THSs

↓

$\frac{\text{TF expression rank in mesophyll microarray}}{\text{TF expression rank in stem cell RNA-seq}} \geq 2$

↓

129 TFs more highly expressed in mesophyll

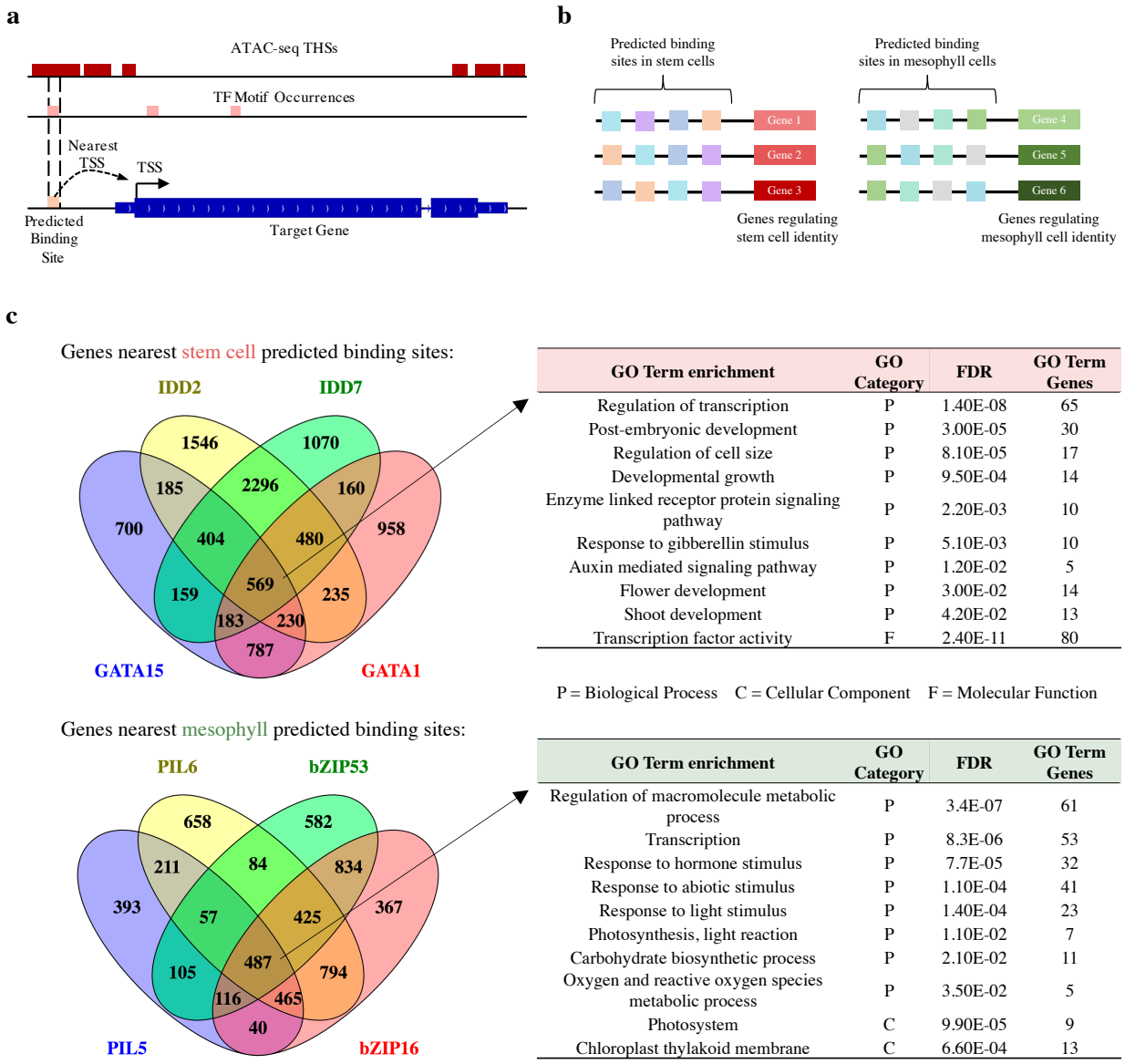
b

Stem Cell Enriched Motif Binding TFs		
Motif	**PWM**	**E-Value**
AT3G06740 (GATA15)		1.92E-02
AT3G50700 (IDD2)		9.84E-03
AT3G24050 (GATA1)		7.30E-05
AT1G55110 (IDD7)		1.35E-02
AT1G06180 (MYB13)		2.90E-02
AT5G03150 (JKD)		3.38E-05

Mesophyll Enriched Motif Binding TFs		
Motif	**PWM**	**E-Value**
AT5G17300 (RVE1)		3.22E-02
AT3G62420 (bZIP53)		9.70E-05
AT3G59060 (PIL6)		5.00E-33
AT2G20180 (PIL5)		1.00E-35
AT2G35530 (bZIP16)		4.80E-19
AT1G75080 (BZR1)		6.26E-06

Figure 4.3 Sequence motifs identified in cell type-enriched THSs. (a) Cell type-enriched THS sequences were centered and scaled to 300 bp, repeat masked, and analyzed with MEME-ChIP. Motifs that had an E-value equal to or less than 0.05 were considered significant. The 364 and 291 transcription factors (TFs) associated with overrepresented motifs from stem cell- and mesophyll-enriched THSs, respectively, were further separated by their ranked expression difference between previously reported stem cell RNA-seq and mesophyll microarray data. Only those TFs that had at least a two-fold higher expression rank difference for the cell type their motif was identified in were kept (Table S4.5). **(b)** Six TFs that potentially regulate transcriptional networks for each cell type, their position weight matrix (PWM), expression rank difference between the two cell types, and E-value from the MEME-ChIP analysis are shown for the stem cell (left) and mesophyll (right). The TFs are ranked by their difference in expression rank between the two cell types, with the highest expression rank difference for the corresponding cell type at the top.

Fig 4.4



GO Term enrichment	GO Category	FDR	GO Term Genes
Regulation of transcription	P	1.40E-08	65
Post-embryonic development	P	3.00E-05	30
Regulation of cell size	P	8.10E-05	17
Developmental growth	P	9.50E-04	14
Enzyme linked receptor protein signaling pathway	P	2.20E-03	10
Response to gibberellin stimulus	P	5.10E-03	10
Auxin mediated signaling pathway	P	1.20E-02	5
Flower development	P	3.00E-02	14
Shoot development	P	4.20E-02	13
Transcription factor activity	F	2.40E-11	80

P = Biological Process C = Cellular Component F = Molecular Function

GO Term enrichment	GO Category	FDR	GO Term Genes
Regulation of macromolecule metabolic process	P	3.4E-07	61
Transcription	P	8.3E-06	53
Response to hormone stimulus	P	7.7E-05	32
Response to abiotic stimulus	P	1.10E-04	41
Response to light stimulus	P	1.40E-04	23
Photosynthesis, light reaction	P	1.10E-02	7
Carbohydrate biosynthetic process	P	2.10E-02	11
Oxygen and reactive oxygen species metabolic process	P	3.50E-02	5
Photosystem	C	9.90E-05	9
Chloroplast thylakoid membrane	C	6.60E-04	13

Figure 4.4 Predicted binding sites and target genes for cell type-enriched TFs. (a) Schematic for identifying predicted binding sites using ATAC-seq THSs and FIMO-identified TF motif occurrences in the genome. These predicted sites were used to identify the nearest TSS to define the target gene potentially regulated by the TF. **(b)** Schematic for using the predicted binding sites (as shown in **a**) to identify genes that regulate cell identity, and which TFs control the expression of these genes. **(c)** Overlap of predicted target genes of IDD2, IDD7, GATA15, and GATA1 (top, left). The genes targeted by all four TFs (569) were analyzed with AgriGO, and the resulting GO terms that had an FDR value of 0.05 or less were retained.

A subset of these enriched GO terms is shown (top, right). Overlap of predicted target genes of PIL6, PIL5, bZIP53, and bZIP16 (bottom, left). The genes targeted by all four factors (487) were analyzed with AgriGO, and the resulting GO terms that had an FDR value of 0.05 or less were retained. A subset of these enriched GO terms is shown (bottom, right).

Fig 4.5

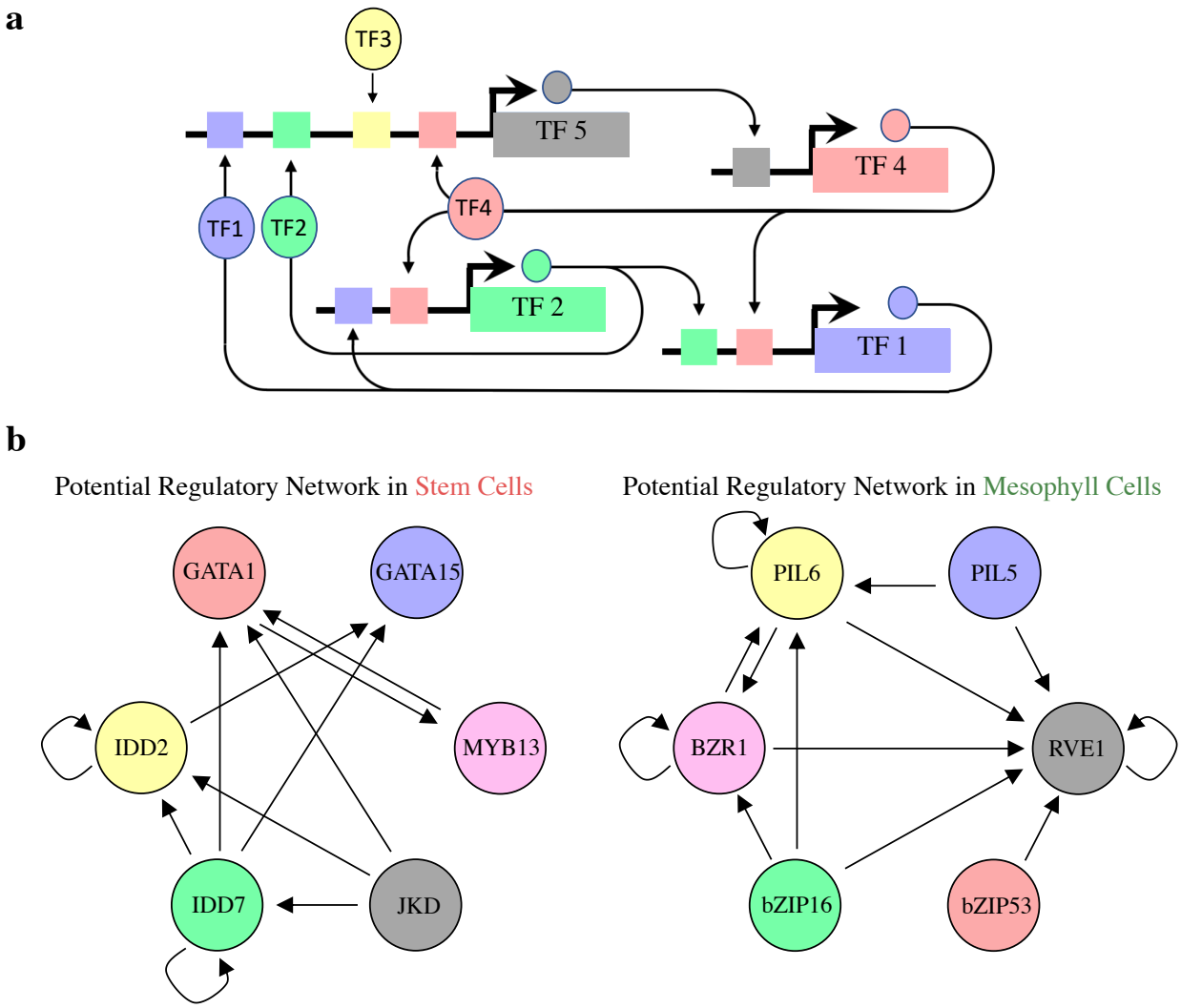


Figure 4.5 Proposed regulatory pathways for key transcription factors in stem cells and mesophyll cells. (a) Schematic for identifying regulatory interactions between transcription factors (TFs). A predicted binding site for a TF, such as TF5, may regulate the expression of another TF, such as TF4. Subsequently regulated TFs may regulate other TFs, making up a transcription factor network that is active within a cell type. **(b)** The putative regulatory networks for stem cells (left) and mesophyll (right) are shown. Each TF circle has regulatory inputs (stem cell or mesophyll predicted TF binding site within its proximal regulatory regions) and regulatory outputs (that TF's predicted binding site in the other TF gene's proximal regulatory

regions). For example, IDD7 has four regulatory outputs to IDD2, GATA1, GATA15, and itself, and one regulatory input to itself.

LITERATURE CITED

Alvarez, J.P., Furumizu, C., Efroni, I., Eshed, Y., and Bowman, J.L. (2016). Active suppression of a leaf meristem orchestrates determinate leaf growth. eLife **5**.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics **31,** 166-169.

Bajic, M., Maher, K.A., and Deal, R.B. (2018). Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. Methods in Molecular Biology **1675,** 183-201.

Balcerowicz, D., Schoenaers, S., and Vissenberg, K. (2015). Cell Fate Determination and the Switch from Diffuse Growth to Planar Polarity in Arabidopsis Root Epidermal Cells. Frontiers in Plant Science **6,** 1163.

Barton, M.K. (2010). Twenty years on: the inner workings of the shoot apical meristem, a developmental dynamo. Developmental Biology **341,** 95-113.

Behringer, C., and Schwechheimer, C. (2015). B-GATA transcription factors - insights into their structure, regulation, and role in plant development. Frontiers in Plant Science **6,** 90.

Besnard, F., Vernoux, T., and Hamant, O. (2011). Organogenesis from stem cells in planta: multiple feedback loops integrating molecular and mechanical signals. Cellular and Molecular Life Sciences **68,** 2885-2906.

Birkenbihl, R.P., Kracher, B., and Somssich, I.E. (2017). Induced Genome-Wide Binding of Three Arabidopsis WRKY Transcription Factors during Early MAMP-Triggered Immunity. The Plant Cell **29,** 20-38.

Bolt, S., Zuther, E., Zintl, S., Hincha, D.K., and Schmulling, T. (2017). ERF105 is a transcription factor gene of Arabidopsis thaliana required for freezing tolerance and cold acclimation. Plant, Cell & Environment **40,** 108-120.

Brand, U., Grunewald, M., Hobe, M., and Simon, R. (2002). Regulation of CLV3 expression by two homeobox genes in Arabidopsis. Plant Physiology **129,** 565-575.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology **109,** 21.29.21-29.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods **10,** 1213-1218.

Burton, A., and Torres-Padilla, M.E. (2014). Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. Nature reviews. Molecular Cell Biology **15,** 723-734.

Chen, D., Xu, G., Tang, W., Jing, Y., Ji, Q., Fei, Z., and Lin, R. (2013). Antagonistic basic helix-loop-helix/bZIP transcription factors form transcriptional modules that integrate light and reactive oxygen species signaling in Arabidopsis. The Plant Cell **25,** 1657-1673.

Chen, H., Lai, Z., Shi, J., Xiao, Y., Chen, Z., and Xu, X. (2010). Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress. BMC Plant Biology **10,** 281.

Chen, J., Nolan, T.M., Ye, H., Zhang, M., Tong, H., Xin, P., Chu, J., Chu, C., Li, Z., and Yin, Y. (2017). Arabidopsis WRKY46, WRKY54, and WRKY70 Transcription Factors Are Involved in Brassinosteroid-Regulated Plant Growth and Drought Responses. The Plant Cell **29,** 1425-1439.

Clough, S.J., and Bent, A.F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. The Plant Journal **16,** 735-743.

Cui, D., Zhao, J., Jing, Y., Fan, M., Liu, J., Wang, Z., Xin, W., and Hu, Y. (2013). The arabidopsis IDD14, IDD15, and IDD16 cooperatively regulate lateral organ morphogenesis and gravitropism by promoting auxin biosynthesis and transport. PLoS Genetics **9,** e1003759.

Danisman, S. (2016). TCP Transcription Factors at the Interface between Environmental Challenges and the Plant's Growth Responses. Frontiers in Plant Science **7,** 1930.

De Rybel, B., Vassileva, V., Parizot, B., Demeulenaere, M., Grunewald, W., Audenaert, D., Van Campenhout, J., Overvoorde, P., Jansen, L., Vanneste, S., Moller, B., Wilson, M., Holman,

T., Van Isterdael, G., Brunoud, G., Vuylsteke, M., Vernoux, T., De Veylder, L., Inze, D., Weijers, D., Bennett, M.J., and Beeckman, T. (2010). A novel aux/IAA28 signaling cascade activates GATA23-dependent specification of lateral root founder cell identity. Current Biology **20,** 1697-1706.

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

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

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research **38,** W64-70.

Endo, M., Shimizu, H., Nohales, M.A., Araki, T., and Kay, S.A. (2014). Tissue-specific clocks in Arabidopsis show asymmetric coupling. Nature **515,** 419-422.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics **27,** 1017-1018.

He, H.H., Meyer, C.A., Hu, S.S., Chen, M.W., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H., Long, H., Liu, X.S., and Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nature Methods **11,** 73-78.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell **38,** 576-589.

Heyndrickx, K.S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. (2014). A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. The Plant Cell **26,** 3894-3910.

Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., and Li, L. (2013). PAVIS: a tool for Peak Annotation and Visualization. Bioinformatics **29,** 3097-3099.

John, S., Sabo, P.J., Canfield, T.K., Lee, K., Vong, S., Weaver, M., Wang, H., Vierstra, J., Reynolds, A.P., Thurman, R.E., and Stamatoyannopoulos, J.A. (2013). Genome-scale mapping of DNase I hypersensitivity. Current Protocols in Molecular Biology **Chapter 27,** Unit 21.27.

Koyama, T., Furutani, M., Tasaka, M., and Ohme-Takagi, M. (2007). TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in Arabidopsis. The Plant Cell **19,** 473-484.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods **9,** 357-359.

Leivar, P., and Quail, P.H. (2011). PIFs: pivotal components in a cellular signaling hub. Trends in Plant Science **16,** 19-28.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25,** 2078-2079.

Li, S. (2015). The Arabidopsis thaliana TCP transcription factors: A broadening horizon beyond development. Plant Signaling & Behavior **10,** e1044192.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15,** 550.

Lu, Z., Hofmeister, B.T., Vollmers, C., DuBois, R.M., and Schmitz, R.J. (2016). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Research **45,** e41

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics **27,** 1696-1697.

Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D., Zumstein, K., Woodhouse, M., Bubb, K.L., Dorrity, M.W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S.M., and Deal, R. (2017). Profiling of accessible chromatin regions across multiple plant species and cell

types reveals common gene regulatory principles and new control modules. The Plant Cell **30,** 15-36.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D.M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2015). RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Research **43,** W50-56.

Moreno-Risueno, M.A., Sozzani, R., Yardimci, G.G., Petricka, J.J., Vernoux, T., Blilou, I., Alonso, J., Winter, C.M., Ohler, U., Scheres, B., and Benfey, P.N. (2015). Transcriptional control of tissue formation throughout root development. Science **350,** 426-430.

Murashige, T., and Skoog, F. (1962). A Revised Medium for Rapid Growth and Bioassays with Tobacco Tisue Cultures. Physiol. Plantarum **15,** 473-497.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. Cell **150,** 1274-1286.

O'Malley, R.C., Huang, S.S., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165,** 1280-1292.

Pfeiffer, A., Shi, H., Tepperman, J.M., Zhang, Y., and Quail, P.H. (2014). Combinatorial complexity in a transcriptionally centered signaling hub in Arabidopsis. Molecular Plant **7,** 1598-1618.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research **44,** W160-165.

Ranftl, Q.L., Bastakis, E., Klermund, C., and Schwechheimer, C. (2016). LLM-Domain Containing B-GATA Factors Control Different Aspects of Cytokinin-Regulated Development in Arabidopsis thaliana. Plant Physiology **170,** 2295-2311.

Richter, R., Behringer, C., Zourelidou, M., and Schwechheimer, C. (2013). Convergence of auxin and gibberellin signaling on the regulation of the GATA transcription factors GNC and GNL in Arabidopsis thaliana. Proceedings of the National Academy of Sciences **110,** 13192-13197.

Saini, S., Sharma, I., and Pati, P.K. (2015). Versatile roles of brassinosteroid in plants in the context of its homoeostasis, signaling and crosstalks. Frontiers in Plant Science **6,** 950.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics **11,** 415.

Sawchuk, M.G., Donner, T.J., Head, P., and Scarpella, E. (2008). Unique and overlapping expression patterns among members of photosynthesis-associated nuclear gene families in Arabidopsis. Plant Physiology **148,** 1908-1924.

Scarpeci, T.E., Frea, V.S., Zanor, M.I., and Valle, E.M. (2017). Overexpression of AtERF019 delays plant growth and senescence, and improves drought tolerance in Arabidopsis. Journal of Experimental Botany **68,** 673-685.

Schiefelbein, J., Huang, L., and Zheng, X. (2014). Regulation of epidermal cell fate in Arabidopsis roots: the importance of multiple feedback loops. Frontiers in Plant Science **5,** 47.

Schoof, H., Lenhard, M., Haecker, A., Mayer, K.F., Jurgens, G., and Laux, T. (2000). The stem cell population of Arabidopsis shoot meristems in maintained by a regulatory loop between the CLAVATA and WUSCHEL genes. Cell **100,** 635-644.

Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., and Furey, T.S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Research **23,** 777-788.

Singh, A.P., and Savaldi-Goldstein, S. (2015). Growth control: brassinosteroid activity gets context. Journal of Experimental Botany **66,** 1123-1132.

Son, G.H., Wan, J., Kim, H.J., Nguyen, X.C., Chung, W.S., Hong, J.C., and Stacey, G. (2012). Ethylene-responsive element-binding factor 5, ERF5, is involved in chitin-induced innate immunity response. Molecular Plant-Microbe Interactions **25,** 48-60.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nature reviews. Genetics **13,** 613-626.

Stephenson, P.G., Fankhauser, C., and Terry, M.J. (2009). PIF3 is a repressor of chloroplast development. Proceedings of the National Academy of Sciences **106,** 7654-7659.

Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P., Stergachis, A.B., Vernot, B., Johnson, A.K., Haugen, E., Sullivan, S.T., Thompson, A., Neri, F.V., 3rd, Weaver, M., Diegel, M., Mnaimneh, S., Yang, A., Hughes, T.R., Nemhauser, J.L., Queitsch, C., and Stamatoyannopoulos, J.A. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. Cell Reports **8,** 2015-2030.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., and von Mering, C. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research **45,** D362-d368.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics **14,** 178-192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y.,

Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., and Stamatoyannopoulos, J.A. (2012). The accessible chromatin landscape of the human genome. Nature **489,** 75-82.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Research.

Toledo-Ortiz, G., Johansson, H., Lee, K.P., Bou-Torrent, J., Stewart, K., Steel, G., Rodriguez-Concepcion, M., and Halliday, K.J. (2014). The HY5-PIF regulatory module coordinates light and temperature control of photosynthetic gene transcription. PLoS Genetics **10,** e1004416.

Weirauch, M.T., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.C., Galli, M., Lewsey, M., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F.Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., and Hughes, T.R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158,** 1431-1443.

Yadav, R.K., Girke, T., Pasala, S., Xie, M., and Reddy, G.V. (2009). Gene expression map of the Arabidopsis shoot apical meristem stem cell niche. Proceedings of the National Academy **106,** 4941-4946.

Yadav, R.K., Tavakkoli, M., Xie, M., Girke, T., and Reddy, G.V. (2014). A high-resolution gene expression map of the Arabidopsis shoot meristem stem cell niche. Development **141,** 2735-2744.

Yang, Z., Tian, L., Latoszek-Green, M., Brown, D., and Wu, K. (2005). Arabidopsis ERF4 is a transcriptional repressor capable of modulating ethylene and abscisic acid responses. Plant Molecular Biology **58,** 585-596.

Yoshida, H., Hirano, K., Sato, T., Mitsuda, N., Nomoto, M., Maeo, K., Koketsu, E., Mitani, R., Kawamura, M., Ishiguro, S., Tada, Y., Ohme-Takagi, M., Matsuoka, M., and Ueguchi-Tanaka, M. (2014). DELLA protein functions as a transcriptional activator through the DNA

binding of the indeterminate domain family proteins. Proceedings of the National Academy of Sciences **111,** 7861-7866.

You, Y., Sawikowska, A., Neumann, M., Pose, D., Capovilla, G., Langenecker, T., Neher, R.A., Krajewski, P., and Schmid, M. (2017). Temporal dynamics of gene expression and histone marks at the Arabidopsis shoot meristem during flowering. Nature Communications **8,** 15120.

Yu, X., Li, L., Zola, J., Aluru, M., Ye, H., Foudree, A., Guo, H., Anderson, S., Aluru, S., Liu, P., Rodermel, S., and Yin, Y. (2011). A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana. The Plant Journal **65,** 634-646.

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 Genetics **9,** e1003244.

Zhong, J., Luo, K., Winter, P.S., Crawford, G.E., Iversen, E.S., and Hartemink, A.J. (2016). Mapping nucleosome positions using DNase-seq. Genome Research **26,** 351-364.

Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature **456,** 125-129.

CHAPTER 5: EVOLUTIONARY FLEXIBILITY IN FLOODING RESPONSE CIRCUITRY IN ANGIOSPERMS

Mauricio A. Reynoso*, Kaisa Kajala*, Marko Bajic*, Donnelly A. West*, Germain Pauluzzi*, Andrew Yao, Kathryn Hatch, Kristina Zumstein, Margaret Woodhouse, Joel Rodriguez-Medina, Neelima Sinha, Siobhan M. Brady, Roger B. Deal, Julia Bailey-Serres

This work is in press at *Science* (2019) 6459:1291-1295. doi:10.1126/science.aax8862.

Supplemental tables can be found in the online publication.

*These authors contributed equally to this work.

SHORT TITLE

Differential wiring under submergence

ONE SENTENCE SUMMARY

Conserved submergence-activated gene families display flexibility in regulatory circuitry.

ABSTRACT

Flooding due to extreme weather threatens crops and ecosystems. To understand variation in gene regulatory networks activated by submergence, we conducted a high-resolution analysis of chromatin accessibility and gene expression at three scales of transcript control in four angiosperms, ranging from a dryland-adapted wild species to a wetland crop. The data define a cohort of conserved submergence-activated genes with signatures of overlapping *cis*-regulation by four transcription factor families. Syntenic genes are more highly expressed than non-syntenic genes, yet the latter can possess the *cis*-motifs and chromatin accessibility associated with submergence upregulation. While the flexible circuitry spans the eudicot-monocot divide, the frequency of specific *cis*-motifs, extent of chromatin accessibility,

and the degree of submergence-activation is more prevalent in the wetland crop and may have adaptive significance.

MAIN TEXT

Climate change has increased the frequency and intensity of floods that impact agricultural productivity. Of major crops, only rice (*Oryza sativa*) is resilient to waterlogging of roots and submergence of aerial tissue, due to adaptation to a semi-aquatic habitat. Other angiosperms experience intermittent flooding and are not adapted to these conditions. Submergence triggers signaling in plant cells as a consequence of entrapment of the gaseous hormone ethylene and depletion of available O_2, leading to inefficient anaerobic metabolism and energy starvation (*1*) To understand the variation in response to submergence, we studied rice as a representative monocot and flood resilient species, the legume *Medicago truncatula*, and two *Solanum* species, domesticated tomato (*S. lycopersicum* cv. M82) and its dryland-adapted wild relative *S. pennellii* (Figure 5.1A). Roots are the first responders to flooding, and we thus monitored the early response of seedling apical root tips to complete seedling submergence. By monitoring the sentinel response gene family *ALCOHOL DEHYDROGENASE* (*ADH*), required for anaerobic production of ATP (*1*) (Figure 5.1B), we identified 2 hours, the mid-point of maximal upregulation, as a physiologically relevant time to compare initiation of the submergence response across species.

To conserve energy under hypoxia, stress-induced mRNAs are preferentially translated over transcripts associated with development in the model *Arabidopsis thaliana* (*2–4*). We therefore considered both transcriptional and post-transcriptional regulation under submergence across the species surveyed. To do so, we deployed Isolation of Nuclei TAgged in specific Cell Types (INTACT) (*5*) and Translating Ribosome Affinity Purification (TRAP) (*6*), using constitutive promoters. INTACT was used to profile chromatin accessibility by ATAC-seq (Assay for Transposase-Accessible Chromatin) (*7*), and measure the abundance of nuclear RNA (nRNA). TRAP was used to monitor ribosome-associated polyadenylated mRNA (TRAP RNA) and evaluate the position of individual ribosomes along transcripts (Ribo-seq) (*8*) (Figure 5.1C, S5.1-5.2). We also profiled total polyadenylated mRNA (polyA RNA).

Multidimensional scaling analysis confirmed the reproducibility and distinctness of each of these RNA sub-populations and their changes following submergence (Figure S5.3, Table S5.1-5.2).

Flood-adapted rice displayed the greatest plasticity in terms of the number of differentially up- and downregulated transcripts (Figure 5.1D, S5.4, Table S5.3). Cultured hairy roots (*Sl*-HR) were used as a contrast to intact roots of tomato (*Sl*) plants, and were more responsive. The clustering of modulated RNAs resolved variation in regulation in all four species (Figure 5.1D, S5.5-5.9). Rice gene regulation was coordinated across scales (except in clusters 7 and 8 where transcripts were enriched or depleted in the nucleus). In *M. truncatula* and tomato, regulation of gene activity was more evident in the ribosome-associated RNAs, whereas in the dryland-adapted *S. pennellii* regulation was evident as nRNA enrichment or depletion.

Selection likely acts on species-specific traits and adaptation to specific environments that are largely regulated by a common set of gene families. The root meristem is frequently oxygen-deprived due to high metabolic activity and periodic soil inundation; therefore, its capacity to transiently upregulate anaerobic metabolism might be expected in all species. Yet, rice may have evolved a higher proportion of gene family members that are regulated by submergence than flooding-sensitive species. We leveraged gene families (*9*) to investigate conservation in submergence-responsive genes of the four species, focusing on the shared families (6685) plus those conserved between the two *Solanums* (3301) (Figure 5.1E, Table S5.4). Next, we tabulated the submergence-responsive gene family members of each species, identifying families with at least one member differentially controlled in any of the RNA populations evaluated (Figure 5.1F, S5.10, Table S5.5). This uncovered a set of 68 submergence-upregulated families (SURFs: 249 genes in *Os*, 121 in *Mt*, 137 in *Sl*, 181 in *Sl*-HR and 92 in *Sp*). The 68 SURFs include 17 of the 49 ubiquitously hypoxia-responsive genes of Arabidopsis seedlings (*6*), demonstrating evolutionarily conservation of gene families activated by submergence and hypoxia (Table S5.5).

The 68 SURFs include one to 13 upregulated genes per family, leading us to investigate whether similar proportions of these families are elevated in each species (Figure S5.11, Table S5.6). Consistent with overall numbers, rice had the highest and *S. pennellii* the lowest proportion of upregulated genes per

family. The restrained response of wild tomato was evident from the 412 *Solanum*-specific gene families that were up-regulated in tomato but not in *S. pennellii*. This motivated exploration of the aerial tissue (shoot apex) response in the *Solanums* and found more gene families and family members upregulated in shoots of wild than domesticated tomato (Figure S5.12, Table S5.7). The shoot response of *S. pennellii* showed greater overlap with Arabidopsis shoot-specific hypoxia-responsive genes (*10*). Distinctions between the two *Solanums* included genes involved in cell elongation and auxin signaling, which predominated in *S. pennellii*.

We reasoned that dynamics in chromatin accessibility and transcriptional activation may be coordinated and conserved for SURF members across species. ATAC-seq exposed open chromatin regions of rice and *M. truncatula* primarily within 1 kb upstream of the transcription start site (TSS) and downstream of the polyadenylation (pA) site of genes (Figure 5.2A, Table S5.8). By contrast, *Solanum* roots showed a majority of intergenic ATAC-seq reads (Figure S5.13). The rice and *M. truncatula,* transposase hypersensitive sites (THS) (*11*) uncovered a preference for opening of chromatin in response to submergence (Figure 5.2B, S5.13), with increases in 3,497 and 7,501 THSs, respectively. Highly submergence-upregulated genes had elevated accessibility 5’ of their TSS and 3’ of their pA sites (Figure 5.2C, S5.5-5.6, 5.14), demonstrating nucleosome depletion accompanies activation of transcript production under submergence. Downregulated genes had lower chromatin accessibility overall, particularly in rice (Figure 5.2C, S5.5-5.6, 5.14).

We leveraged the ATAC-seq data to explore conservation in gene regulatory circuitry. A pipeline was developed to identify transcription factor (TF) binding site motif enrichment within promoters and their THS regions of the upregulated SURFs (Figure 5.2D). Four significantly enriched TF motifs were identified. These included the Hypoxia Responsive Promoter Element (HRPE), transactivated by low oxygen-stabilized ethylene response group VII (ERFVII) TFs that upregulate genes key to anaerobic metabolism and flooding survival in Arabidopsis (*12–14*), a basic Helix-Loop-Helix (bHLH), a MYB, and a WRKY-type motif (Figure 5.2D, S5.15-5.16A, Table S5.9). At least one of the four motifs was present in over 84% of the upregulated SURF genes of rice and *M. truncatula* and over 68% of those of

the *Solanums*. HRPE and bHLH motifs predominated near the TSS in all species, with the MYB near the TSS in tomato and WRKY motifs more evenly distributed across the upstream region (Figure S5.16B). Differential wiring of upregulated SURFs was evident from the HRPE enrichment in rice (55%) versus the MYB or bHLH motif enrichment in the eudicots (Figure S5.16A, Table S5.9).

Accessibility of chromatin in response to abiotic stress can be rapid and transient (*15*, *16*). We hypothesized that concordance between a TF binding site and a THS would be representative of a more static regulatory architecture while discordance could reflect the transient propagation of a stress signal. Chromatin accessibility increased during submergence around HRPE and bHLH sites in rice and *M. truncatula* (Figure S5.17). A more modest increase was observed for MYB and WRKY sites, potentially representing more rapid and/or transient regulatory interactions (Figure S5.17). The co-occurrence of an HRPE and THS corresponded with more pronounced polyA RNA upregulation, with a similar trend observed for bHLH sites in rice and *M. truncatula* (Figure 5.2E, S5.18, Table S5.10). In *M. truncatula,* the presence of a THS alone in the proximal promoter was associated with greater elevation of polyA RNA, and co-occurrence of a MYB and THS corresponded with higher upregulation than the presence of the motif alone (Figure 5.2F, S5.18). Repetitive motifs of the same type within accessible regions coincided with greater up-regulation than with repetitive motifs outside THSs. The presence of multiple HRPE or WRKY motifs corresponded with higher upregulation in tomato, whereas only an HRPE or multiple bHLH motifs corresponded with upregulation in *S.pennellii*. These results establish a link between the four conserved motifs, chromatin accessibility and transcriptional activation under submergence.

The discovery of the SURFs and four conserved *cis*-regulatory TF binding motifs in submergence-accessible chromatin regions motivated us to evaluate if the conservation prevails in genes maintained at syntenic chromosomal regions (syntelogs). To do so, the gene activity data were re-clustered for the differentially regulated syntelogs across the four species (711) that included 22 of the 68 SURFs (Figure 5.3A, S5.19, Table S5.11). Syntelog clusters 2 and 3 had coordinated upregulation across the scales of gene activity in all species. These comprised seven SURFs with functions in anaerobic

metabolism, nutrient transport, abscisic acid (ABA) perception and survival of extreme stress. The upregulated syntelogs included 32 and 53 SURFs in all three eudicots and the two *Solanums*, respectively (Figure S5.20-5.22, Table S5.11).

Next, we explored conservation on more recent evolutionary timescales by evaluating the activity of genes with syntelogs of related species (Figure 5.3B, S5.23, Table S5.12-5.14). Syntenic genes had higher transcript abundance than non-syntenic genes, as reported previously (*17*). This was evident in all RNA populations under both conditions, with the most pronounced difference between syntenic and non-syntenic genes in the *Solanums*. Rice and *M. truncatula* syntenic gene control regions had slightly higher chromatin accessibility than non-syntenic genes at the global scale (Figure S5.14), consistent with their higher expression. Transcript elevation was similar for syntenic and non-syntenic SURF genes, especially for the *Solanums* (Figure 5.3C, S5.24, Table S5.14), indicating that upregulated non-syntenic genes have maintained or acquired features enabling their stress activation. Consistent with this, most highly expressed syntenic and non-syntenic SURF genes contained at least one of the four TF motifs recognized (80% rice, 80% *M. truncatula*, >70% *Solanums)* (Figure 5.3C, S5.24, Table S5.15). Most TF motifs were coincident with THSs in rice and *M. truncatula*. Although the number of highly expressed but non-syntenic SURF genes was fewer than six in the *Solanums*, all from *S. lycopersicum* contained at least one motif. The four identified TF motifs are therefore a broadly conserved feature of both syntenic and non-syntenic submergence-responsive genes.

To appraise conservation in gene regulation across eudicots and monocots, we built networks that associate TF motif presence with each upregulated SURF gene for each species (Figure 5.4A, S5.25-S5.28, Table S5.16). The individual species networks emphasize the presence of species-specific motif biases. The combinatorial nature of target gene regulation was also evident (overlapping outer circles of network) with over 70% of the genes with more than one of the four motifs. Syntenic upregulated SURF genes across the four species (represented with black borders) expose a single conserved putative regulatory network (Figure 5.4B, S5.29, Table S5.16). This network illustrates conservation of TF motifs of syntelogs of responsive genes, in addition to the HRPE regulated by ERFVIIs in Arabidopsis.

As oxygen levels decline below a threshold, constitutively synthesized ERFVIIs accumulate due to attenuation of their conversion into an N-degron for active turnover (*1*). The unified SURF network uncovered HRPE conservation across eudicots-monocots in promoters of genes essential to anaerobic metabolism and hypoxia survival including *PLANT CYSTEINE OXIDASE* (*PCO*) genes (Figure 5.4C, S5.30, Table S5.17), which catalyze the oxygen-promoted degradation of ERFVIIs to temper the adaptive response (*18*). The upregulated SURF genes included *ERFVII*s in all four species, with at least one with an HRPE motif, suggesting possible autoregulation (Figure S5.31).

The syntelog network also identified conservation of bHLH motif enrichment in genes not well associated with submergence (i.e., *PYRABACTIN RESISTANCE 1* / *PYR1-LIKE* [*PYL*]) (Figure 5.4B, S5.32) and MYB motif enrichment in genes that contribute to hypoxia tolerance (*14*) (Figure S5.33-34). The upregulation of these genes often coincided with a TF motif within a region of submergence-enhanced chromatin accessibility (Figure 5.4D-I, S5.33), supporting functionality of the regulatory sequences. As for the ERFVIIs, the upregulated SURF genes included bHLH, MYB and WRKY family members (Figure S5.31).

Information from single genes is used in breeding or modifying crops for stress tolerance. The use of multi-scale gene regulatory information of gene families across flowering plant clades to infer regulatory networks demonstrates that conservation of flooding resilience mechanisms is complex and involves diverse regulatory mechanisms. Targeted manipulation of the four submergence-activated modules and seven SURF loci discovered here with the greatest interspecies conservation might be used to enhance flooding tolerance of susceptible crops.

ACKNOWLEDGEMENTS

We thank members of our labs for support and discussions, Jérémie Bazin, Dan Koenig and Dan Kliebenstein for guidance and Hokuto Nakayama for illustration. **Funding:** Supported by United States National Science Foundation Plant Genome Research Program (IOS-1238243) to R.B.D., N.R.S., S.M.B.

and J.B.-S., a Finnish Cultural Foundation fellowship to K.K. and an HHMI Faculty Scholar Fellowship to S.M.B. **Author contributions:** M.A.R, K.K., M.B., G.P., D.A.W.,N.S., S.B., R.B.D. and J.B.-S. conceived and designed experiments and analyzed data; M.A.R, K.K., M.B., G.P., D.A.W., A.Y., K.H., K.Z. and M.W. performed experiments. M.A.R, K.K., M.B., N.S., S.B., R.D. and J.B.-S. wrote the manuscript. **Competing Interests:** Authors declare no competing interests**. Data and material availability:** Sequence data deposited in GEO (accession GSE128680). All other data needed to evaluate the conclusions are in Supplementary Materials. **Plasmids and genetic materials:** rice, J.B.-S.; *Medicago*, R.B.D.; tomato, S.B., *S. pennellii*, N.S..

MATERIALS AND METHODS

Plant material and transformation

Rice (*Oryza sativa japonica cv.* Nipponbare), tomato (*Solanum lycopersicum var. M82,* LA3475), *Solanum pennellii* (LA0716), and *Medicago truncatula* (ecotype Jemalong A17) were used. The previously reported stable transgenic rice lines used were *35S:OsNTF2-7* for INTACT (*20*) and *35S:His6-FLAG:OsRPL18-2* for TRAP (*21*). Production of transgenic INTACT lines was described previously for all but *S. pennellii (11, 20, 22)*. For the *Solanum* species and *M. truncatula,* the INTACT construct with the Arabidopsis WPP domain was used (*22*). These carry a constitutively expressed dicot codon-optimized *BirA* driven by the *S. lycopersicum SlACT2* promoter (*SlACT2p:mBirA*) for the *Solanum* species and *Arabidopsis thaliana AtACT2* promoter (*AtACT2p:mBirA*) for *M. truncatula*. The TRAP construct for *Solanum* species was *His_6-FLAG-GFP-AtRPL18* (*22*). For *M. truncatula,* the TRAP construct was identical to that of tomato except the *M. truncatula* 60S ribosomal protein RPL18-3 gene (*Medtr1g083460*) replaced *Arabidopsis* RPL18, and 2) the pB7WG backbone (https://gateway.psb.ugent.be/vector/show/pB7WG/search/index/), conferring phosphinothricin (BASTA) resistance, was used instead of the kanamycin resistance conferring pK7WG backbone.

Stable transgenic *Solanum* species were produced by use of *Agrobacterium tumefaciens* transformation by UC Davis Plant Transformation Facility (*11*). The specific *Solanum* lines used were

tomato *35S:INNTF-1*; *35S:His6-FLAG-GFP-AtRPL18-5*, and *S. pennellii 35S:NTF-3; 35S:His6-FLAG-GFP-AtRPL18B-1*.

Hairy root cultures of *Agrobacterium rhizogenes* transformed *S. lycopersicum* roots were initiated and cultivated as described (*22*).

Hairy root composite *M. truncatula* plants were initiated by injecting the primary roots with *A. rhizogenes* K599. To do so, the day before injection, a small volume of glycerol stock of *A. rhizogenes* K599 carrying either *35S:NTF;AtACT2p:mBirA* or *35S:His6-FLAG-GFP-MtRPL18-3* plasmid was used to inoculate 5 ml of Yeast Extract Beef media (5 g/L tryptone, 1 g/L yeast extract, 5 g/L nutrient broth, 5 g/L sucrose, 0.49 g/L $MgSO_4x7H_2O$, 15 g/L agar, pH 7.2) additionally containing 100 mg/L Spectinomycin and grown overnight at 28°C at 200 rpm. When the OD_{600} was equal to 1.0, cultures were centrifuged at 5,000 rpm for 5 min. The supernatant was poured off and the pellet was resuspended in Injection Media (IM) (1X PBS, 100 uM acetosyringone, 1/10,000 (v/v) Silwet). Seedlings germinating for 2.5-days were injected by placing them in 5 mL of IM-resuspended *A. rhizogenes* poured out on a Petri dish and stabbing the root a few times with an 18G1 needle. Additionally, 1-2 mm of the primary root tip was cut off. Injected seedlings were moved to slanted Fahräeus Media (FM) plates (0.5 mM $MgSO_4$, 0.7 mM KH_2PO_4, 0.8 mM Na_2HPO_4, 50 nM FeEDTA, 0.5 mM NH_4NO_3, 1mM $CaCl_2$, 0.1 mg of $MnSO_4$, $CuSO_4$, Zn SO_4, H_3BO_3, and Na_2MoO_4, 8 g/L Phytoblend agar, pH 6.5) with no selection and were grown horizontally for 3 days, and then were moved to FM plates with 5 mg/L phosphinothricin and were grown vertically for 3 weeks before transfer to 1X MS media without vitamins (1% w/v agar, 1% w/v sucrose).

Growth conditions and submergence treatment

For rice, seeds were dehulled and surface sterilized in 50% (v/v) bleach solution for 30 min, rinsed ten times with sterile distilled water and grown on plates (100 cm^2) containing 0.5x Murashige and Skoog medium (MS), 1% (w/v) agar 1% (w/v) sucrose for 7 days (16h day / 8h night; at 28°C/25°C day/night; 110 $\mu Em^{-2}s^{-1}$). For tomato, seeds were surface sterilized in 50% (v/v) bleach solution for 5 min (*S. pennellii*) or 20 min (*S. lycopersicum*) and then rinsed three times with sterile distilled water. Growth was on vertical plates

(10 cm x 10 cm) containing full-strength MS without vitamins, with 1% (w/v) agar (w/v) and 1% (w/v) sucrose. *S. lycopersicum* hairy root cultures transformed with *A. rhizogenes* were subcloned using a 2-cm hairy root segment, grown on horizontal plates (10 cm x 10 cm) containing full-strength MS with vitamins, with 1% (w/v) agar (w/v) and 3% (w/v) sucrose, 200 mg/L kanamycin and 200 mg/L cefotaxime. All tomato root cultures or germinating seeds were grown for 7 days in a growth chamber (at 25°C, 16h day/8h night; 60-65 $\mu Em^{-2}s^{-1}$).

M. truncatula seeds were surface-sterilized by incubating in concentrated sulfuric acid for 8 minutes with gentle stirring, washing 3 times with 4°C sterile, distilled water, then 4-8 minutes in 3% (w/v) hypochlorite (diluted bleach), washing 4 times with sterile, distilled water, and finally placing the seeds onto moist filter paper. Seeds were germinated without stratification on moist filter papers in inverted Petri dishes wrapped with surgical tape wrapping. These were kept in the growth room in the dark at 20°C for 2 days. The seedlings were then injected with *A. rhizogenes*, and composite plants with transformed hairy roots were obtained three weeks later. After the composite plant transformation protocol, the plants were grown vertically for one week (at 20°C, 16h day/8h night; 150 $\mu Em^{-2}s^{-1}$) before being used for the submergence experiment. The day before the experiment was performed, root tips from one plate were collected and visualized on a fluorescence stereomicroscope to check for GFP expression. Transformation efficiency was calculated as the percentage of root tips with ubiquitous GFP expression.

Four independent biological replicates were grown for each species. For whole plant submergence, plates were placed horizontally, opened and the seedlings covered with 5 cm of autoclaved distilled water at ZT 4h. Root tips (apical 1 cm including the meristem, elongation zone and early differentiation zone; all four species) and the shoot apical meristem region (*Solanum* species) were harvested at ZT 6h (2h before relative noon). Control plates were not opened, but positioned horizontally for the duration of the treatment and harvested at ZT 6h. Oxygen partial pressure was measured with the NeoFox Sport O_2 sensor and probe (Ocean Optics). The dissolved oxygen content in the water covering the plants remained above 18% (v/v) for the 2 h duration of the stress treatment.

Quantitative real-time reverse transcriptase PCR (qRT-PCR)

Three independent biological replicates of submergence and control time courses were conducted. Root tips (apical 1 cm) were harvested every two hours after submergence for qRT-PCR. For the *Solanum* species, RNA was extracted by polyA mRNA extraction (*23*), cDNA was synthesised by Superscript III (Invitrogen) and qRT-PCR was performed with SensiFast SYBR Hi-RIX kit (Bioline) with CFX384 Touch™ Real-Time PCR Detection System (Bio-Rad), all as per the manufacturer's instructions. For *M. truncatula*, RNA was extracted using the RNeasy Plant Mini Kit (Qiagen). Genomic DNA was removed using the TURBO DNA-free Kit (Ambion) and first-strand synthesis was performed using SuperScript III on 150 ng of RNA. qRT-PCR was performed with Power SYBR Green Master Mix (Applied Biosystems) on the StepOnePlus Real-Time PCR System (Applied Biosystems). For rice, RNA was extracted using Direct-zol RNA Miniprep (Zymo Research). qRT-PCR was performed with 500 ng of RNA pre-treated with DNAse I (Thermo Scientific), reverse transcribed using Maxima reverse transcriptase (Thermo Fisher) in a total volume of 20 µl, according to the manufacturer's instructions. cDNA was diluted with 180 µl of ddH_20 and real-time (RT) PCR was performed with 5 µl of the diluted cDNA using SsoAdvanced™ Universal SYBR Green Supermix (Bio-Rad) and the CFX96 Touch Real-Time PCR Detection System (Bio-Rad) using defined primers (Table S5.18). *ADH* mRNA abundance was compared to *UBCII* (*O. sativa*), *ACT2* (*S. lycopersicum* and *S. pennelli*) and *RPL2* (*M. truncatula*). The ΔΔCt method (*24*) was used to calculate relative in RNA abundance, using a Student's *t* test to evaluate significance using three technical and three biological replicates.

Nuclei purification by INTACT for ATAC-seq and nRNA-seq

Nuclei were purified from frozen and pulverized tissue as described previously for *A. thaliana* (*25*) with minor modifications (*20*). Tagmentation using Tn5 insertion and ATAC-seq libraries were prepared using 20,000-50,000 nuclei as previously described (*11*)(*26*), with slight modifications. For rice, minor modifications in nuclei purification include: 1) the use of a 30 µm filter to exclude 30 to 70 µm cellular debris from the crude extract and extended centrifugation times (*27*), and 2) using AMPureXP beads instead

of columns to purify amplified libraries. For nRNA, samples were processed as described (*20*, *28*). In brief, nuclei captured by INTACT were processed using the RNeasy Micro kit (Qiagen), treated with Turbo DNase I (ThermoFisher Scientific) and concentrated using Agencourt RNAClean XP beads (Beckman Coulter). To remove contaminating pre-rRNA/rRNA a subtraction step using biotinylated oligos tiling pre-rRNA/rRNA and high temperature double-stranded nuclease was performed (*20*). Samples were re-treated with Turbo DNase I (ThermoFisher Scientific) and cleaned up with Agencourt RNAClean XP beads prior to use in library construction as described below.

Polysomal mRNA purification by TRAP and Total RNA isolation and RNA-seq library construction

TRAP was performed as previously described (*29*, *30*) with the following modifications: α-FLAG conjugated IgG Dynabeads were used for binding; after magnetic collection and washing the polysomes were removed from the magnetic beads by the addition of Lysis and Binding Buffer (LBB) buffer for polyA mRNA isolation using biotinylated oligo-dT primers and streptavidin magnetic beads (NEB) (*23*). Total RNA was extracted from frozen tissue using polysome extraction buffer (*29*) followed by LBB polyA mRNA isolation using biotinylated oligo(dT) and streptavidin magnetic beads (*23*). Random primer-primed RNA-seq library construction for nRNA (pre-rRNA and rRNA digested), polyadenylated total RNA and polyadenylated TRAP RNA was performed as described (*23*) in at least four biological replicates for each condition and species.

Ribo-seq library construction

Ribo-seq libraries were generated as described by (*31*) but with ribosome isolation by TRAP as described by (*32*) starting with pulverized frozen root tip tissue (~1,000 1 cm root tip) thawed in 5 mL of Polysome Extraction Buffer and using α-FLAG conjugated IgG Dynabeads for binding instead of anti-FLAG M2 magnetic beads. Manipulations were as previously described by (*32*) through to the generation of ribosome footprint fragments (RFs) and on-magnetic bead digestion of 1 mL of resuspended beads with 2,000 units of RNase I (Ambion; ca. 15 U/µg RNA) by incubation for 180 min at 23-25 °C. RFs of 26-34 nt were gel

purified, dephosphorylated using T4 polynucleotide kinase, ligated 500 ng preadenylylated miRNA cloning linker (IDT, miRNA cloning linker #1). The ligated-RFs were excised, recovered and resuspended in 10.0 µl of 10 mM Tris (pH 8). After this step, rRNA removal of RFs was done by use of Ribo-Zero rRNA Removal Kit (Plant; Illumina) probe solution. Library construction continued as described by (*32*) and the resultant 130 nt RF cDNAs were circularized and contaminating rRNA was subtracted by a second hybridization with custom-designed biotinylated oligos corresponding to pre-rRNA and rRNA as described (*32*). rRNA-subtracted circularized fragments were used for 12 cycles of 10s at 98°C, 10s at 60°C, and 5s at 72 °C PCR amplification including library and indexing primers. *M. truncatula* and *S. lycopersicum* libraries used indexing primers described (*23*), rice libraries used the same primers as described by Ingolia et al., 2012 (*33*). The amplified RF library (~175-180 bp) was excised and recovered from the gel, purified, analysed on Agilent BioAnalyzer DNA 1000 chip, multiplexed and sequenced.

Short read processing, quality assessment, alignment to genomes, and read coverage

For rice, *S. lycopersicum* and *S. pennellii*, nRNA, total poly(A)$^+$ and TRAP libraries and all Ribo-seq libraries, including *M. truncatula*, were sequenced on the Illumina HiSeq 3000 to obtain 50 nt single-end reads at the UC Davis DNA Technologies Core. Raw reads were filtered to remove adapter-only or polyA-pulldown primer sequences. For *M. truncatula*, nRNA, total poly(A)+ and TRAP libraries were sequenced on the Illumina NextSeq 500 at the Georgia Genomics and Bioinformatics Core at UGA to obtain 75 nt single-end reads for nRNA libraries and 36 nt paired-end reads for total poly(A)$^+$ and TRAP libraries.

Rice and *M. truncatula* data analysis steps were performed on the University of California, Riverside Institute for Integrative Genome Biology high performance bioinformatics cluster (http://www.bioinformatics.ucr.edu/), supported by NSF MRI DBI 1429826 and NIH S10-OD016290. R packages from Bioconductor including systemPipeR (*34*) were used. Quality reports of raw reads were generated with the FastQC package (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptors were removed from Ribo-seq reads and RNA raw reads were mapped with the splice junction aware short read alignment suite Bowtie2/Tophat2 to the IGRSP1.0-30 genome for rice

(http://plants.ensembl.org/Oryza_sativa/Info/Index) and Mt4.0 genome for *M. truncatula*, allowing unique alignments with ≤2 nt mismatches. nRNA, polyadenylated total RNA and polyadenylated TRAP RNA was filtered by mapping first to the mitochondrial and chloroplast genomes before mapping to the nuclear genome. Expression analyses were performed by generating read count data for features of exons-by-genes using the summarizeOverlaps function from the GenomicRanges package(*4*).

For the *Solanum* samples, mapping was executed in the CyVerse Discovery Environment. STAR v2.4.0.1 was used to align nRNA-seq reads to organelle sequences to filter out reads that map to organelles (*S. lycopersicum* AFYB01.1 mitochondrial sequence, *S. lycopersicum* NC_007898.3 chloroplast sequence, and *S. pennellii* HG74452 chloroplast sequence, all downloaded from NCBI); alignment parameters were set to include unmapped reads as the output (--outSAMunmapped Within, --outReadsUnmapped Fastx). For all RNA-seq libraries, reads were then mapped to either *S. lycopersicum* ITAG3.10 or *S. pennellii* cDNA (exonic) sequences (downloaded from SolGenomics). For both *Solanum* species the 90 gene models identified to potentially encode for rRNA were masked. STAR GenerateGenomeIndex was set to --limitGenomeGenerateRAM 36000000000 to account for the large number of cDNA inputs) using the STAR v2.4.0.1 aligner (default parameters, SAM output). Estimated count (est_counts) abundances were calculated using eXpress v1.5.1, default parameters. For visualization of read mapping, the reads were also mapped to either the *S. lycopersicum* ITAG3.10 or the *S. pennellii* genome using STAR v2.4.0.1, default parameters.

For all species, reads were converted from SAM to sorted BAM files using Samtools 1.0.9 and the BAM files were converted to bigwig files using BedTools 2.26 genomeCoverageBed followed by UCSC bedGraphtoBigWig 332--0, default parameters.

Ribo-seq reads periodicity analysis was performed using Ribotaper (*35*). Modified GTF files for each species to select protein-coding annotated genes and TRAP RNA expressed genes were used to create annotation files (create_ annotation_files.bash). Merged bam files were used to generate periodicity plots with the script create_metaplots.bash.

RNA-seq differential expression

Statistical analysis of differentially expressed genes was carried out using limma-voom (*36*) Bioconductor package in R. Raw RNA-seq read counts were normalized with voom using the *quantile* method. The functions lmfit, contrasts.fit, and ebayes were used to calculate differential gene expression for different contrasts, including $\log_2$ Fold Change (FC) values and adjusted P values (adj.P.Val). The *fdr* method was used for controlling the false discovery rate (FDR). Multidimensional scaling (MDS) plots were generated using Glimma Bioconductor package in R (*37*) using genes with more than 0.5 count per million in at least 3 replicates. Following normalization, count data were used to calculate reads per kilobase per million reads (rpkM). Enrichment analysis of Gene Ontology (GO) terms was performed with systemPipeR (*34*) using the GO definitions from the BioMart database for rice and *M. truncatula*, and with goseq R package (*38*) using the category list specifically built for *S. lycopersicum* (*39*) for the *Solanum* species.

Transcript read (average) coverage plots were calculated and produced over selected groups of genes with the functions computeMatrix and plotHeatmap from deepTools (http://deeptools.readthedocs.io/en/latest/index.html) using bigwig files. For each group of genes the 5% most highly and 5% lowly expressed genes were removed from the coverage plots, as highly covered regions can bias the mean. Log_2 FC values for genes identified as differentially expressed were clustered using the Partitioning Around Medoids (PAM) method with $k = 12$ to 24 clusters. The method of clustering and the k values were optimum for resolution of correlated genes based on the evaluation of results with multiple k values. Clustering and heatmaps plotting the mean $\log_2$ FC values were created using the following R packages: gplots, cluster, e1071, and RColorBrewer installed from https://cran.r-project.org.

Identification of genes families and syntenic orthologous genes (syntelogs) across species

Predicted angiosperm gene families were extracted from Phytozome v11 including *O. sativa*, *M. truncatula* and *S. lycopersicum* (*11*). These families were established by Phytozome using InParanoid, which uses BLAST alignment between two related protomes to identify orthology groups, defined by the developers as homologs from a speciation event (*40*). Gene families shared between *S. lycopersicum* ITAG3.10 and *S.*

pennellii were independently generated in three ways: 1) using the synteny aligner CoGe SynMap, megablast, with an e-value of 0.001; 2) a two-directional syntenic alignment using CoGe SynFind, default parameters, both with ITAG3.10 as the query and *S. pennellii* as the query; and 3) the best blastn hit, default parameters, between the cDNAs of ITAG3.10 and *S. pennellii*. The gene identifier (ID) obtained was used to associate the gene to a gene family. As the Phytozome list used the ITAG 2.4 annotation, for the genes annotated in ITAG3.10, which do not have an assigned gene ID in ITAG2.4, we performed a blastp (-max_target_seqs 1) search between the cDNAs to find the closest related gene in *A. thaliana.* The resultant gene ID was used to associate the *Solanum* gene to a gene family. To identify conserved responses across species, the overlaps in families containing differentially expressed genes were produced with the function overLapper from systemPipeR. Venn diagrams were plotted with the function vennPlot.

Syntenic orthologs (syntelogs) were identified using a combination of CoGe SynFind (https://genomevolution.org/CoGe/SynFind.pl) with default parameters, and CoGe SynMap (https://genomevolution.org/coge/SynMap.pl) with the QuotaAlign feature selected and a minimum of six aligned pairs required (*41*, *42*). Rice and *Brachypodium distachyon,* or *Zea mays* syntelogs were obtained from ensembl plants (http://plants.ensembl.org/compara_analyses.html). To identify corregulated syntelogs, each gene which had a syntelog in the comparison species and was differentially expressed was clustered by the PAM method as described. To simplify the visualization the median of the cluster was calculated and plotted using the functions geom_line and geom_point from the R package ggplots2.

To evaluate the differences in transcript (nRNA, polyA RNA or TRAP RNA) levels between syntenic and non syntenic genes, the mean of normalized rpkM values were plotted for genes with transcript abundance > 0.5 rpM using the function geom_violin. Significant differences of means were evaluated using the Student's *t* test, comparing a single population of transcripts under one condition (*i.e.*, nRNA in control samples) for syntenic and non-syntenic genes. An F-test was used to evaluate the difference in the variances of the two populations using the R function var.test (Table S5.13).

Enrichment for regulated genes in gene families in each species

For each upregulated gene family in any species and comparison, the number of upregulated and non-regulated genes were quantified to generate a contingency table including the overall number of regulated genes. A linear model was generated to test for the enrichment in a gene family in a species compared to others (mod=glm(formula = family ~ species + degs,data=my.data, family = binomial(logit), weights=Freq)).

Mapping of chromatin accessibility, identification of Transposase Hypersensitive Sites, and evaluation of accessibility changes between conditions

Rice libraries were sequenced on the HiSeq 3000 at the UC Davis DNA Technologies Core to obtain 50 nt single-end reads. *M. truncatula* ATAC-seq libraries were sequenced on the NextSeq 500 at the UGA Georgia Genomics and Bioinformatics Core to obtain 36 nt paired-end reads. *S. lycopersicum*, *S. pennellii*, and a few rice ATAC-seq libraries were sequenced on the NextSeq 500 at the UC Davis DNA Technologies Core to obtain 36 nt paired-end reads. Genomic DNA ATAC-seq libraries from root tips for each species were sequenced on the NextSeq 500 at the UGA Georgia Genomics and Bioinformatics Core to obtain 36 nt paired-end reads. Sequencing reads were mapped using Bowtie2 software (*43*) with default parameters to each species' corresponding genome build; rice was mapped to IGRSP1.0-30, *M. truncatula* was mapped to Mt4.0, *S. lycopersicum* was mapped to ITAG3.10, and *S. pennellii* was mapped to the genome assembly of Bloger et al. (2014) (*44*). Mapped reads were processed as previously described (*45*), which included converting to *.bam* format using Samtools 0.1.19 (*46*), sorting and filtering to retain only reads that had a mapping quality score of 2 or higher, and filtering to retain only the reads that mapped to nuclear chromosomes and scaffolds.

Peak calling was done using the "*Findpeaks*" function of the HOMER package (*47*) with the parameters "*-minDist 150*" "*-region*" and "*-regionRes 1*". Peaks called between replicates were kept if they replicated at least once between replicates given the condition that they overlap by at least 50%. This was done using the Bedtools software (*48*) and the "*intersect*" function. Reproducible peaks that overlapped by 150 bp, half the size of the mean peak sizes called in each species, were merged together using the Bedtools

"*merge*" function to give the final list of reproducible, non-redundant chromatin accessible regions identified in each species. These regions are referred to as Transposase Hypersensitive Sites (THSs).

Read alignments, referred to as counts, present in the coordinates of identified THSs were quantified in each species for control and submergence samples using HTSeq's *htseq-count* script (*49*). At least two replicates of each condition were counted and the counts were statistically evaluated using DESeq2 (*50*). THSs that had a log fold change value of 1 or more, or -1 or less, and a p-value $\leq$ 0.05 were identified as THSs that are either upregulated, or downregulated, during submergence stress. Upregulated THSs refer to chromatin regions where chromatin was more accessible during submergence stress, compared to control conditions.

For visualization of chromatin accessibility data, *.bam* files were converted to bigiwg files using the deeptools 2.0 (*51*) "*bamCoverage*" script, using the bin size of 1 bp and RPKM normalization parameters, and UCSC's "*bigWigMerge*" and "*bedGraphToBigWig*" programs. Replicates for a specific condition were processed such that each replicate had the same number of mapped reads before merging. This was done using the Samtools "*view -c*" command to count the number of aligned reads within the replicate and the Samtools "*view -S*" command to scale down globally the number of reads for that replicate. Heatmap and metaplots of chromatin accessibility data were generated using the deepTools "*computeMatrix*" "*plotHeatmap*" and "*plotProfile*" functions.

Annotation of Transposase Hypersensitive Sites

For each THS, distance to and identity of the nearest gene was assigned using the "*TSS*" function of the PeakAnnotator 1.4 program (*52*). Each THS was assigned to a genomic feature (upstream, exon, intron, downstream, or intergenic) using HOMER's "*annotatePeaks.pl*" program. Genomic features were defined using published annotation files for the genomes used for alignment, as well as HOMER's "*parseGTF.pl*" and "*assignGenomeAnnotation*" programs. The "upstream" regions were defined as the 2,000 bp upstream of the transcription start site, and "downstream" regions were defined as the 1,000 bp downstream of the transcription end site.

Identification of enriched regulatory motifs

Two methods of cis-element enrichment were used. (Method 1) *De novo* discovery: To identify motifs enriched in promoter regions of SURFs, the 2 kb upstream region of the ATG for all the upregulated genes in a family were evaluated for sequence enrichment using MEME (*53*). Motifs significantly enriched (E-value<0.01) were compared to databases of known transcription factors, including DAP-seq (*54*) and CIS-BP (*55*), by using TOMTOM (*56*) on the MEME output. As a control, genes not regulated from each family were processed in the same way to detect putative regulatory elements in control conditions that are non-submergence specific. Detected motifs were screened in all annotated genes using RSAT and the enrichment of SURFs were evaluated using a Fisher's exact test using the fisher.test() function in R. (Method 2) *De novo* discovery in promoter-bound accessible regions of SURFs: To identify motifs enriched in accessible sites found in promoter regions of SURFs, THS sequences found in the 2 kb upstream and +500 bp region of the TSS, highest density of THS localization for rice and *M. truncatula*, for all upregulated genes in a family were evaluated for sequence enrichment using AME (*57*). This enrichment analysis was done using both DAP-seq and CIS-BP databases to match enriched motifs to known transcription factors. Motifs with an E-value<0.01 were considered as significantly enriched motifs found in SURF promoter-bound THSs.

Motif mapping and validation of SURF-regulated enrichment

The FIMO program (*58*) was used to map motifs throughout repeat-masked genome sequences of rice (IGRSP1.0-30), *M. truncatula* (Mt4.0), *S. lycopersicum* (ITAG3.10), and *S. pennellii* genome assembly of Bloger et al. (2014) (*44*). The default parameters for FIMO mapping were adjusted to account for memory considerations and p-value scoring bias for smaller, more precise positional weight matrices. This was done by using the "*--max-stored-scores 100000000*" option and the "*--thresh*" option. The p-value threshold was set manually by choosing a low significance value, such as "*--thresh 0.005*" and then visually examining the results to determine at which cutoff there was more than 1 base pair mismatch between the identified

motif and the sequence reported in the positional weight matrix. The only p-value cutoff changed for the main motifs described in this work was bHLH, which was set to 0.0002 instead of the default value of 0.0001 to account for the smaller size of the motif.

A total of 23 motifs were identified as potential regulators of increased SURF expression. To remove false-positive motifs that are found in high abundance within all gene promoters we used the Fisher's exact test to compare the percentage of SURF promoters of a given species that had a motif versus the percentage of all other genes' promoters that had a motif within that species. Through this validation process, bHLH, HRPE, and MYB were found to be significantly enriched in SURF promoters of all four species. WRKY was significantly enriched in SURF promoters of rice, *M. truncatula*, and *S. lycopersicum*, but not *S. pennellii*.

Network analyses

The Cytoscape software (*59*) was used to build gene regulatory networks between the four TF-binding sequence motifs (bHLH, HRPE, MYB, and WRKY) and the SURF genes of each species. The TF-binding sequence motif must be located within the -2 kb upstream and +500 bp downstream region of the TSS of the SURF gene being regulated in order for the TF-gene connection to be made.

Phylogenetic analyses

Gene family conservation was inferred by using a Maximum Likelihood method based on the JTT matrix-based model in the software MEGA 7 (*60*). The consensus tree is generated from 1000 bootstrap replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position.

Accessible datasets

The data reported are accessible from Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE128680). The scripts used are available at: https://github.com/plant-plasticity/Evolutionary-flexibility-in-flooding-response-2019.

Extended acknowledgements

We thank Ralston Mataki, Sean Cabanlit and Elise Viox for help in tissue collection, Kelly Tran for help in initial experiments, Sonja Winte for advice on protein phylogeny generation, Maureen Gateas Hummel, Travis Lee, Mike Covington and Sharon Gray for discussions and advice on bioinformatics tools.

FIGURES

Fig 5.1

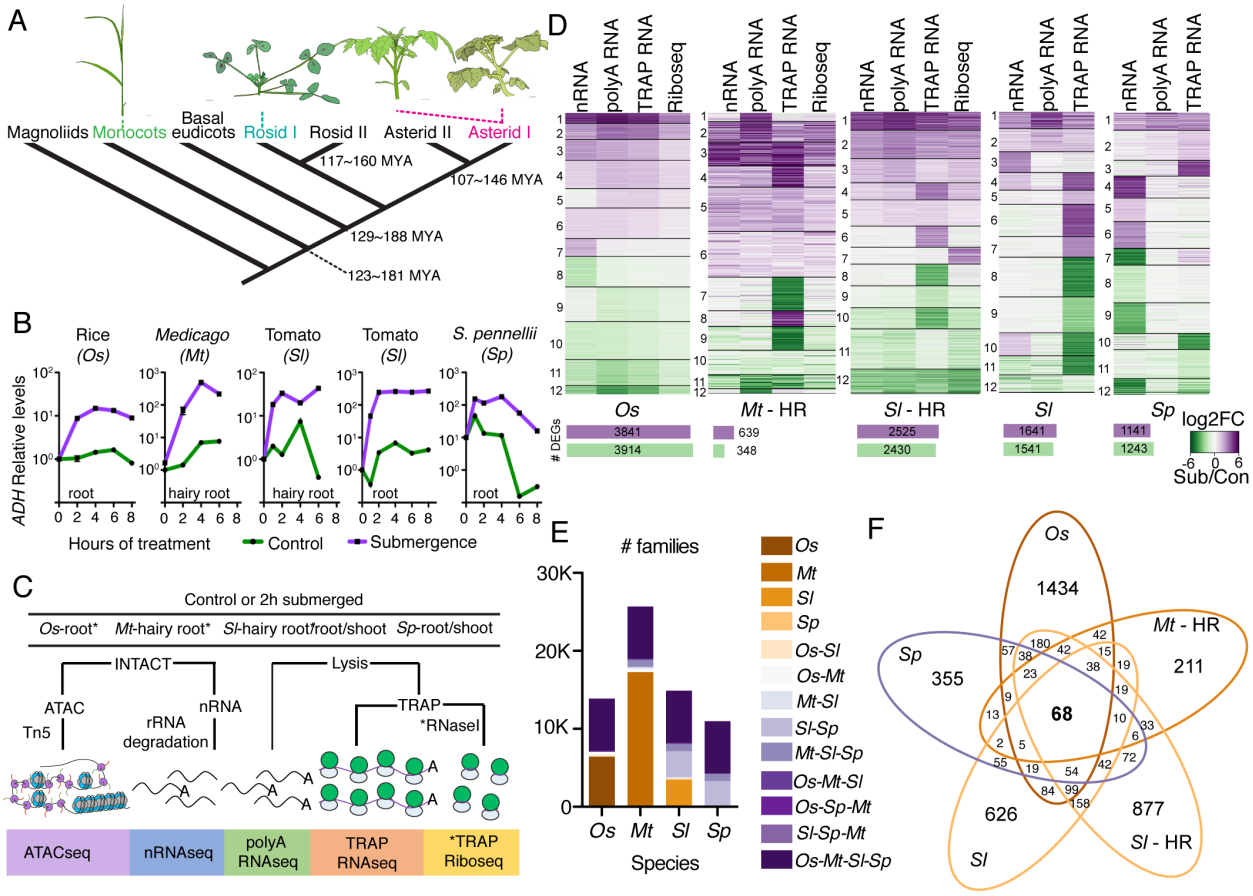


Figure 5.1. Multi-tier evaluation of gene activity in four angiosperms identifies highly conserved submergence-upregulated genes. (**A**) Relatedness of target species (*19*). (**B**) *ALCOHOL DEHYDROGENASE* (*ADH*) transcript levels of submerged seedlings. (**C**) Overview of experimental strategy. (**D**) Cluster analysis heatmap of $\log_2$ fold change (FC; submergence vs. control RNA) of differentially expressed genes (DEGs; $|\log_2 FC|>1$ and padj<0.01). Below, Bars indicate number of up or down DEGs after submergence $|\log_2 FC|>1$ and padj<0.05. (**E**) Gene families per species and their overlap. (**F**) Conserved submergence upregulated gene families (SURFs).

Fig 5.2

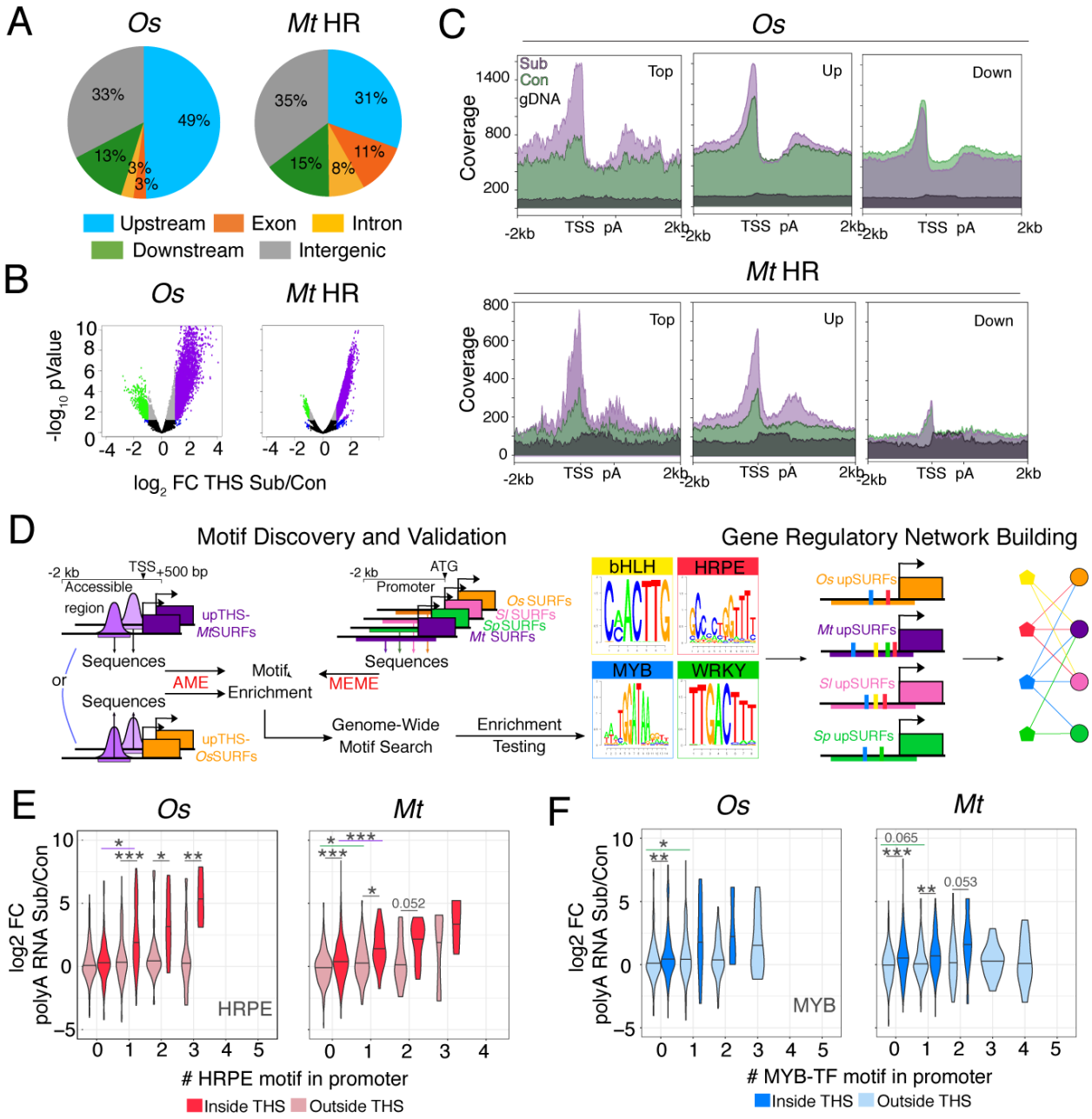


Figure 5.2. Enhanced chromatin accessibility and motif enrichment in responsive genes. (A) Accessible chromatin regions (transposase hypersensitive sites [THS]) measured by ATAC-seq. Categories: 2 kb upstream of the transcription start site (TSS), exons, introns, 1 kb downstream of polyadenylation (pA) site, and intergenic. (**B**) THS change in response to submergence. (**C**) Control (Con) and submergence (Sub) ATAC-seq reads on genes of upregulated (Top; cluster 1; Up) and

downregulated (Down) clusters from Figure 5.1D. gDNA is ATAC-seq on naked DNA. (**D)** Discovery pipeline for enriched transcription factor motifs present in upregulated THSs and SURF promoters, using unsupervised (MEME) and supervised (TOMTOM, AME, FIMO) methods. (**E**) **and** (**F**) Distribution of log_2FC polyA RNA Sub/Con for SURFs arranged by presence and number of HRPE or MYB motif upstream of the ATG, inside or outside THSs. Student's *t* test; * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$; values ≤ 0.1.

Fig 5.3

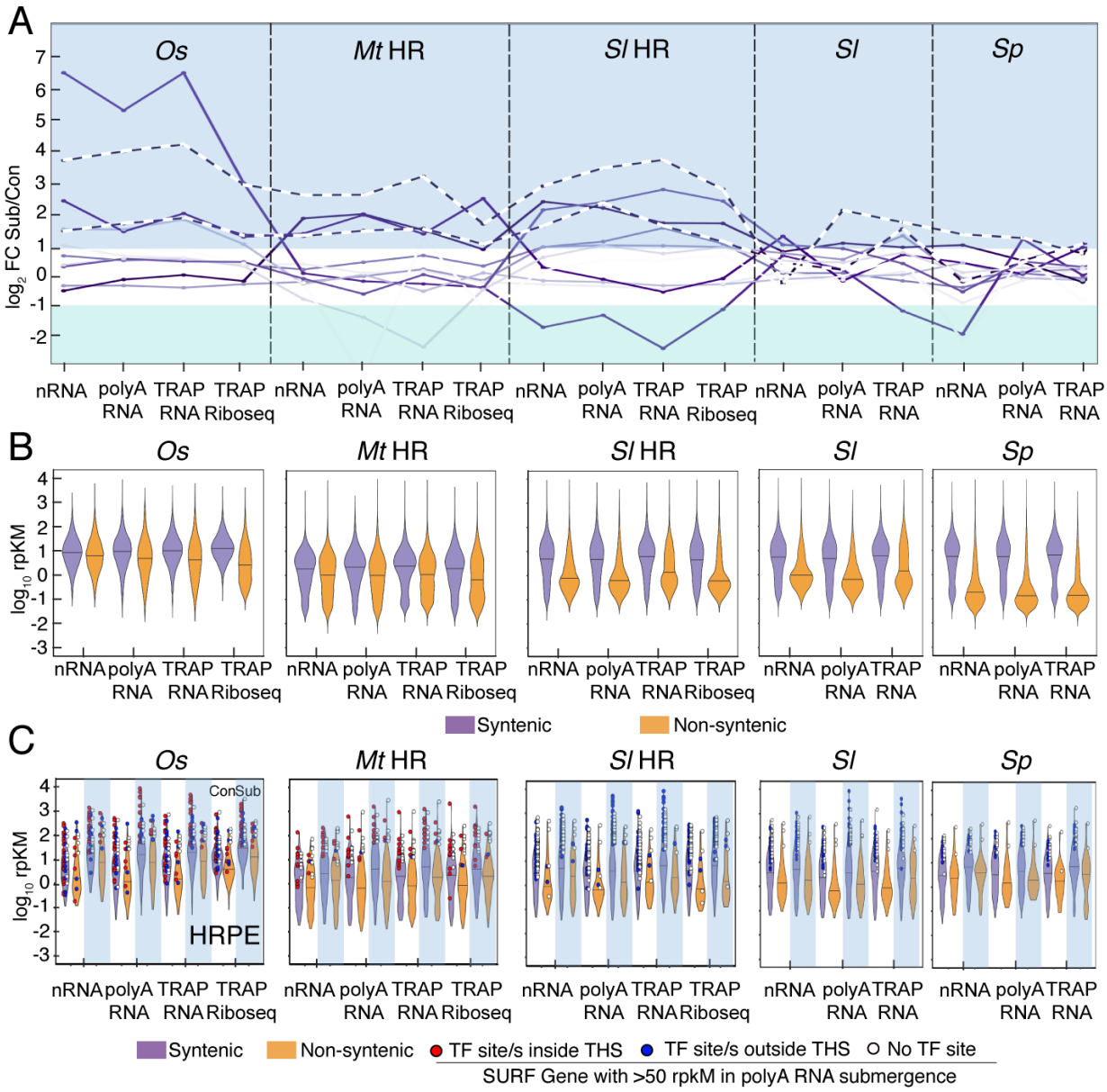


Figure 5.3. Syntenic genes are more highly expressed. (**A**) Median log_2FC of syntenic genes across four species for eight upregulated clusters. Dashed lines indicate two clusters with conserved interspecies up-regulation. (**B**) Plot of log10 rpKM for all detected syntenic and non-syntenic genes under control condition. Rice synteny was evaluated to *Brachypodium distachyon*, *M. truncatula* to *S. lycopersicum,* and between *Solanums*. Variances between syntenic and non-syntenic genes are significant in every RNA population (F-test). (**C**) Control (white columns) and submergence (blue columns) plots for SURF genes.

Highly expressed SURF genes under submergence (>50 rpKM) with a Hypoxia Responsive Promoter Element (HRPE) are depicted as a red or blue dot for those located within or outside a THS, respectively. Central horizontal lines indicate median values.

Fig 5.4

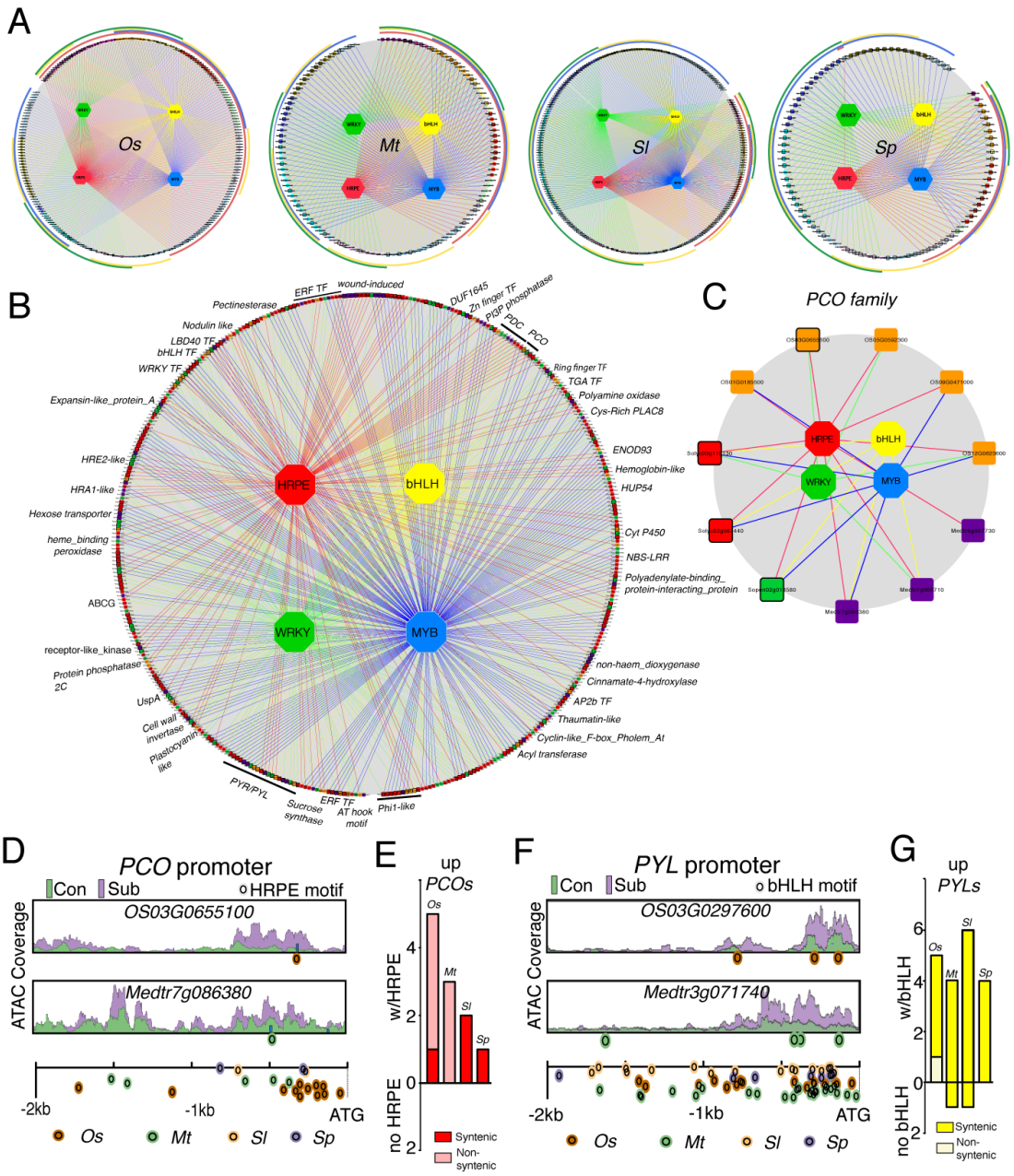


Figure 5.4. Conserved transcription factor motifs in SURFs and accompanying chromatin dynamics. (**A**) Regulatory networks for upregulated SURF genes (expanded in Figures S5.25-S5.28). Hexagons, transcription factors (TFs); rectangles, genes; colored lines (edges), interactions of promoter and TF based on motif presence. Outer circles: genes grouped with shared motifs. Genes with black borders have a syntenic ortholog (rice to *M. truncatula*; *M. truncatula* to *S. lycopersicum;* and between

Solanums). (**B**) Network for syntenic conserved SURF genes across species (expanded in Figure S5.29). Genes of alternating families have alternating grey or black borders. Families represented in three species are labeled. (**C**) Regulatory network of *PLANT CYSTEINE OXIDASE* (*PCO)* up-regulated genes. Syntenic orthologs have black borders. (**D**) **and** (**F**) Chromatin accessibility in promoters of syntenic *PCO* and *PYL (PYRABACTIN RESISTANCE 1 (PYR1) / PYR1-LIKE (PYL) / REGULATORY COMPONENTS OF ABA RECEPTORS (RCAR*) genes. ATAC coverage scale is the same for genes shown in each panel. Below: locations of HRPE or bHLH motifs for four species. (**E**) **and** (**G**) Number of upregulated genes containing motifs classified by syntenic and non-syntenic.

LITERATURE CITED

1. L. A. C. J. Voesenek, J. Bailey-Serres, Flood adaptive traits and processes: an overview. *New Phytol.* **206**, 57–73 (2015).

2. C. Branco-Price, K. A. Kaiser, C. J. H. Jang, C. K. Larive, J. Bailey-Serres, Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in *Arabidopsis thaliana*. *Plant J.* **56**, 743–755 (2008).

3. R. Sorenson, J. Bailey-Serres, Selective mRNA sequestration by OLIGOURIDYLATE-BINDING PROTEIN 1 contributes to translational control during hypoxia in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2373–2378 (2014).

4. P. Juntawong, T. Girke, J. Bazin, J. Bailey-Serres, Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E203–12 (2014).

5. R. B. Deal, S. Henikoff, The INTACT method for cell type–specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nat. Protoc.* **6**, 56–68 (2010).

6. A. Mustroph *et al.*, Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18843–18848 (2009).

7. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*. **10**, 1213–1218 (2013).

8. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. **324**, 218–223 (2009).

9. D. M. Goodstein *et al.*, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).

10. M. Klecker *et al.*, A Shoot-Specific Hypoxic Response of Arabidopsis Sheds Light on the Role of the Phosphate-Responsive Transcription Factor PHOSPHATE STARVATION RESPONSE1. *Plant Physiol.* **165**, 774–790 (2014).

11. K. A. Maher *et al.*, Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell.* 30, 15–36, (2017).

12. P. Gasch *et al.*, Redundant ERF-VII Transcription Factors Bind to an Evolutionarily Conserved cis-Motif to Regulate Hypoxia-Responsive Gene Expression in Arabidopsis. *Plant Cell*. **28**, 160–180 (2016).

13. S. C. Lee *et al.*, Molecular characterization of the submergence response of the *Arabidopsis thaliana* ecotype Columbia. *New Phytol.* **190**, 457–471 (2011).

14. A. Mustroph *et al.*, Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiol.* **152**, 1484–1500 (2010).

15. B. O. R. Bargmann *et al.*, TARGET: a transient transformation system for genome-wide transcription factor target discovery. *Mol. Plant*. **6**, 978–980 (2013).

16. A. Para *et al.*, Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10371–10376 (2014).

17. J. W. Walley *et al.*, Integration of omic networks in a developmental atlas of maize. *Science*. **353**, 814–818 (2016).

18. D. A. Weits *et al.*, Plant cysteine oxidases control the oxygen-dependent branch of the N-end-rule pathway. *Nat. Commun.* **5**, 3425 (2014).

19. J. Barba-Montoya, M. dos Reis, H. Schneider, P. C. J. Donoghue, Z. Yang, Constraining uncertainty

in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytol.* **218**, 819–834 (2018).

20. M. A. Reynoso *et al.*, Nuclear transcriptomes at high resolution using retooled INTACT. *Plant Physiol.* **176**, 270–281 (2018).

21. D. Zhao *et al.*, Analysis of ribosome-associated mRNAs in rice Reveals the importance of transcript size and GC content in translation. *G3* . **7**, 203–219 (2017).

22. M. Ron *et al.*, Hairy root transformation using *Agrobacterium rhizogenes* as a tool for exploring cell type-specific gene expression and function using tomato as a model. *Plant Physiol.* **166**, 455–469 (2014).

23. B. T. Townsley, M. F. Covington, Y. Ichihashi, BrAD-seq: Breath Adapter Directional sequencing: a streamlined, ultra-simple and fast library preparation protocol for strand specific mRNA library construction. *Front. Plant Sci.* 6, 366 (2014) (available at http://europepmc.org/articles/pmc4441129).

24. J. S. Yuan, A. Reed, F. Chen, C. N. Stewart Jr, Statistical analysis of real-time PCR data. *BMC Bioinformatics*. **7**, 85 (2006).

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

26. M. Bajic, K. A. Maher, R. B. Deal, Identification of open chromatin regions in plant genomes using ATAC-seq. *Methods Mol. Biol.* **1675**, 183–201 (2018).

27. M. A. Reynoso *et al.*, Nuclear transcriptomes at high resolution using retooled INTACT. *Plant Physiol.* **176**, 270–281 (2018).

28. M. Reynoso *et al.*, Isolation of Nuclei in Tagged Cell Types (INTACT), RNA extraction and

ribosomal RNA degradation to prepare material for RNA-seq. *BIO-PROTOCOL*. **8** (2018), doi:10.21769/BioProtoc.2458.

29. A. Mustroph, P. Juntawong, J. Bailey-Serres, Isolation of plant polysomal mRNA by differential centrifugation and ribosome immunopurification methods. *Methods Mol. Biol.* **553**, 109–126 (2009).

30. M. A. Reynoso *et al.*, Translating Ribosome Affinity Purification (TRAP) followed by RNA sequencing technology (TRAP-SEQ) for quantitative assessment of plant translatomes. *Methods Mol. Biol.* **1284**, 185–207 (2015).

31. P. Juntawong, M. Hummel, J. Bazin, J. Bailey-Serres, Ribosome profiling: a tool for quantitative evaluation of dynamics in mRNA translation. *Methods Mol. Biol.* **1284**, 139–173 (2015).

32. J. Bazin *et al.*, Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10018–E10027 (2017).

33. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).

34. T. Girke, systemPipeR: NGS workflow and report generation environment. *UC Riverside. https://github. com/tgirke/systemPipeR* (2014) (available at http://www.bioconductor.org/packages/release/bioc/html/systemPipeR.html).

35. L. Calviello *et al.*, Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods*. **13**, 165–170 (2016).

36. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

37. S. Su *et al.*, Glimma: interactive graphics for gene expression analysis. *Bioinformatics*. **33**, 2050–

2052 (2017).

38. M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).

39. D. Koenig *et al.*, Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2655–62 (2013).

40. E. L. L. Sonnhammer, G. Östlund, InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–9 (2015).

41. E. Lyons, M. Freeling, How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008).

42. E. Lyons, B. Pedersen, J. Kane, M. Freeling, The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the Rosids. *Trop. Plant Biol.* **1**, 181–190 (2008).

43. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

44. A. Bolger *et al.*, The genome of the stress-tolerant wild tomato species *Solanum pennellii. Nat. Genet.* **46**, 1034–1038 (2014).

45. P. Sijacic, M. Bajic, E. C. McKinney, R. B. Meagher, R. B. Deal, Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant J.* **94**, 215–231 (2018).

46. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

47. S. Heinz *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*. **38**, 576–589 (2010).

48. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

49. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. **31**, 166–169 (2015).

50. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

51. F. Ramírez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

52. M. Salmon-Divon, H. Dvinge, K. Tammoja, P. Bertone, PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*. **11**, 415 (2010).

53. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994) (available at http://www.cs.toronto.edu/~brudno/csc2417_15/10.1.1.121.7056.pdf).

54. R. C. O'Malley *et al.*, Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*. **165**, 1280–1292 (2016).

55. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. **158**, 1431–1443 (2014).

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

57. R. C. McLeay, T. L. Bailey, Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*. **11**, 165 (2010).

58. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics*. **27**, 1017–1018 (2011).

59. P. Shannon *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

60. S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

61. M. A. Reynoso, F. A. Blanco, J. Bailey-Serres, M. Crespi, M. E. Zanetti, Selective recruitment of mRNAs and miRNAs to polyribosomes in response to rhizobia infection in *Medicago truncatula*. *Plant J.* **73**, 289–301 (2013).

62. S. Gonzali *et al.*, Universal stress protein HRU1 mediates ROS homeostasis under anoxia. *Nat Plants*. **1**, 15151 (2015).

63. S. Hartman *et al.*, Ethylene-mediated nitric oxide depletion pre-adapts plants to hypoxia stress. *bioRxiv* (2019), p. 705194.

64. F. Schröder, J. Lisso, C. Müssig, EXORDIUM-LIKE1 promotes growth during low carbon availability in Arabidopsis. *Plant Physiol.* **156**, 1620–1630 (2011).

65. D. J. Gibbs *et al.*, Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants. *Nature*. **479**, 415–418 (2011).

66. F. Licausi *et al.*, HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in Arabidopsis thaliana. *Plant J.* **62**, 302–315 (2010).

67. B. Giuntoli *et al.*, A trihelix DNA binding protein counterbalances hypoxia-responsive transcriptional activation in Arabidopsis. *PLoS Biol.* **12**, e1001950 (2014).

68. M. D. White *et al.*, Plant cysteine oxidases are dioxygenases that directly enable arginyl transferase-catalysed arginylation of N-end rule targets. *Nat. Commun.* **8**, 14690 (2017).

69. M. D. White, J. J. A. G. Kamps, S. East, L. J. Taylor Kearney, E. Flashman, The plant cysteine oxidases from Arabidopsis thaliana are kinetically tailored to act as oxygen sensors. *J. Biol. Chem.* **293**, 11786–11795 (2018).

70. C. Pucciariello, S. Parlanti, V. Banti, G. Novi, P. Perata, Reactive oxygen species-driven transcription in Arabidopsis under oxygen deprivation. *Plant Physiol.* **159**, 184–196 (2012).

71. E. Yeung *et al.*, A stress recovery signaling network for enhanced flooding tolerance in *Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. 115,* E6085–E6094 (2018).

72. S. R. Cutler, P. L. Rodriguez, R. R. Finkelstein, S. R. Abrams, Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* **61**, 651–679 (2010).

73. M. González-Guzmán *et al.*, Tomato PYR/PYL/RCAR abscisic acid receptors show high expression in root, differential sensitivity to the abscisic acid agonist quinabactin, and the capability to enhance plant drought resistance. *J. Exp. Bot.* **65**, 4451–4464 (2014).

CHAPTER 6: CHROMATIN ACCESSIBILITY CHANGES IN DIFFERENTIATING CELLS OF THE LEAF

Marko Bajic, Liz Dreggors, Juan Dong, Eric Lam, Roger Deal

This work is in the preliminary phase of data analysis and has not been published

SUMMARY

Transcriptional networks orchestrate cellular composition, identity, and response to environmental stimuli. The precise timing and amount of each transcript in a network is regulated by different transcription factors (TFs). The structure of chromatin around a protein coding gene can affect the regulation of that gene by specific TFs. We used the Isolation of Nuclei TAgged in specific Cell Types followed by Assay for Transposase Accessible Chromatin (INTACT-ATAC-seq) to identify chromatin accessibility changes among different stages of differentiation in cells of the stomatal lineage in *Arabidopsis thaliana*. Using previously reported chromatin accessibility profiles for stem cells of the shoot apical meristem (SAM) and differentiated mesophyll cells, along with chromatin profiles generated in this study (asymmetrically dividing meristemoid cells, symmetrically dividing guard mother cells (GMCs), and differentiated guard cells), we identified thousands of genomic regions preferentially accessible to the Tn5 transposase in each cell type. These genomic regions reveal regulatory elements, and nearby regulated genes, that maintain the function and intermediary differentiation potential specific for each cell type. These cell type-specific transposase hypersensitive sites (THSs) were found primarily upstream of genes associated with known biological processes indicative of each cell type. Analysis of the sequences found within cell type-enriched THSs identified hundreds of overrepresented TF binding motifs, several of which were found to be overrepresented only in one cell type. However, further work remains to be done to evaluate the validity of these TF binding motifs, to optimize the approach used to identify these motifs, as well as to build gene regulatory networks for each cell type.

INTRODUCTION

Proper gas exchange between most land plants and their environment is facilitated by the opening and closing of stomatal pores on the leaves of the plant (Zhao et al. 1998). The stomata is made up of two kidney bean-shaped epidermal guard cells that enlarge or shrink in response to changing light, temperature, water, pathogen, or carbon dioxide levels (Lee et al. 2019). Constricted guard cells create an opening in the stomata that releases water vapor and oxygen from the leaf and allows for the intake of carbon dioxide from the environment (Pillitteri et al. 2013). Stomatal guard cells are produced by a dedicated and specialized lineage that is prevalent in developing leaves but becomes inactive after epidermal maturation (Geisler et al. 2000, Nadeau et al. 2003, Bergmann et al. 2007). This lineage originates from the protodermal cells of the developing leaf primordia at the flanks of the shoot apical meristem (SAM) (Barton 2010, Besnard et al. 2011). The protodermal cells are converted to meristemoid mother cells (MMCs) that undergo an asymmetric division to create meristemoid cells. Within the developing leaf, the meristemoid cells may undergo up to three self-renewing asymmetric divisions, creating additional meristemoid cells, before converting to guard mother cells (GMCs) (Dong et al. 2009). The GMC undergoes a single symmetric cell division to form the paired guard cells that mature into a stoma.

Transition between each stage of stomatal development is closely regulated by the master regulator basic helix-loop-helix (bHLH) class Ia transcription factors (TFs) *SPEECHLESS (SPCH)*, *MUTE*, and *FAMA* (Ohashi-Ito et al. 2006, Pillitteri et al. 2007, Adrian et al. 2015). *MUTE* becomes active in meristemoid cells and directs the transition of a meristemoid cell to a GMC, after which *FAMA* becomes expressed in GMCs and directs the transition of a GMC to a guard cell (Ohashi-Ito et al. 2006, Pillitteri et al. 2008, Han et al. 2018). The differentiated guard cell fate is regulated by a R2R3-MYB transcription factor *AtMYB60*, which is specifically expressed in guard cells (Cominelli et al. 2005). We utilized the promoters of these cell state-specific TFs in the context of the INTACT (Isolation of Nuclei TAgged in specific Cell Types) technique to isolate nuclei specifically from meristemoid, GMC, or guard cells (Deal et al. 2010, Deal et al. 2011). This technique utilizes a cell type-specific promoter to drive the expression

of a nuclear targeting fusion (NTF) protein in a specific cell type. The NTF is comprised of a nuclear envelope-targeting domain, a green fluorescent protein, and a biotin ligase recognition peptide (BLRP). The BLRP is biotinylated by a constitutively expressed *BirA* gene, and the NTF localized nuclei can be affinity purified using streptavidin-coated magnetic beads.

We utilized Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) to measure DNA accessibility in the isolated nuclei from the different cell types (Buenrostro et al. 2013, Maher et al. 2018). This combined INTACT-ATAC-seq has been successfully utilized in seedlings and roots of different plant species, as well as mesophyll cells and stem cells of the central zone and the peripheral zone of the shoot apical meristem of *Arabidopsis thaliana* (Lu et al. 2017, Bajic et al. 2018, Sijacic et al. 2018, Frerichs et al. 2019). In this study, we analyzed chromatin accessibility levels in stem cells of the shoot apical meristem and fully-differentiated mesophyll cells (Sijacic et al. 2018) along with three distinct cell states of the stomatal lineage: proliferating meristemoid cells, round guard mother cells, and fully-differentiated guard cells of the leaf epidermis. The comparison of these five different cell types illuminates chromatin regulation during cell differentiation along two differentiation paths, from the starting point, two intermediary points, and two different endpoints. Our results show that while most Transposase Hypersensitive Sites (THSs) have similar levels of accessibility in the three cell types of the stomatal lineage, thousands of regions could be identified that were quantitatively more accessible in each of the cell types analyzed relative to the others. Furthermore, we identified transcription factors (TFs) that likely bind overrepresented sequence motifs in each cell type-enriched group of THSs. For each group of THSs, we picked the single most significantly enriched motif and mapped its occurrence throughout the genome. The motif occurrences were then intersected with the different THS groups to validate the cell type-specific enrichment of each motif. The initial results did not support the observation that the most significantly enriched motif discovered by AME occurred at higher amounts in a cell type-specific group of THSs compared to the other groups of THSs. Further work is needed to apply transcriptomic data to the cis-motif analysis pipeline and identify TF-binding motifs that are specifically enriched in one group of cell type-enriched THSs, as well as combining the transcriptomic data with the ATAC-seq results to build gene

regulatory networks for each cell type. Understanding chromatin dynamics within sequential differentiation states of the three stomatal lineage cell types will help us better understand how TFs organize chromatin, what other TFs are important for maintaining the identity of each cell type, and how guard cells are developed and maintained.

RESULTS

Validation of cell type-specific NTF expression and INTACT-ATAC-seq data

The *MUTE*, *FAMA*, and *AtMYB60* genes are exclusively expressed in the meristemoid, guard mother cells, and guard cells, respectively (Cominelli et al. 2005, Ohashi-Ito et al. 2006, Pillitteri et al. 2008). The upstream regulatory sequences of these genes were used to drive the expression of the nuclear targeting fusion (NTF) transgene selectively in the meristemoid, guard mother cells, and guard cells. The creation of *MUTEp::NTF;ACT2p::BirA*, *FAMAp::NTF;ACT2p::BirA*, and *AtMYB60::NTF;ACT2p::BirA* constructs and the transformation and establishment of transgenic *Arabidopsis thaliana* T2 plants with these vectors was done in Eric Lam and Juan Dong's labs at Rutgers University. Cell type-specific expression of the NTF was confirmed using fluorescence microscopy. The Green Fluorescent Protein (GRP), which is part of the NTF, was detected specifically in meristemoid, guard mother cells, and guard cells in *MUTEp::NTF;ACT2p::BirA*, *FAMAp::NTF;ACT2p::BirA*, and *AtMYB60::NTF;ACT2p::BirA* transgenic lines, respectively (Figure 6.1A).

Nuclei were purified from *MUTE*, *FAMA*, or *AtMYB60* promoter-driven INTACT lines as previously described (Bajic et al. 2018). A total of 37,500 freshly isolated nuclei for each replicate were used for ATAC-seq, and three biological replicates were performed per cell type. At least 14 million reads were obtained using paired-end sequencing for each replicate (Figure S6.1A). On average, 95% of sequenced reads aligned to the *Arabidopsis thaliana* TAIR10 genome, 66% of these reads aligned to the nuclear genome, and 51% of these reads had a quality score of 2 or more (Figure S6.1A). Processed alignment files for the three stomatal lineage ATAC-seq triplicates, as well as previously published stem

cell and mesophyll triplicates (Sijacic et al. 2018), revealed segregation of replicates by cell type, with minimal variation among replicates of the stem cell and mesophyll datasets, but with higher variation in the stomatal lineage samples (Figure S6.1B). This separation of samples indicated the high similarity of genome-wide chromatin accessibility in stomatal lineage cell types compared to the more distinct profiles of the undifferentiated stem cells of the shoot apical meristem or the alternate, fully differentiated mesophyll cells. We analyzed the fragment size distribution for each sample to determine the distribution of nucleosome-containing (>150 bp) and nucleosome-free reads (<150 bp). In all the meristemoid, GMC, and guard cell replicates we observed fragments primarily of 100 bp in size (Figure S6.1C), which is consistent with ATAC-seq fragment sizes observed before (Bajic et al. 2018, Sijacic et al. 2018). This distribution is indicative of nucleosome-free reads, which are chromatin regions where transcription factors (TFs) are more likely to be bound (Buenrostro et al. 2013).

To further confirm cell type-specific expression of NTF we used Integrated Genome Viewer (IGV) to visualize chromatin accessibility around genes of promoters used to drive the expression of NTF. Chromatin accessibility was highest around the *CLV3*, *MUTE*, *FAMA*, *MYB60*, and *RBCS2B* genes in stem cell, meristemoid, GMC, guard cell, and mesophyll ATAC-seq datasets (Figure 6.1B). In summary, sufficient numbers of high-quality reads, indicative of input cell state, were obtained from each cell type using INTACT-ATAC-seq.

Identification and characterization of cell type-specific regions of chromatin accessibility

To characterize regions where chromatin accessibility differs among the five different cell types we used the peak calling function of HOMER (Heinz et al. 2010) to identify transposase hypersensitive sites (THSs) in each cell type. Reproducible THSs, those that overlapped by at least 50% between replicates, were identified in each cell type. This resulted in 53,426 THSs reproducibly called in meristemoid samples, 53,474 in guard mother cell samples, and 58,550 in guard cell samples. Averaged chromatin accessibility in meristemoid, GMC, and guard cell samples was plotted across the three different sets of THSs, validating

distinct sites of high chromatin accessibility at the center of all the THSs (Figure S6.2A-C). HOMER was also used to annotate each THS to its genomic feature, and the THSs reproducibly identified in each cell type were found primarily in promoters (0 to 2 kb upstream from the transcription start site), with little variation of distribution between the three groups of THSs (Figure S6.2D-F).

THSs from meristemoid, GMC, and guard cell were combined with THSs called in stem cell and mesophyll cells, using the same peak calling parameters, and THSs in this combined set were merged together if they overlapped by at least 60 base pairs. This resulted in 44,313 THSs that represent the accessible sites present in at least one cell type. We used HTSeq's *htseq-count* script (Anders et al. 2015) to calculate the read amount at each THS in all five cell types and used DESeq2 (Love et al. 2014) to normalize and statistically evaluate the accessibility levels in all five cell types. Principal component analysis comparing chromatin accessibility levels for the five cell types across the 44,313 THSs revealed clear separation of cell type-specific accessibility patterns among the five cell types and their tightly associated replicates, with the meristemoid, GMC, and guard cell accessibility profiles still clustering close to each other (Figure S6.3A). We utilized k-means clustering to identify THSs that are most accessible in one cell type. This identified 4,363 THSs in k-means Cluster 1, 8,881 THSs in cluster 3, 4,099 THSs in cluster 4, 6,279 THSs in cluster 2, and 1,673 THSs in cluster 6 that are predominantly enriched in stem cells, guard mother cells, guard cells, or mesophyll cells, respectively (Figure 6.1C). Chromatin accessibility in all five cell types, as well as genomic DNA (Bajic et al. 2018), was plotted across the 2 kb window of each stem cell-enriched, meristemoid-enriched, GMC-enriched, guard cell-enriched, and mesophyll-enriched THSs (Figure 6.2A, S6.3B). The metaplots and heatmaps confirmed cell type-specific enrichment of chromatin accessibility at the THS sites designated as enriched within that cell type.

The five different groups of THSs were assigned to genomic features and to the nearest protein coding gene (PCG). The different groups of THSs were primarily found in promoters of genes, specifically just upstream of the transcription start site (TSS) (Figure S6.3C). Metaplots also confirmed that chromatin accessibility around the genes nearest to the cell type-enriched groups of THSs was more accessible within

that cell type, but this enrichment across the gene body was not drastically different among the five cell types (Figure 6.2B). Genes nearest to the different cell type-enriched THSs were separated by cell type using a Venn diagram to identify groups of genes that had a nearby THS or THSs that were enriched in only one cell type (Figure S6.4A). These unique groups of genes were analyzed using AgriGO to find enriched gene ontology (GO) terms. The resulting GO terms that had a p-value of 0.05 or less were clustered to find terms only found in one cell type and these unique terms were visualized using ReviGO (Supek et al. 2011) (Figure S6.4B, 6.2C). Cell type-specific terms, such as "regulation of developmental process" in stem cells, "anatomical structure arrangement" in meristemoid, "callose localization" in GMC, "developmental growth" in guard cells, and "circadian rhythm" in mesophyll cells, further support the cell type-specific accessibility profiles assigned to the 5 clusters of THSs found near these genes. Overall, we were able to identify chromatin accessibility sites in the five different cell types and determine which THSs had accessibility that was highest in each of the five different cell types.

Identification and validation of enriched motifs in cell type-enriched THSs

The cell type-enriched THSs represent chromatin regions that are most accessible in each cell type, and TF-binding sequences that are found within these regions are most likely to be occupied by the sequence-specific TFs to regulate nearby genes within the cell types where these sites are most accessible. We used Analysis of Motif Enrichment (AME) (McLeay et al. 2010) on repeat-masked sequences within these THS regions to identify specific transcription factors that may regulate gene expression through these accessible sites. The DAP-seq (O'Malley et al. 2016) and CIS-BP (Weirauch et al. 2014) databases were used to match overrepresented sequences to possible TFs that bind these sequences (Figure 6.3A). Venn diagrams of TFs that bind overrepresented motifs identified from the two databases were overlapped to identify TFs-binding sequences that are unique to only one cell type (Figure S6.5 A-C).

There were 32, 22, 59, 44, and 14 TF-binding sequences found only in stem cell, meristemoid, GMC, guard cell, or mesophyll-enriched THSs (Figure S6.6A, S6.7A, S6.8A, S6.9A, S6.10A). The

transcription factors that bind these sequences had overrepresented GO terms representative of the specific cell type where the motif they bind was found to be overrepresented. This includes terms such as “meristem initiation” for shoot apical meristem stem cells (Figure S6.6B), “phyllome development” for meristemoid cells (Figure S6.7B), “cell differentiation” for guard mother cells(Figure S6.9A), “regulation of cell differentiation” for guard cells (Figure S6.9B), and “regulation of circadian rhythm” for mesophyll cells (Figure S6.10B). The full list of all the overrepresented motifs, the corresponding E-value and positional weight matrices found in THSs enriched in stem cells, meristemoid cells, guard mother cells, guard cells, and mesophyll cells can be found in Figures S6.11, S6.12, and S6.13.

To verify that the highest scoring motifs identified in only one group of cell type-enriched THSs were specifically enriched in only that group of THSs we used FIMO to map the genomic distribution of these motifs (Grant et al. 2011) (Figure 6.3B). By intersecting the genomic locations of each motif with the cell type-enriched THSs we identified the distribution of each motif within each group of THSs. Finally, we computed the percentage of genomic motifs that were found within a group of THSs as well as the percentage of THSs that had at least one motif (Figure 6.3C). The *EARLY-PHYTOCHROME-RESPONSIVE1* (*EPR1*) TF was the only motif that had a clear motif enrichment within the cell type where it was identified as uniquely enriched. All of the other cell type-specific motifs that were analyzed were found at high percentages in additional groups of cell type-enriched THSs, or at high amounts in the wrong group of THSs. The latter is the case for MYB61, a TF whose binding sequence was found to be enriched only in guard mother cell-enriched THSs but the highest proportion of THSs with that motif are found in meristemoid-enriched THSs. Overall, the motifs identified within each set of cell type-enriched THS sequences appear to be specific to functions associated with that cell type, but simply choosing the motif with the highest E-value may not identify motifs that are statistically more enriched within that set of THSs. Additional work to identify functional, overrepresented motifs needs to be done by incorporating RNA-seq results to first confirm that the TF that binds the motif of interest is actually expressed within that cell type. Secondly, page rank analyses can be done to distinguish among the different cell types to identify TFs that

are expressed at elevated levels within specific cell types. Finally, RNA-seq results can be used to determine whether genes found near an accessible motif are regulated at different levels within that cell type compared to the other cell types where the motif is accessible at lowered levels.

DISCUSSION

Prior work to characterize chromatin changes between the starting point in stem cells and the end point in differentiated mesophyll cells identified clear differences in chromatin organization, but it also highlighted a high level of similar chromatin accessibility between the two cell types (Sijacic et al. 2018). Through this work we built upon prior work by introducing three additional cell states that are developmentally highly similar to each other but distinct from the two cell types studied before. By incorporating data from all five cell types we sought to further evaluate the chromatin accessibility profiles that do not change between the stem cells and mesophyll cells, potential accessible sites near leaf specific genes or homeostasis genes, and to also distinguish among the different cell types and find chromatin accessible sites that are specifically utilized in one cell type compared to the rest. We observed that chromatin accessibility enriched specifically in each of the five cell types analyzed had moderate accessibility in the other cell types (Figure 6.2A). This is particularly true of the three stomatal lineage cell states, where the regions of accessibility that were identified as specific to one cell state had moderate accessibility in the other cell types, but almost identical levels of accessibility in the stem cell. The observation that accessible sites in undifferentiated stem cells become more accessible in subsequent differentiating cell types was previously reported in the original analysis (Sijacic et al. 2018). This observation is further supported here, indicating the differentiation potential present in stem cells. These sites may already be bound by transcription factors that are not yet being utilized to promote transcription. Alternatively, these accessible sites are being used to repress transcription in stem cells, either through the dual regulatory ability of the same transcription factor or by another TF binding the same sequence or a neighboring sequence that cannot be easily discerned using our approach. The ability to validate cell type-enriched THSs, the cis-regulatory elements found within these

sequences, and which genes are being regulated is dependent on complimentary transcriptiomic information from each cell type. While this information exists for each cell type (Endo et al. 2014, Adrian et al. 2015, You et al. 2017), there are several considerations that limit how this information can be used to further evaluate the chromatin accessibility information reported here. These considerations include: 1) sequencing techniques used, 2) age of plants used, 3) specificity of cell types analyzed, and 4) cellular localization of the RNA molecules. Current work is being done within our lab to generate RNA-seq libraries from each cell type using nuclear RNA as the starting material. RNA-seq information from the nuclei of age-matched plants of the same lines that were used to create the ATAC-seq libraries would reduce the confounding variability between datasets and increase the ability to make accurate connections between accessible regulatory elements and genes that are being regulated by the TFs that bind these regulatory elements. The proposed improvements for identifying cis-regulatory elements and using them to build gene regulatory networks (GRNs) specific to each cell type are detailed in Figure 6.4. Additionally, having chromatin accessibility and gene expression data for each cell state in the stomatal lineage would be a valuable resource to the science community because specific networks of genes could be selected for the pipeline in Figure 6.4 and used to find TFs that regulate those networks, such as response to a specific stress.

Overall, the chromatin accessibility data generated for the three stomatal lineage cell states fits in nicely with the quality of data already reported in stem cells and mesophyll cells (Sijacic et al. 2018). This is true for the number of reads, fragment size distribution, THS sizes, and genomic localization of THSs in the three new cell states. We were able to identify THSs that are specifically enriched in each cell type. This set of coordinates will be used in future experiments to compliment RNA-seq information to build gene regulatory networks that can shed light in the differentiation control intrinsic to the stomatal lineage and development in *Arabidopsis* leaves.

METHODS

Plasmid DNA constructs and transformation

We used promoters of *MUTE*, *FAMA*, and *MYB DOMAIN PROTEIN 60 (MYB60)* genes, known to be transcribed specifically in meristemoid, guard mother cells, and guard cells, respectively (Cominelli et al. 2005, Ohashi-Ito et al. 2006, Han et al. 2018), to drive cell-type specific expression of the *Nuclear Targeting Fusion (NTF)* gene. Transgenic *Arabidopsis thaliana MUTE::NTF*, *FAMA::NTF*, and *MYB60::NTF* lines were created by the members of the Eric Lam and Juan Dong laboratories at Rutgers University.

Plant growth conditions

Arabidopsis thaliana plants of the Columbia (Col-0) ecotype carrying either the *MUTE::NTF*, *FAMA::NTF*, or *MYB60::NTF* constructs were stratified at 4°C for 2 days before being moved to the growth chamber. These plants were then grown for 19 days under fluorescent lights, with 16 hour light-8 hour dark cycle at 20°C, at which point tissue was harvested for nuclei isolation.

Microscopy

Developing (smaller than 1cm) and mature (bigger than 1 cm) leaves from three week old *MUTE::NTF*, *FAMA::NTF*, and *MYB60::NTF* plants were imaged using an OLYMPUS BX53 fluorescent microscope with 20x objectives to confirm proper expression and localization of the NTF protein. GFP fluorescence was visualized using the FITC lens filter. For each sample, the leaf tissue was immersed in water, covered with a cover slip, and the tops and bottoms of each leaf were imaged.

Nuclei isolation using INTACT

Nuclei purification from specific cell types using the Isolation of Nuclei TAgged in specific Cell Types (INTACT) method was performed as described previously (Bajic et al. 2018). Between 80 mg and 200 mg of fresh leaf tissue was harvested from each transgenic line, in triplicate, for nuclei isolation. Young leaves were collected from *MUTE::NTF* and *FAMA::NTF* plants, and old leaves were collected from *MYB60::NTF* plants. The collected tissue was tissue was finely chopped with a razor blade in Nuclei Purification Buffer (NPB) on ice. For each preparation, a volume of 10 µl of Streptavidin M280 magnetic

beads was used to capture biotinylated nuclei. Captured nuclei were imaged and counted using a hemocytometer, yielding between 37,500 to 160,000 nuclei.

Assay for transposase accessible chromatin (ATAC) and library preparation

Transposase tagmentation and sequencing library preparation was carried out on freshy isolated, never frozen, nuclei as previously described (Sijacic et al. 2018). Briefly, 37,500 freshly purified nuclei from each sample were resuspended in 50 µl of transposition reaction mix and incubated at 37°C for 30 minutes, with gentle mixing every 10 minutes, using Nextera reagents (Illumina, FC-121-1030). Tagmented DNA was purified using the Qiagen PCR purification kit and eluted in 11 µl of elution buffer. The samples were stored at -20°C before library preparation. Libraries were amplified using the 2X high fidelity PCR mix (NEB) with custom barcoded primers (Table 6.1). Quantitative PCR (qPCR) was used to determine the optimal number of additional amplification cycles necessary for each sample, resulting in 13-14 total cycles of amplification for the different samples. Libraries were quantified using qPCR with the NEBNext Library Quant Kit from Illumina. Library size distribution was determined for each sample using the Agilent 4200 TapeStation system.

High throughput sequencing

Next-generation sequencing was done at the Georgia Genomics Facility at the University of Georgia using the NextSeq 500 instrument (Illumina). All libraries were pooled and sequenced in the same flow cell using paired-end 150 nt reads.

Sequence read mapping, processing, and visualization

The read number, quality, and lengths of ATAC-seq reads were assessed using the FastQC app (v0.11.5, www.bioinformatics.babraham.ac.uk/projects/fastqc). Sequencing reads were processed to remove adapter sequences by trimming each read down to 36 bases using Trimmomatic (Bolger et al. 2014). The trimmed reads were mapped to the *Arabidopsis thaliana* genome (version TAIR10) using the Bowtie2 software

(Langmead et al. 2012) with default parameters. Mapped reads in *.sam* format were converted to *.bam* format, sorted, and indexed using Samtools 0.1.19 (Li et al. 2009). Further filtering was done using Samtools to retain only those reads that had a mapping quality score of 2 or higher and were mapped to nuclear chromosomes. For visualization, the filtered, sorted, and indexed *.bam* files were scaled to the same number of reads and were converted to the bigwig format using the *"bamcoverage"* script in deepTools 2.0 (Ramirez et al. 2014) with a bin size of 1 bp and RPKM normalization. Heatmaps and average plots displaying the ATAC-seq data were generated using SeqPlots (Stempor et al. 2016). Genome browser images were generated using the Integrative Genomics Viewer (IGV) 2.3.68 (Thorvaldsdottir et al. 2013).

Peak calling to detect transposase hypersensitive sites (THSs)

Peak calling on ATAC-seq data was done by employing the "*Findpeaks*" function of the HOMER package (Heinz et al. 2010) using the "*-minDist 150",* "*-region*", and *'regionRes 1"* options to identify sites of high accessibility. Called peaks, referred to as Transposase Accessible Sites, or THSs, were processed using Bedtools (Quinlan et al. 2010) to identify THSs called in at least on other replicate for a given cell type. This was done by keeping any THSs that overlapped by at least 50% between the replicates of that cell type. The retained THSs were concatenated and then merged together if they overlapped by at least 60 base pairs (50% of the majority of the peak sizes).

Genomic distribution of THSs

The distribution of THSs relative to genomic features was identified using the "annotatePeaks.pl" package from HOMER (Heinz et al. 2010) with "upstream" regions set as the 2,000 bp upstream of the annotated transcription start site, and "downstream" regions set as the 1,000 bp downstream of the transcript end site. Pie charts depicting the genomic distribution were generated using the ggplot2 package for R (Wickham 2009).

Assignment of THSs to nearby genes

For each group of accessible sites, the THS coordinates were assigned to nearest target protein-coding genes using the *"TSS"* function of the PeakAnnotator 1.4 program (Salmon-Divon et al. 2010). In addition to identifying the nearest protein-coding gene for each THS, this program also reports the distance from the peak center to the assigned TSS.

THSs enriched in a specific cell type

In order to identify THSs that are more accessible in a specific cell type, all of the reproducible THSs identified in the different cell types (Stem Cell, Meristemoid, Guard Mother Cell, Guard Cell, and Mesophyll) were concatenated and then merged if they overlapped by at least 150 base pairs. This set of THSs represents the chromatin landscape where accessibility was observed at least twice in one cell type. The amount of accessibility at each THS was calculated using HTSeq's *htseq-count* script (Anders et al. 2015) by identifying the number of reads (counts) present in each THS. Three ATAC-seq replicates of each cell type were counted and the counts were processed using DESeq2 (Love et al. 2014). THSs were identified as enriched in a specific cell type by plotting k-means clustered count data for all the THSs across all the cell types using the *"cluster"* (https://CRAN.R-project.org/package=cluster) and *"pheatmap"* (https://CRAN.R-project.org/package=pheatmap) packages in R.

Gene ontology analysis

Gene ontology (GO) analysis on specific gene lists was done using either the AgriGO GO Analysis Toolkit (Du et al. 2010, Tian et al. 2017), with default parameters, or by using the built in GO analysis toolkit of STRING (Szklarczyk et al. 2017). GO terms were considered significant if they had a false discovery rate (FDR) of 0.05 or less. Multiple GO terms were visualization using ReviGO (Supek et al. 2011).

Transcription factor motif analysis

Cell type-enriched THSs were used for motif analysis. The sequences present in the THS coordinates were isolated using the Bedtools "getfasta" command and a repeat masked TAIR10 genome sequence of

Arabidopsis thaliana. The isolated sequences were analyzed using Analysis of Motif Enrichment (AME) (McLeay et al. 2010), with default parameters, to identify motifs enriched in each group of THSs. The motif enrichment was done using both DAP-seq (O'Malley et al. 2016) and CIS-BP (Weirauch et al. 2014) databases to match enriched motifs to known transcription factors. Motifs were considered significant if they had an E-value of 0.05 or less.

Transcription Factor interaction analysis using STRING

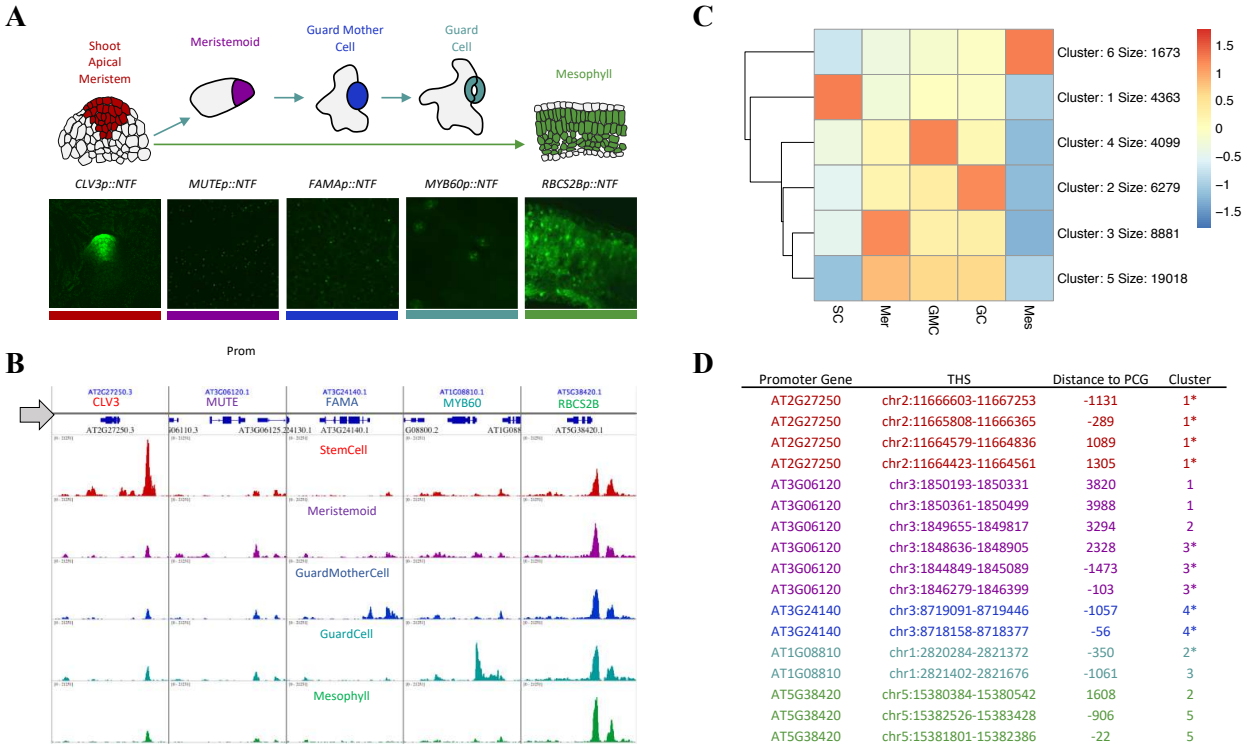
Groups of TFs that have predicted interactions were analyzed using the STRING database. Interactions are predicted based on co-expression data, publication co-ocurrences, colocalization, gene orthology, and experimental data. The interacting TF connections were visualized under default parameters.

Mapping genomic locations of motifs

FIMO (Grant et al. 2011) was used to map positional weight matrices within the repeat-masked sequence of the *Arabidopsis thaliana* TAIR10 genome. Significant TF motif occurrences were those with a p-value of 0.0001 or less. Bedtools was used to intersect genomically mapped predicted binding sites with cell type-enriched THS coordinates to identify which THSs contained motifs and which motifs were found in THSs

TABLES AND FIGURES

Fig 6.1



D

Promoter Gene	THS	Distance to PCG	Cluster
AT2G27250	chr2:11666603-11667253	-1131	1*
AT2G27250	chr2:11665808-11666365	-289	1*
AT2G27250	chr2:11664579-11664836	1089	1*
AT2G27250	chr2:11664423-11664561	1305	1*
AT3G06120	chr3:1850193-1850331	3820	1
AT3G06120	chr3:1850361-1850499	3988	1
AT3G06120	chr3:1849655-1849817	3294	2
AT3G06120	chr3:1848636-1848905	2328	3*
AT3G06120	chr3:1844849-1845089	-1473	3*
AT3G06120	chr3:1846279-1846399	-103	3*
AT3G24140	chr3:8719091-8719446	-1057	4*
AT3G24140	chr3:8718158-8718377	-56	4*
AT1G08810	chr1:2820284-2821372	-350	2*
AT1G08810	chr1:2821402-2821676	-1061	3
AT5G38420	chr5:15380384-15380542	1608	2
AT5G38420	chr5:15382526-15383428	-906	5
AT5G38420	chr5:15381801-15382386	-22	5

Figure 6.1 Identification of cell type-enriched THSs. (**A**) Cartoon of stages in leaf development with confocal or fluorescent microscope images of INTACT lines used for nuclei isolation. Specific promoters used to isolate nuclei from the indicated cell types are: *CLV3p::NTF*, shoot apical meristem, red; *MUTEp::NTF*, meristemoid, purple; *FAMAp::NTF*, guard mother cell, red; *MYB60p::NTF*, guard cell, teal; and *RBC2S2Bp::NTF*, mesophyll, green. Confocal images of *CLV3p::NTF* and *RBC2SBp:NTF* are from previously published work from our lab (Sijacic et al. 2018) and are shown here for reference, the rest of the images were captured on a fluorescent microscope. Green fluorescence is observed specifically in nuclei where the Nuclear Targeting Fusions (NTF) is expressed. **(B)** Integrated Genome Viewer snapshots of normalized ATAC-seq reads from shoot apical stem cells (red), meristemoid cells (purple), guard mother cells (blue), guard cells (teal), and mesophyll cells (green). Reads from each cell type are shown around the different genes (grey arrow), shown as different columns, for which the promoter sequence was used to drive cell type-specific expression of NTF. **(C)** Heatmap of chromatin accessibility

levels across 44,313 Transposase Hypersensitive Sites (THSs) identified in at least one cell type. The read amount at each THS was calculated and normalized for each THS in all 5 cell types (SC = Stem Cell, Mer = Meristemoid, GMC = Guard Mother Cell, GC = Guard Cell, Mes = Mesophyll). The heatmap was clustered using 6 k-means clusters. The number of THSs in each cluster is shown at the end of each row, referred to as "Size." The read amounts are averaged across the row and the variation from this average, corresponding to a value of 0, is colored as red if the value is more than the average and blue if the value is lower than the average. The increased or decreased value corresponds to the fold difference in the averaged read amounts of a specific square compared to the average value across all the squares of that row. **(D)** THSs found nearest to the genes whose promoters were used to drive expression of each cell type (AT2G27250, CLV3, red; AT3G06120, MUTE, purple; AT3G24140, FAMA, blue; AT1G08810, MYB60, teal; AT5G38420, RBCS2B, green) were identified and are listed in the THS column, along with the distance of that THS to the nearest Protein Coding Gene (PCG) and which cluster that THS falls in for C. Asterisk indicates THSs that had highest chromatin accessibility in cell types where the promoter was expected to be most accessible.

Fig 6.2

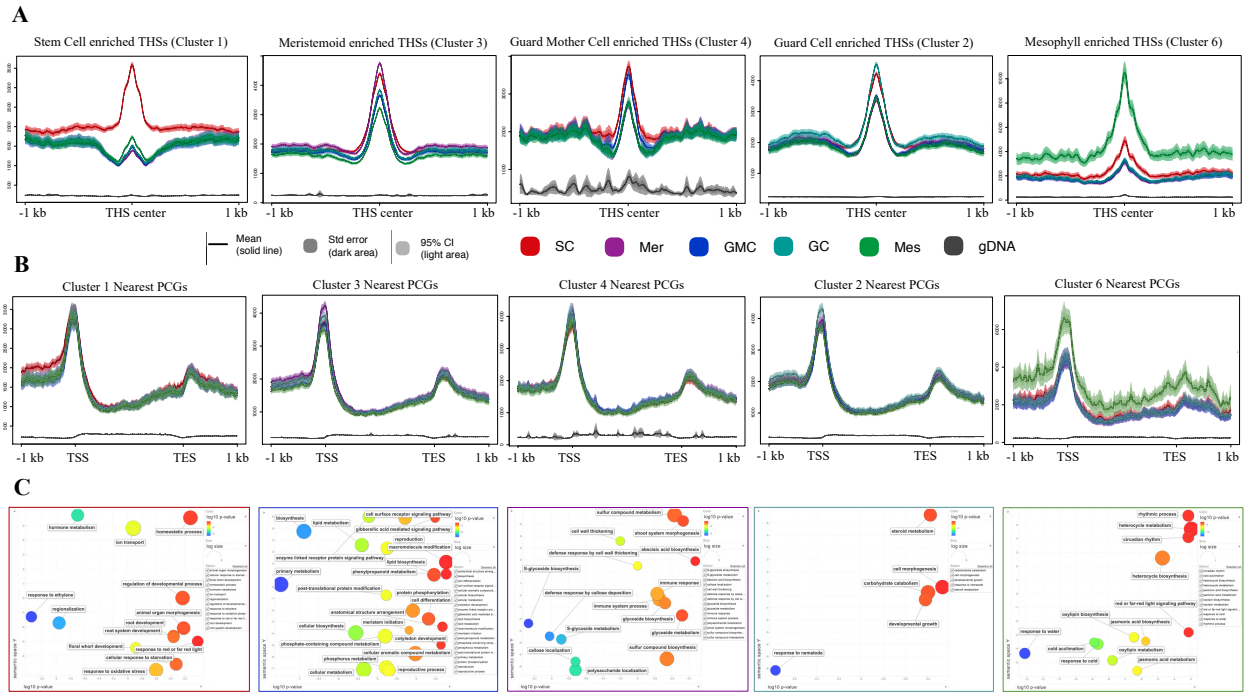


Figure 6.2 Cell type-enriched THSs and nearest genes. (**A**) Profile plots of normalized ATAC-seq reads at THSs from the 5 cell type-enriched clusters from figure 6.1C (Cluster 1 = stem cell enriched, Cluster 3 = Meristemoid enriched, Cluster 4 = Guard Mother Cell enriched, Cluster 2 = Guard Cell enriched, Cluster 6 = Mesophyll enriched). Plots are centered on each THS and show read density across 2 kb relative to the peak center. Each line represents the mean of read density with standard error shaded in dark color and the 95% confidence interval shaded in lighter coloration. Read density from the different normalized libraries is color coded: SC, stem cell, red; Mer, meristemoid, purple; GMC, guard mother cell, blue; GC, guard cell, teal; Mes, mesophyll, green; gDNA, genomic DNA, black. **(B)** Profile plots of the nearest protein coding genes (PCGs) of the THSs from the cell type-specific clusters indicated. The plot is anchored to -1kb from the Transcription Start Site (TSS) and +1kb from the Transcription End Site (TES). **(C)** Same groups of protein coding genes as in B were analyzed using STRING and the unique GO terms and corresponding p-values were plotted using ReviGO.

Fig 6.3

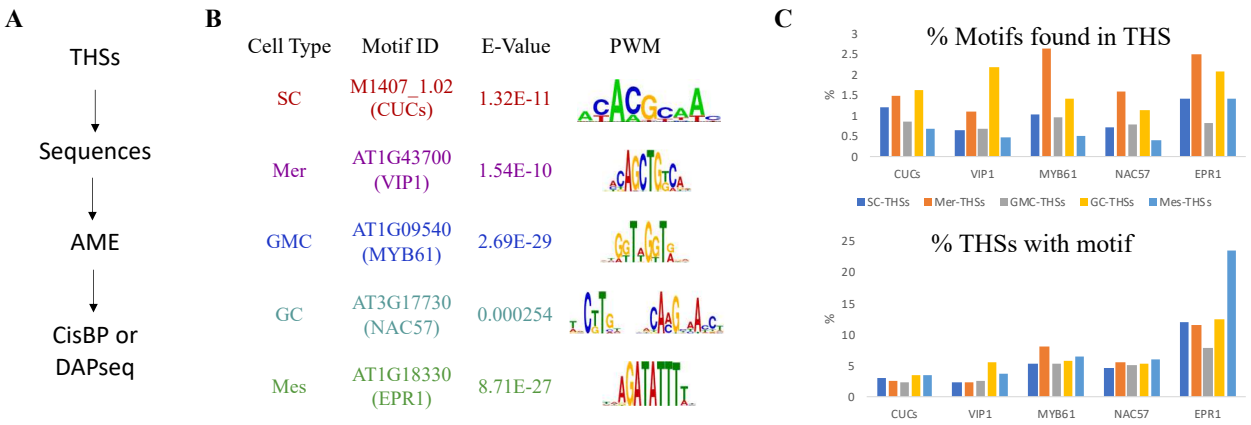


Figure 6.3 Identification and validation of most enriched motifs in each cell type. (**A**) Pipeline for discovery of motifs enriched in cell type-enriched THSs. Sequences that correspond to coordinates of cell type-enriched THSs identified in 1C were isolated and analyzed using the Analysis of Motif Enrichment (AME) tool with either the CisBP or Dap-seq libraries as the databases of queried motifs. **(B)** Identified enriched motifs with an E-value of 0.05 or less were identified and compared among cell types, keeping only those that were observed only in one cell type. Shown are the identified motifs, unique to one cell type, with the most significant E-score value within that cell type. **(C)** Find Individual Motif Occurences (FIMO) was used to map motif locations of the five cell type specific motifs throughout the Arabidopsis genome. Top panel represents the percentage of total motifs in the Arabidopsis genome that are found within a cell type-enriched THS, and the bottom panel represents the percentage of cell type-enriched THSs that contain a motif. Cell type-enriched THSs are stem cell-enriched THSs from cluster 1 (SC-THSs, blue), meristemoid-enriched THSs from cluster 3 (Mer-THSs, orange), Guard mother cell-enriched THSs from cluster 4 (GMC-THSs, grey), guard cell-enriched THSs from cluster 2 (GC-THSs, yellow), and mesophyll-enriched THSs from cluster 6 (Mes-THSs, light blue).

Fig 6.4

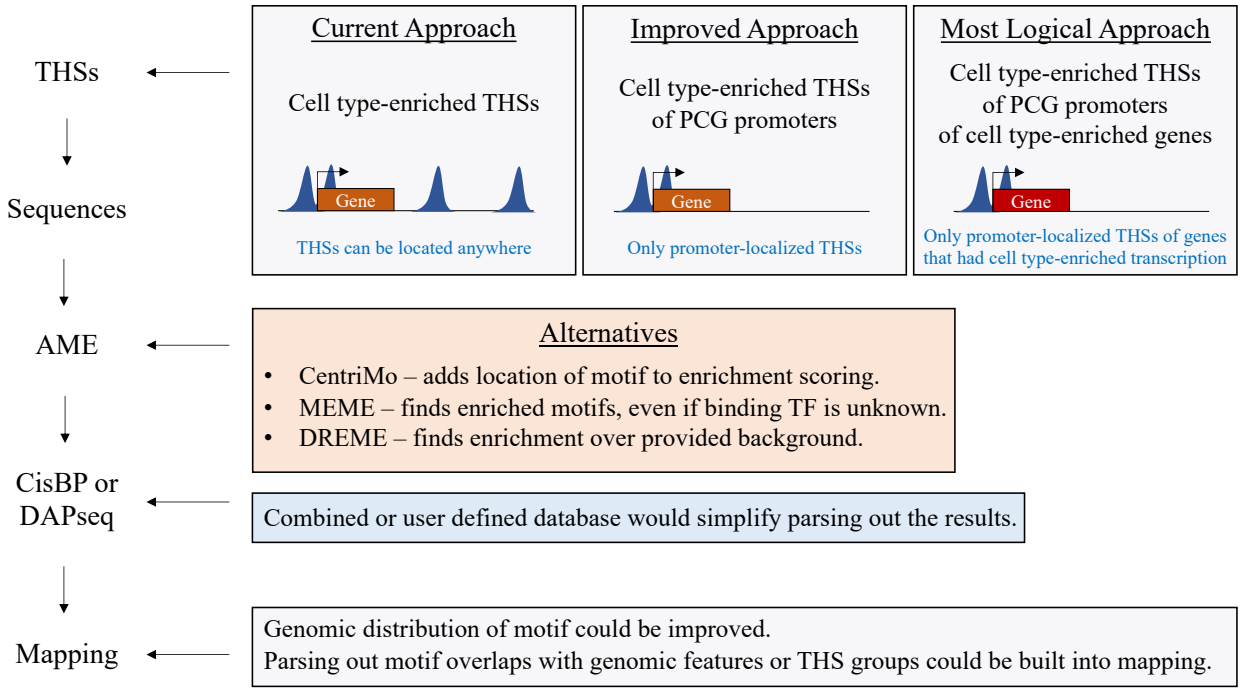


Figure 6.4 Pipeline and potential improvements for the discovery of cell type-specific motif sequences. Current pipeline (left) for the discovery and genomic mapping of enriched sequence motifs is reliant on four stages that can be optimized for future applications. Stage 1 (top diagrams) involves the choosing of which THS sequences are analyzed for motif enrichment. Current approach is to analyze any cell-type enriched THSs regardless of genomic location. Two improvements to this approach include keeping only promoter-localized THSs and using RNA-seq data to subset the THSs even further, keeping only promoter-localized THSs that are in promoters of genes that have increased transcription in the cell type where the THSs of concern are being analyzed. Stage 2 (second diagram from the top) involves the analysis program to find motif enrichment. Analysis of Motif Enrichment (AME) could be replaced with location considering enrichment of CentriMo, novel discovering program MEME, or utilize the background sequence occurrence rates by utilizing DREME. Stage 3 (third diagram from the top) involves the choice of database that enriched motifs are queried against to discover known factors that bind this sequence. For additional result matching, several databases should be queried at the same time.

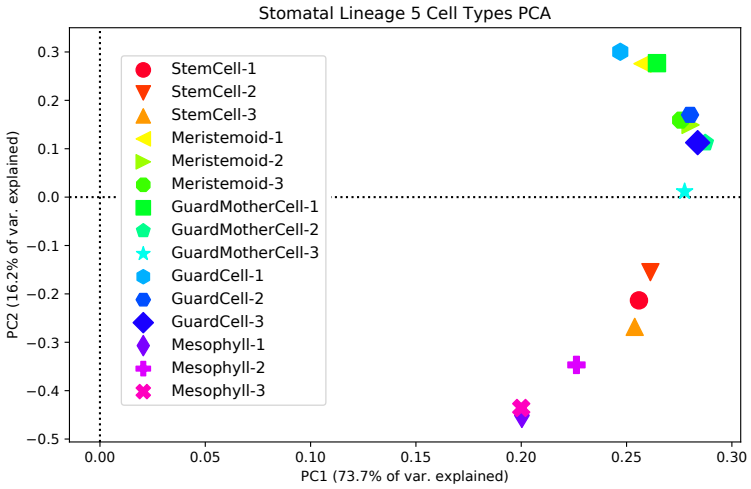
Stage 4 (bottom diagram) involves the mapping of identified motifs. Positional weight matrices can be mapped with other programs genome wide and the process of overlapping the mapped sites with specific genomic features should be built into the mapping process.

Fig S6.1

A

Sample	Aligned	q2-WithOrg	q2-OrgFree	THSs	FRIP	THS-Size
Mer1	26485930	19487756	13991017	36838	39.48	189
Mer2	14047900	8549268	3340456	25002	31.54	159
Mer3	17008670	11125991	4945148	27628	31.47	170
GMC1	32624342	24700453	18755054	39398	39.88	194
GMC2	26202984	16061607	6399073	30099	31.82	169
GMC3	19329628	11520227	4200579	21812	24.40	161
GC1	52065604	39919561	30621083	41090	43.56	223
GC2	52221338	32878114	13266945	36594	37.09	185
GC3	21758596	13810257	5623819	28312	31.92	172

B



C

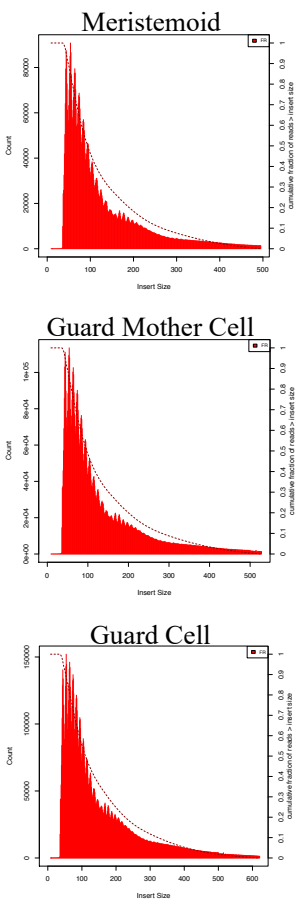


Figure S6.1 ATAC-seq read alignment quality, sample variation, and fragment size distribution.

(A) Sequencing summary for the nine samples. ATAC-seq samples of the three different cell types were done in triplicates and are listed (Mer = meristemoid, GMC = guard mother cell, GC = guard cell). All the reads that aligned to the *Arabidopsis thaliana* TAIR10 genome are reported in the Aligned column. Column "q2-WithOrg" represents all aligned reads that had a mapping quality score of 2 or more, and "q2-OrgFree" represents the subset of these that did not align to the chloroplast or mitochondria. The number of THS called in each replicate, the Fraction of Reads in Peaks (FRIP), and the average THS size in base pairs of each replicate are reported. **(B)** Principal component analysis for all the ATAC-seq samples sequenced in this study, as well as the ATAC-seq samples for stem cell and mesophyll cells

previously reported (The Plant Journal). **(C)** Fragment sizes for meristemoid, guard mother cell, and guard cell aligned, quality filtered reads. Dotted line indicates the trendline.

Fig S6.2

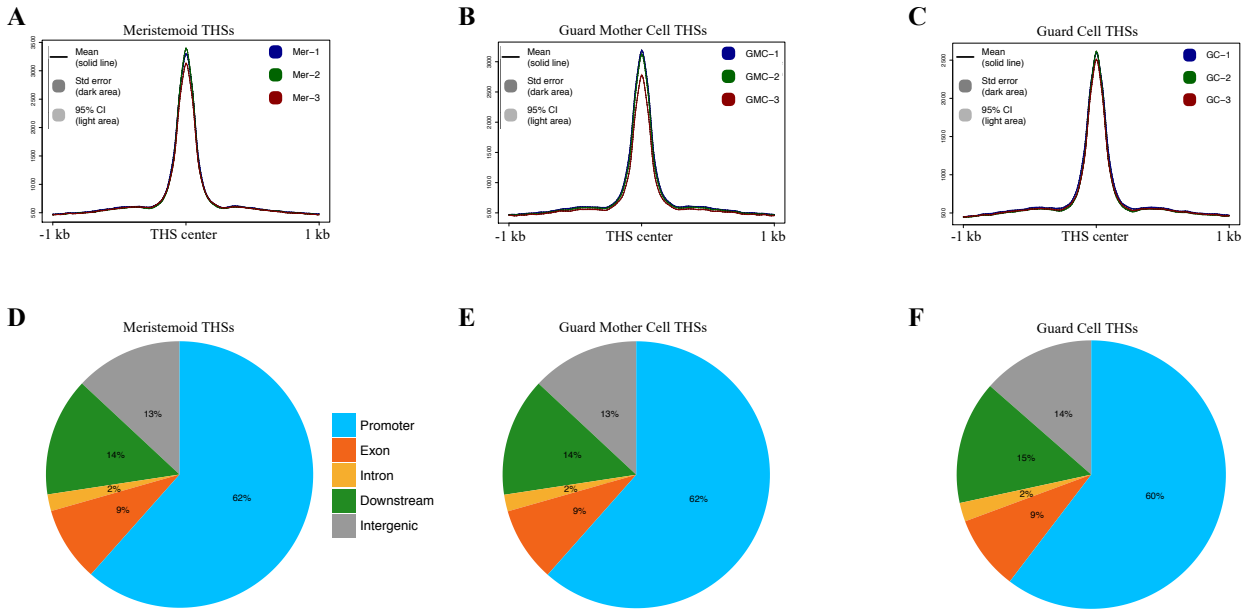


Figure S6.2 Accessibility profiles and genomic distribution of THSs reproducibly identified in each cell type. (A) Metaplot of chromatin accessibility in the three meristemoid ATAC-seq replicates at the 53,426 THSs reproducibly identified in the meristemoid samples. **(B)** Metaplot of chromatin accessibility in the three guard mother cell ATAC-seq replicates at the 53,474 THSs reproducibly identified in guard mother cell samples. **(C)** Metaplot of chromatin accessibility in the three guard cell ATAC-seq replicates at the 58,550 THSs reproducibly identified in the guard cell samples. All metaplots in A-C show a 2 kb window around the THS center, with standard error depicted and the 95% confidence interval depicted as darker and lighter shading, respectively, around the solid line depicting the mean. All the samples in A-C were scaled to the same number of reads prior to visualization. **(D-E)** Genomic distribution of THSs reproducibly identified in the **(D)** meristemoid, **(E)** guard mother cell, or **(F)** guard cell ATAC-seq samples. Promoter region was defined as 2 kb upstream of TSS and downstream region was defined as 1 kb downstream of TES. Genomic annotation was done using protein coding genes as genic features.

Fig S6.3

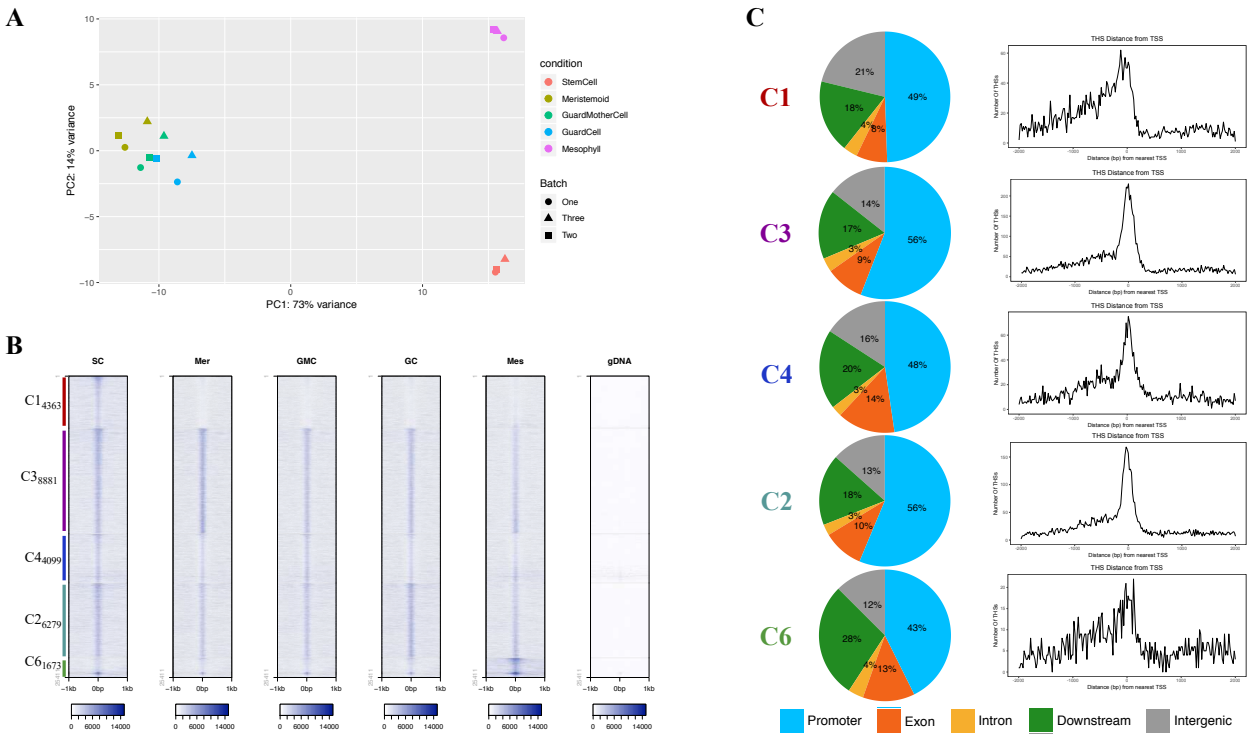


Figure S6.3 Identification and annotation of cell type-enriched THSs. (A) Principal component analysis of chromatin accessibility at the 44,313 THSs compared between the triplicates of the five different cell types. **(B)** Normalized read signal in stem cell (SC), meristemoid (Mer), guard mother cell (GMC), guard cell (GC), mesophyll (Mes), and genomic DNA (gDNA) ATAC-seq samples over the five cell type-enriched clusters from figure 6.1C. The heatmaps show a 2 kb window around the THS center and are all set to the same z-scale range. **(C)** Genomic distribution (left) and distribution of distance to the nearest protein coding gene (read) for the five cell type-enriched clusters from figure 6.1C.

Fig S6.4

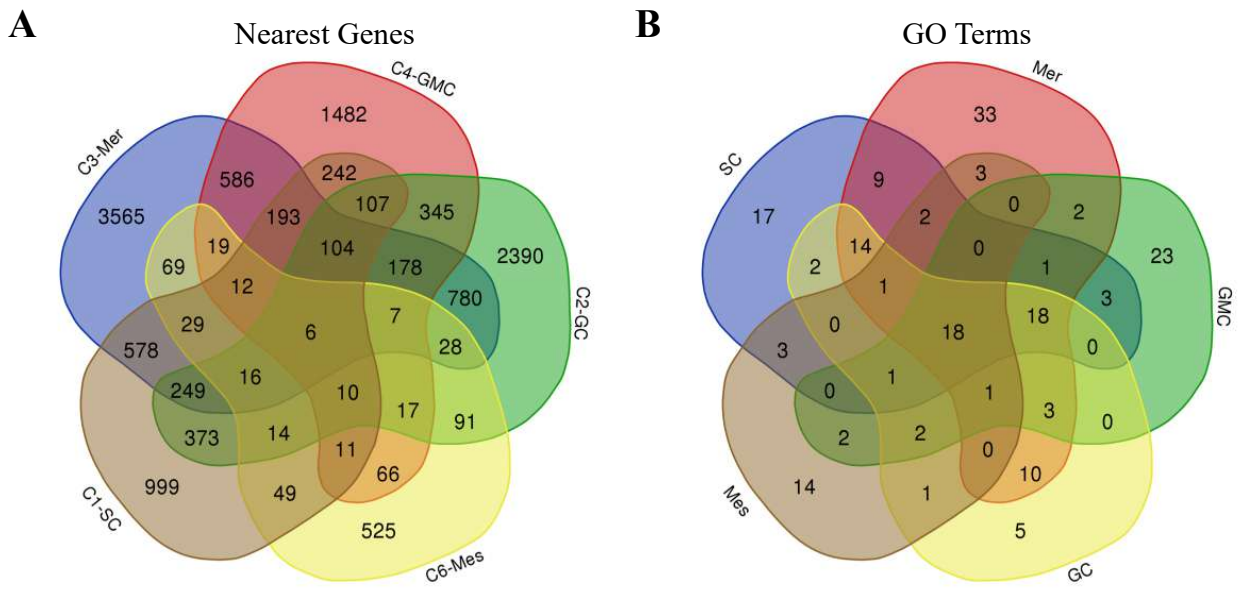


Figure S6.4 Identification and gene ontology annotation of genes nearest to cell type-enriched THS. **(A)** Each set of cell type-enriched THSs was assigned to their nearest genes. The 5-way Venn diagram shows the overlap of the nearest genes for the 5 different cell type-enriched clusters, including genes nearest to cluster 1 stem cell-enriched THSs (C1-SC, brown), cluster 3 meristemoid-enriched THSs (C3-Mer, blue), cluster 4 guard mother cell-enriched THSs (C4-GMC, red), cluster 2 guard cell-enriched THSs (C2-GC, green), or cluster 6 mesophyll-enriched THSs (C6-Mes, yellow). **(B)** Genes nearest to a cell type-enriched set of THSs that did not overlap with any other genes nearest to a different set of cell type-enriched THSs (SC = stem cell, Mer = meristemoid, GMC = guard mother cell, GC = guard cell, Mes = mesophyll) were analyzed for gene ontology (GO) enrichment and the Venn diagram shows an overlap of GO biological process terms that had a false discovery value of 0.05 or less.

Fig S6.5

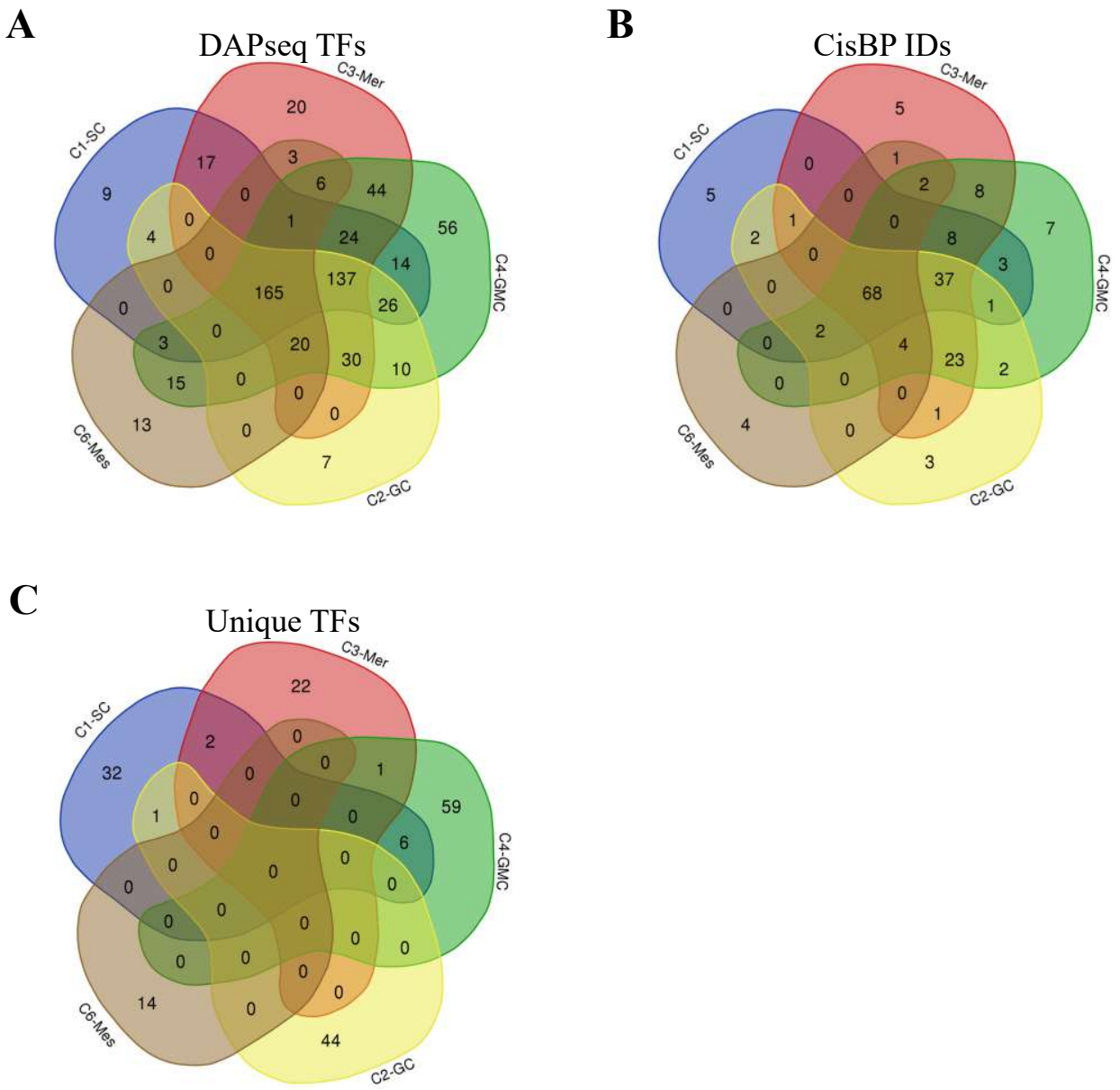
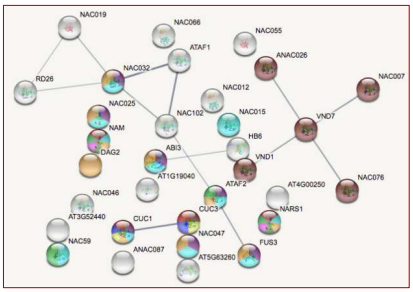


Figure S6.5 Identification of cell type-enriched motifs only found in one cell type. Venn diagram of the overlap of **(A)** Transcription Factors (TFs) from the DAPseq database or **(B)** CisBP IDs from the CisBP database that bind motifs that had an E-value of 0.05 or less in the sequences of the five different cell type-enriched groups of THSs (C1-SC = cluster 1, stem cell, blue; C3-Mer = cluster 3, meristemoid, red; C4-GMC = cluster 4, guard mother cell, green; C2-GC = cluster 2, guard cell, yellow; C6-Mes =

cluster 6, mesophyll, brown). **(C)** Venn diagram of the overlap of gene IDs that correspond to TFs from DAP-seq or CisBP IDs that were only observed in only one cell type.

Fig S6.6

A



B

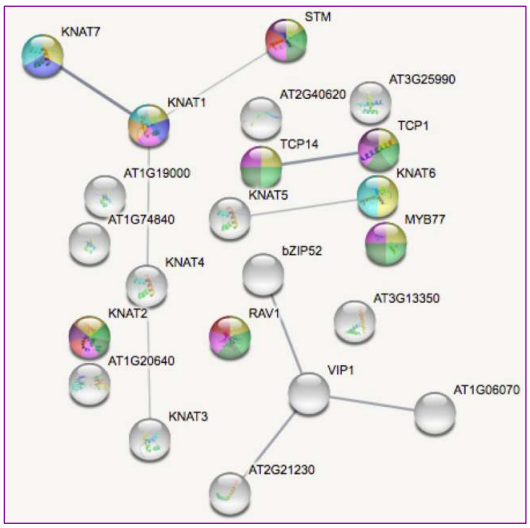
Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:0006351	transcription, DNA-templated	30 of 1957	8.34e-30
GO:0006355	regulation of transcription, DNA-templated	30 of 2167	4.27e-29
GO:0051252	regulation of RNA metabolic process	30 of 2209	4.31e-29
GO:0016070	RNA metabolic process	30 of 3007	1.03e-25
GO:0048522	positive regulation of cellular process	16 of 772	6.57e-16
GO:0048518	positive regulation of biological process	17 of 984	9.45e-16
GO:0032501	multicellular organismal process	19 of 2336	4.74e-12
GO:0031325	positive regulation of cellular metabolic process	12 of 562	7.56e-12
GO:0031328	positive regulation of cellular biosynthetic process	11 of 433	1.28e-11
GO:0007275	multicellular organism development	18 of 2127	1.35e-11
GO:0048759	xylem vessel member cell differentiation	5 of 8	1.47e-11
GO:0051173	positive regulation of nitrogen compound metabolic process	10 of 526	2.08e-09
GO:1901348	positive regulation of secondary cell wall biogenesis	4 of 7	3.14e-09
GO:0009888	tissue development	9 of 481	2.00e-08
GO:0048316	seed development	9 of 485	2.12e-08
GO:0045893	positive regulation of transcription, DNA-templated	8 of 368	5.21e-08
GO:0051254	positive regulation of RNA metabolic process	8 of 398	8.93e-08
GO:1901700	response to oxygen-containing compound	12 of 1398	1.43e-07
GO:0009793	embryo development ending in seed dormancy	7 of 336	5.67e-07
GO:0048731	system development	11 of 1474	2.36e-06
GO:0009791	post-embryonic development	10 of 1164	2.49e-06
GO:0071554	cell wall organization or biogenesis	8 of 639	2.66e-06
GO:2000026	regulation of multicellular organismal development	7 of 433	2.79e-06
GO:1900057	positive regulation of leaf senescence	3 of 13	3.56e-06
GO:0022414	reproductive process	10 of 1348	8.53e-06
GO:0001101	response to acid chemical	9 of 1058	1.01e-05
GO:0062013	positive regulation of small molecule metabolic process	3 of 21	1.17e-05
GO:0042221	response to chemical	13 of 2654	1.59e-05
GO:0051240	positive regulation of multicellular organismal process	4 of 103	2.80e-05
GO:0051094	positive regulation of developmental process	4 of 113	3.91e-05
GO:0009834	plant-type secondary cell wall biogenesis	3 of 40	6.48e-05
GO:0046885	regulation of hormone biosynthetic process	3 of 41	6.88e-05
GO:0009415	response to water	5 of 298	9.24e-05
GO:0048317	seed morphogenesis	2 of 6	0.00013
GO:0009611	response to wounding	4 of 177	0.00020
GO:0010565	regulation of cellular ketone metabolic process	3 of 66	0.00025
GO:0080060	integument development	2 of 10	0.00030
GO:0010371	regulation of gibberellin biosynthetic process	2 of 10	0.00030
GO:0046886	positive regulation of hormone biosynthetic process	2 of 11	0.00035
GO:0045995	regulation of embryonic development	2 of 13	0.00046
GO:0045928	positive regulation of isoprenoid metabolic process	2 of 13	0.00046
GO:0010033	response to organic substance	9 of 1786	0.00050
GO:0090691	formation of plant organ boundary	2 of 15	0.00058
GO:0050896	response to stimulus	15 of 5064	0.00074
GO:0071555	cell wall organization	5 of 503	0.00090
GO:0010035	response to inorganic substance	6 of 795	0.00090
GO:0046889	positive regulation of lipid biosynthetic process	2 of 20	0.00094
GO:0099402	plant organ development	6 of 812	0.00099
GO:0006950	response to stress	11 of 2932	0.00099
GO:0071840	cellular component organization or biogenesis	10 of 2467	0.0011
GO:0048523	negative regulation of cellular process	5 of 561	0.0014
GO:0048519	negative regulation of biological process	6 of 872	0.0014
GO:1901701	cellular response to oxygen-containing compound	5 of 596	0.0018
GO:0010014	meristem initiation	2 of 32	0.0021
GO:0048827	phyllome development	4 of 405	0.0035
GO:0048367	shoot system development	5 of 700	0.0035
GO:0009788	negative regulation of abscisic acid-activated signaling pat...	2 of 43	0.0035
GO:0009753	response to jasmonic acid	3 of 192	0.0040
GO:0009725	response to hormone	7 of 1502	0.0040
GO:0071229	cellular response to acid chemical	4 of 453	0.0048
GO:0070887	cellular response to chemical stimulus	6 of 1245	0.0071
GO:0009628	response to abiotic stimulus	7 of 1699	0.0073
GO:0010150	leaf senescence	2 of 85	0.0111
GO:0009414	response to water deprivation	3 of 292	0.0115
GO:0071310	cellular response to organic substance	5 of 973	0.0122
GO:0009733	response to auxin	3 of 327	0.0149
GO:0065008	regulation of biological quality	5 of 1080	0.0176
GO:0009653	anatomical structure morphogenesis	4 of 699	0.0192
GO:0009743	response to carbohydrate	2 of 123	0.0205
GO:0009908	flower development	3 of 404	0.0255
GO:0016043	cellular component organization	7 of 2271	0.0308
GO:0032870	cellular response to hormone stimulus	4 of 857	0.0361
GO:0009737	response to abscisic acid	3 of 511	0.0447
GO:0042493	response to drug	3 of 533	0.0491

Figure S6.6 STRING network and GO terms of TFs that bind motifs only identified in stem cell-enriched THS sequences. (A) STRING network of the 32 TFs that bind motifs uniquely enriched in stem cell-enriched THS sequences. **(B)** GO enrichment from STRING for the 32 TFs in A with terms indicative of stem cell highlighted and represented in A.

Fig S6.7

A



B

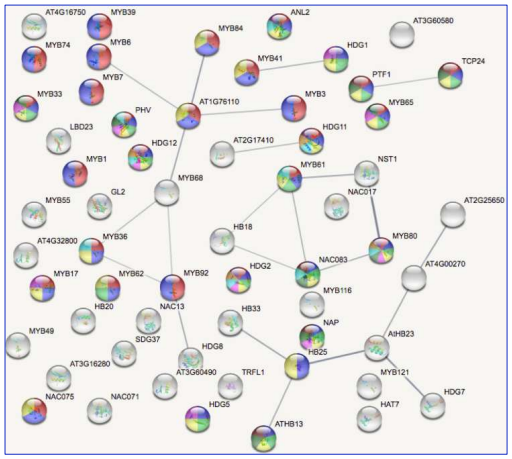
Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:0051252	regulation of RNA metabolic process	17 of 2209	8.73e-13
GO:0006355	regulation of transcription, DNA-templated	17 of 2167	8.73e-13
GO:0006351	transcription, DNA-templated	9 of 1957	0.00014
GO:0009720	detection of hormone stimulus	2 of 6	0.00019
GO:0048856	anatomical structure development	9 of 2361	0.00046
GO:0044249	cellular biosynthetic process	11 of 4013	0.00084
GO:0048367	shoot system development	5 of 700	0.0015
GO:0090696	post-embryonic plant organ development	3 of 152	0.0017
GO:0009791	post-embryonic development	6 of 1164	0.0017
GO:0090567	reproductive shoot system development	4 of 415	0.0019
GO:0032501	multicellular organismal process	8 of 2336	0.0019
GO:0016070	RNA metabolic process	9 of 3007	0.0019
GO:0010089	xylem development	2 of 36	0.0025
GO:0046483	heterocycle metabolic process	10 of 3980	0.0029
GO:0006725	cellular aromatic compound metabolic process	10 of 4074	0.0034
GO:0009735	response to cytokinin	3 of 212	0.0035
GO:0048731	system development	6 of 1474	0.0045
GO:0009725	response to hormone	6 of 1502	0.0048
GO:0007275	multicellular organism development	7 of 2127	0.0051
GO:0048440	carpel development	2 of 64	0.0061
GO:0050793	regulation of developmental process	4 of 622	0.0067
GO:0009723	response to ethylene	3 of 282	0.0067
GO:0048527	lateral root development	2 of 99	0.0121
GO:0099402	plant organ development	4 of 812	0.0145
GO:0048827	phyllome development	3 of 405	0.0156
GO:0009908	flower development	3 of 404	0.0156
GO:0051093	negative regulation of developmental process	2 of 119	0.0157
GO:0009987	cellular process	15 of 10581	0.0173
GO:0003002	regionalization	2 of 130	0.0178
GO:0009888	tissue development	3 of 481	0.0231
GO:0071310	cellular response to organic substance	4 of 973	0.0237
GO:0045892	negative regulation of transcription, DNA-templated	2 of 172	0.0276
GO:0048523	negative regulation of cellular process	3 of 561	0.0319
GO:0051253	negative regulation of RNA metabolic process	2 of 198	0.0343
GO:0071369	cellular response to ethylene stimulus	2 of 204	0.0354

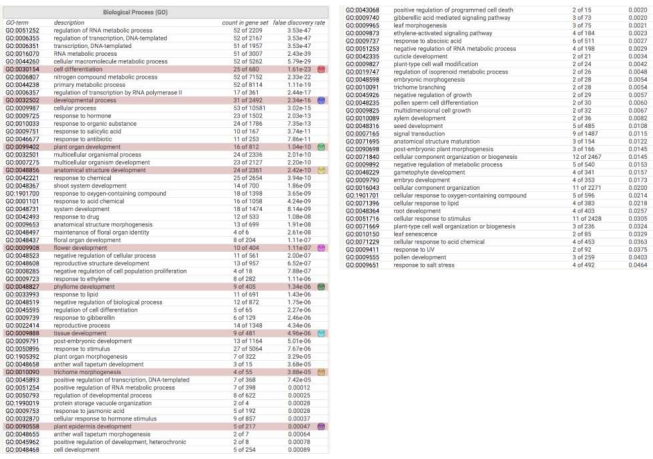
Figure S6.7 STRING network and GO terms of TFs that bind motifs only identified in meristemoid-enriched THS sequences. (A) STRING network of the 22 TFs that bind motifs uniquely enriched in meristemoid-enriched THS sequences. **(B)** GO enrichment from STRING for the 22 TFs in A with terms indicative of meristemoid highlighted and represented in A.

Fig S6.8

A



B



Biological Process (GO)

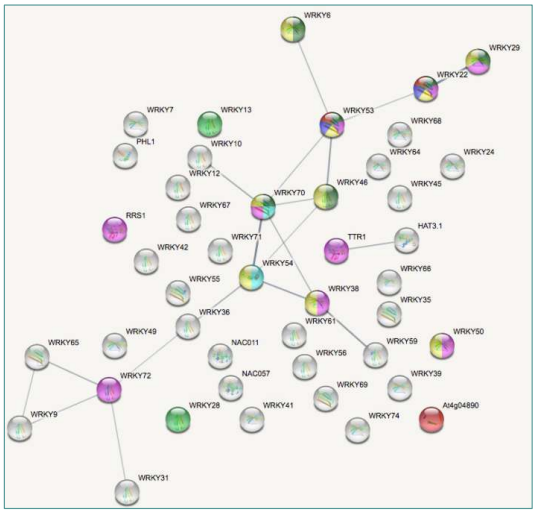
GO-term	description	count in gene set	false discovery rate
GO:0051252	regulation of RNA metabolic process	52 of 2209	3.53e-47
GO:0006355	regulation of transcription, DNA-templated	52 of 2167	3.53e-47
GO:0006351	transcription, DNA-templated	51 of 1957	3.53e-47
GO:0016070	RNA metabolic process	51 of 3007	2.43e-39
GO:0044260	cellular macromolecule metabolic process	52 of 5262	5.79e-29
GO:0030154	cell differentiation	25 of 680	1.61e-23
GO:0006807	nitrogen compound metabolic process	52 of 7152	2.33e-22
GO:0044238	primary metabolic process	52 of 8114	1.11e-19
GO:0006357	regulation of transcription by RNA polymerase II	17 of 361	2.44e-17
GO:0032502	developmental process	31 of 2493	2.34e-16
GO:0009987	cellular process	53 of 10581	3.02e-15
GO:0009725	response to hormone	23 of 1502	2.03e-13
GO:0010033	response to organic substance	24 of 1786	7.35e-13
GO:0009751	response to salicylic acid	10 of 167	3.74e-11
GO:0046677	response to antibiotic	11 of 253	7.86e-11
GO:0099402	plant organ development	16 of 812	1.04e-10
GO:0032501	multicellular organismal process	24 of 2336	2.01e-10
GO:0007275	multicellular organism development	23 of 2127	2.20e-10
GO:0048856	anatomical structure development	24 of 2361	2.42e-10
GO:0042221	response to chemical	25 of 2654	3.94e-10
GO:0048367	shoot system development	14 of 700	1.86e-09
GO:1901700	response to oxygen-containing compound	18 of 1398	3.65e-09
GO:0071101	response to acid chemical	16 of 1058	4.24e-09
GO:0048731	system development	18 of 1474	8.14e-09
GO:0042493	response to drug	12 of 533	1.08e-08
GO:0009653	anatomical structure morphogenesis	13 of 699	1.91e-08
GO:0048497	maintenance of floral organ identity	4 of 6	2.61e-08
GO:0048437	floral organ development	9 of 204	1.11e-07
GO:0009908	flower development	10 of 404	1.11e-07
GO:0048523	negative regulation of cellular process	11 of 561	2.00e-07
GO:0048608	reproductive structure development	13 of 957	6.52e-07
GO:0008285	negative regulation of cell population proliferation	4 of 18	7.88e-07
GO:0009723	response to ethylene	8 of 282	1.11e-06
GO:0048827	phyllome development	9 of 405	1.34e-06
GO:0033993	response to lipid	11 of 691	1.43e-06
GO:0048519	negative regulation of biological process	12 of 872	1.75e-06
GO:0045595	regulation of cell differentiation	5 of 65	2.27e-06
GO:0009739	response to gibberellin	5 of 129	2.46e-06
GO:0022414	reproductive process	14 of 1348	4.34e-06
GO:0009888	tissue development	9 of 481	4.96e-06
GO:0009791	post-embryonic development	13 of 1164	5.01e-06
GO:0009896	response to stimulus	27 of 5064	7.67e-06
GO:1905392	plant organ morphogenesis	7 of 322	3.29e-05
GO:0048658	anther wall tapetum development	3 of 15	3.68e-05
GO:0010090	trichome morphogenesis	4 of 55	3.88e-05
GO:0045893	positive regulation of transcription, DNA-templated	7 of 368	7.42e-05
GO:0051254	positive regulation of RNA metabolic process	7 of 398	0.00012
GO:0090793	regulation of developmental process	8 of 622	0.00025
GO:1990019	protein storage vacuole organization	2 of 4	0.00028
GO:0009750	response to jasmonic acid	5 of 192	0.00028
GO:0032870	cellular response to hormone stimulus	9 of 857	0.00037
GO:0090558	plant epidermis development	5 of 217	0.00047
GO:0048655	anther wall tapetum morphogenesis	2 of 7	0.00064
GO:0045962	positive regulation of development, heterochronic	2 of 8	0.00078
GO:0048468	cell development	5 of 254	0.00089

GO:0043068	positive regulation of programmed cell death	2 of 15	0.0020
GO:0009740	gibberellic acid mediated signaling pathway	3 of 73	0.0020
GO:0009965	leaf morphogenesis	3 of 75	0.0021
GO:0009873	ethylene-activated signaling pathway	4 of 184	0.0023
GO:0009737	response to abscisic acid	6 of 511	0.0027
GO:0051253	negative regulation of RNA metabolic process	4 of 198	0.0029
GO:0042335	cuticle development	2 of 21	0.0034
GO:0009827	plant-type cell wall modification	2 of 24	0.0042
GO:0019747	regulation of isoprenoid metabolic process	2 of 26	0.0048
GO:0048598	embryonic morphogenesis	2 of 28	0.0054
GO:0010091	trichome branching	2 of 28	0.0054
GO:0045926	negative regulation of growth	2 of 29	0.0057
GO:0048235	pollen sperm cell differentiation	2 of 30	0.0060
GO:0009825	multidimensional cell growth	2 of 32	0.0067
GO:0010089	xylem development	2 of 36	0.0082
GO:0048316	seed development	5 of 485	0.0108
GO:0007165	signal transduction	9 of 1487	0.0115
GO:0071695	anatomical structure maturation	3 of 154	0.0122
GO:0090698	post-embryonic plant morphogenesis	3 of 166	0.0145
GO:0071840	cellular component organization or biogenesis	12 of 2467	0.0145
GO:0009892	negative regulation of metabolic process	5 of 540	0.0153
GO:0048229	gametophyte development	4 of 341	0.0157
GO:0009790	embryo development	4 of 353	0.0173
GO:0016043	cellular component organization	11 of 2271	0.0220
GO:1901701	cellular response to oxygen-containing compound	5 of 596	0.0274
GO:0071396	cellular response to lipid	4 of 383	0.0218
GO:0048364	root development	4 of 403	0.0257
GO:0051716	cellular response to stimulus	11 of 2428	0.0305
GO:0071669	plant-type cell wall organization or biogenesis	3 of 236	0.0324
GO:0010150	leaf senescence	2 of 85	0.0329
GO:0071229	cellular response to acid chemical	4 of 453	0.0363
GO:0009411	response to UV	2 of 92	0.0375
GO:0009555	pollen development	3 of 259	0.0403
GO:0009631	response to salt stress	4 of 492	0.0464

Figure S6.8 STRING network and GO terms of TFs that bind motifs only identified in guard mother cell-enriched THS sequences. (A) STRING network of the 59 TFs that bind motifs uniquely enriched in guard mother cell-enriched THS sequences. **(B)** GO enrichment from STRING for the 59 TFs in A with terms indicative of guard mother cell highlighted and represented in A.

Fig S6.9

A



B

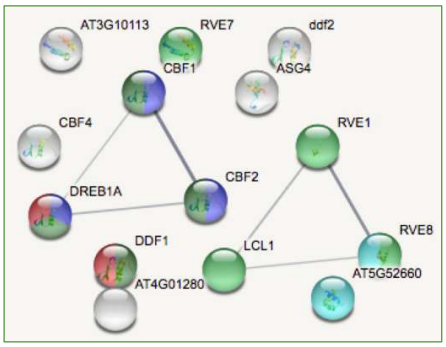
Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:0006351	transcription, DNA-templated	42 of 1957	2.10e-43
GO:0006355	regulation of transcription, DNA-templated	42 of 2167	3.57e-42
GO:0051252	regulation of RNA metabolic process	42 of 2209	4.52e-42
GO:0016070	RNA metabolic process	42 of 3007	4.17e-37
GO:0010200	response to chitin	6 of 113	1.77e-07
GO:0042493	response to drug	8 of 533	9.67e-06
GO:1900056	negative regulation of leaf senescence	2 of 10	0.00074
GO:0009751	response to salicylic acid	4 of 167	0.00074
GO:0006952	defense response	9 of 1277	0.00074
GO:1901700	response to oxygen-containing compound	9 of 1398	0.0014
GO:0031347	regulation of defense response	4 of 224	0.0020
GO:0009863	salicylic acid mediated signaling pathway	2 of 31	0.0051
GO:0042742	defense response to bacterium	4 of 315	0.0064
GO:0010033	response to organic substance	9 of 1786	0.0070
GO:0002218	activation of innate immune response	2 of 41	0.0078
GO:0045892	negative regulation of transcription, DNA-templated	3 of 172	0.0097
GO:0045893	positive regulation of transcription, DNA-templated	4 of 368	0.0100
GO:0009867	jasmonic acid mediated signaling pathway	2 of 54	0.0117
GO:0051254	positive regulation of RNA metabolic process	4 of 398	0.0120
GO:0051253	negative regulation of RNA metabolic process	3 of 198	0.0124
GO:0048518	positive regulation of biological process	6 of 984	0.0135
GO:0045595	regulation of cell differentiation	2 of 65	0.0138
GO:0048731	system development	7 of 1474	0.0210
GO:0010150	leaf senescence	2 of 85	0.0210
GO:0048366	leaf development	3 of 272	0.0227
GO:0098542	defense response to other organism	5 of 847	0.0258
GO:0006950	response to stress	10 of 2932	0.0339
GO:0002376	immune system process	3 of 330	0.0339
GO:0050793	regulation of developmental process	4 of 622	0.0360
GO:0009814	defense response, incompatible interaction	2 of 145	0.0461

Figure S6.9 STRING network and GO terms of TFs that bind motifs only identified in guard cell-enriched THS sequences. (A) STRING network of the 44 TFs that bind motifs uniquely enriched in guard cell-enriched THS sequences. **(B)** GO enrichment from STRING for the 44 TFs in A with terms indicative of guard cell highlighted and represented in A.

Fig S6.10

A



B

Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:0006351	transcription, DNA-templated	14 of 1957	1.80e-14
GO:0051252	regulation of RNA metabolic process	14 of 2209	1.87e-14
GO:0006355	regulation of transcription, DNA-templated	14 of 2167	1.87e-14
GO:0016070	RNA metabolic process	14 of 3007	2.68e-13
GO:0009628	response to abiotic stimulus	10 of 1699	3.20e-09
GO:0010035	response to inorganic substance	7 of 795	2.18e-07
GO:0009651	response to salt stress	6 of 492	3.77e-07
GO:0050896	response to stimulus	12 of 5064	3.92e-07
GO:0007623	circadian rhythm	4 of 87	4.25e-07
GO:0006950	response to stress	10 of 2932	4.98e-07
GO:0009723	response to ethylene	5 of 282	8.13e-07
GO:0046686	response to cadmium ion	5 of 286	8.55e-07
GO:0001101	response to acid chemical	7 of 1058	1.21e-06
GO:0009739	response to gibberellin	4 of 129	1.69e-06
GO:0042221	response to chemical	9 of 2654	3.13e-06
GO:0046885	regulation of hormone biosynthetic process	3 of 41	4.43e-06
GO:0009751	response to salicylic acid	4 of 167	4.43e-06
GO:0009631	cold acclimation	3 of 43	4.83e-06
GO:1901700	response to oxygen-containing compound	7 of 1398	6.73e-06
GO:0009753	response to jasmonic acid	4 of 192	6.96e-06
GO:0009725	response to hormone	7 of 1502	1.04e-05
GO:0009737	response to abscisic acid	5 of 511	1.12e-05
GO:0010371	regulation of gibberellin biosynthetic process	2 of 10	4.13e-05
GO:0009733	response to auxin	4 of 327	4.79e-05
GO:0009409	response to cold	4 of 347	5.87e-05
GO:0019747	regulation of isoprenoid metabolic process	2 of 26	0.00022
GO:0048510	regulation of timing of transition from vegetative to reprodu...	2 of 37	0.00042
GO:0042752	regulation of circadian rhythm	2 of 47	0.00063
GO:0019760	glucosinolate metabolic process	2 of 97	0.0024
GO:0016143	S-glycoside metabolic process	2 of 97	0.0024
GO:0009414	response to water deprivation	2 of 292	0.0189
GO:0045893	positive regulation of transcription, DNA-templated	2 of 368	0.0282
GO:0051254	positive regulation of RNA metabolic process	2 of 398	0.0314

Figure S6.10 STRING network and GO terms of TFs that bind motifs only identified in mesophyll-enriched THS sequences. (A) STRING network of the 14 TFs that bind motifs uniquely enriched in mesophyll-enriched THS sequences. **(B)** GO enrichment from STRING for the 14 TFs in A with terms indicative of mesophyll highlighted and represented in A.

Fig S6.11

A

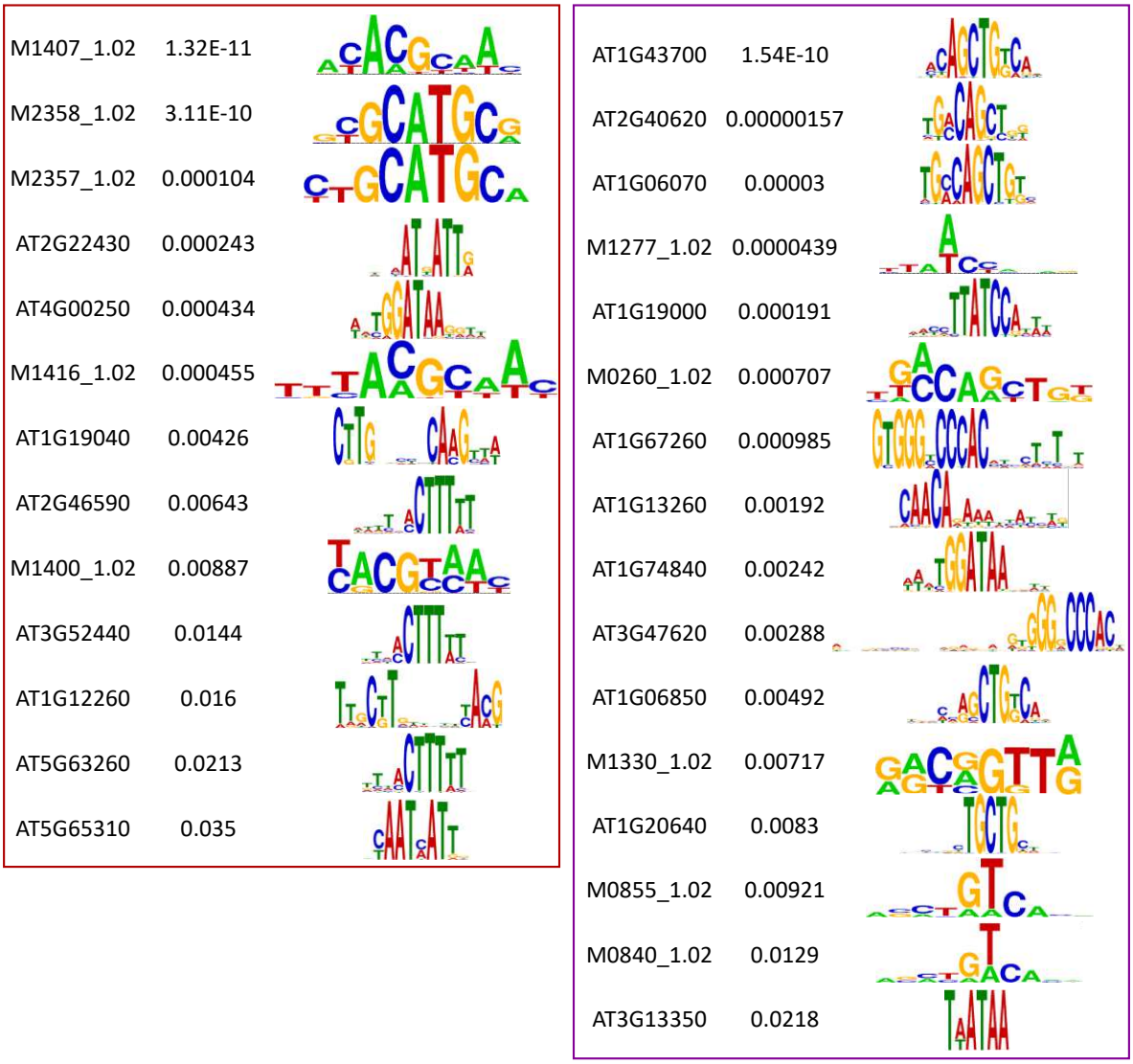


Figure S6.11 Motifs uniquely enriched in stem cell and meristemoid THS sequences. (A) Table of motif sequence binding TFs and CisBP IDs, their E-value, and positional weight matrices for motifs uniquely identified in stem cell (left, red) or meristemoid (right, purple) THS sequences.

Fig S6.12

A

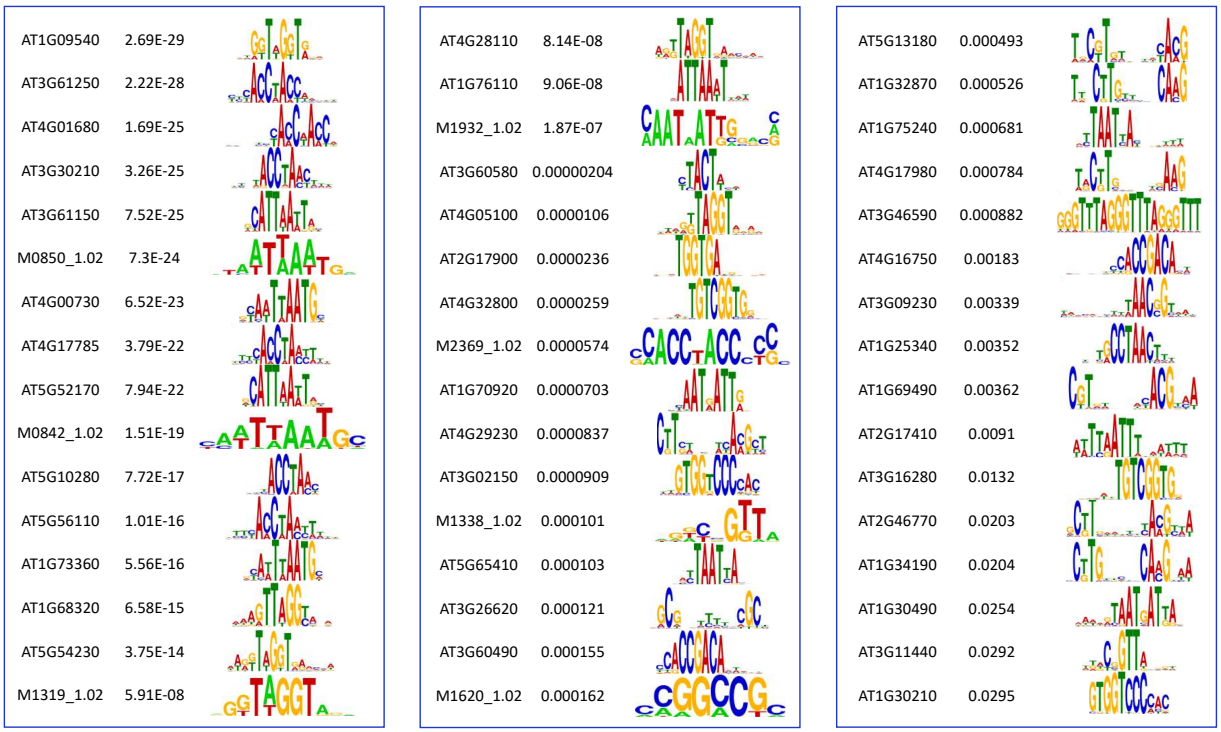


Figure S6.12 Motifs uniquely enriched in guard mother cell THS sequences. (A) Table of motif sequence binding TFs and CisBP IDs, their E-value, and positional weight matrices for motifs uniquely identified in guard mother cell THS sequences.

Fig S6.13

A

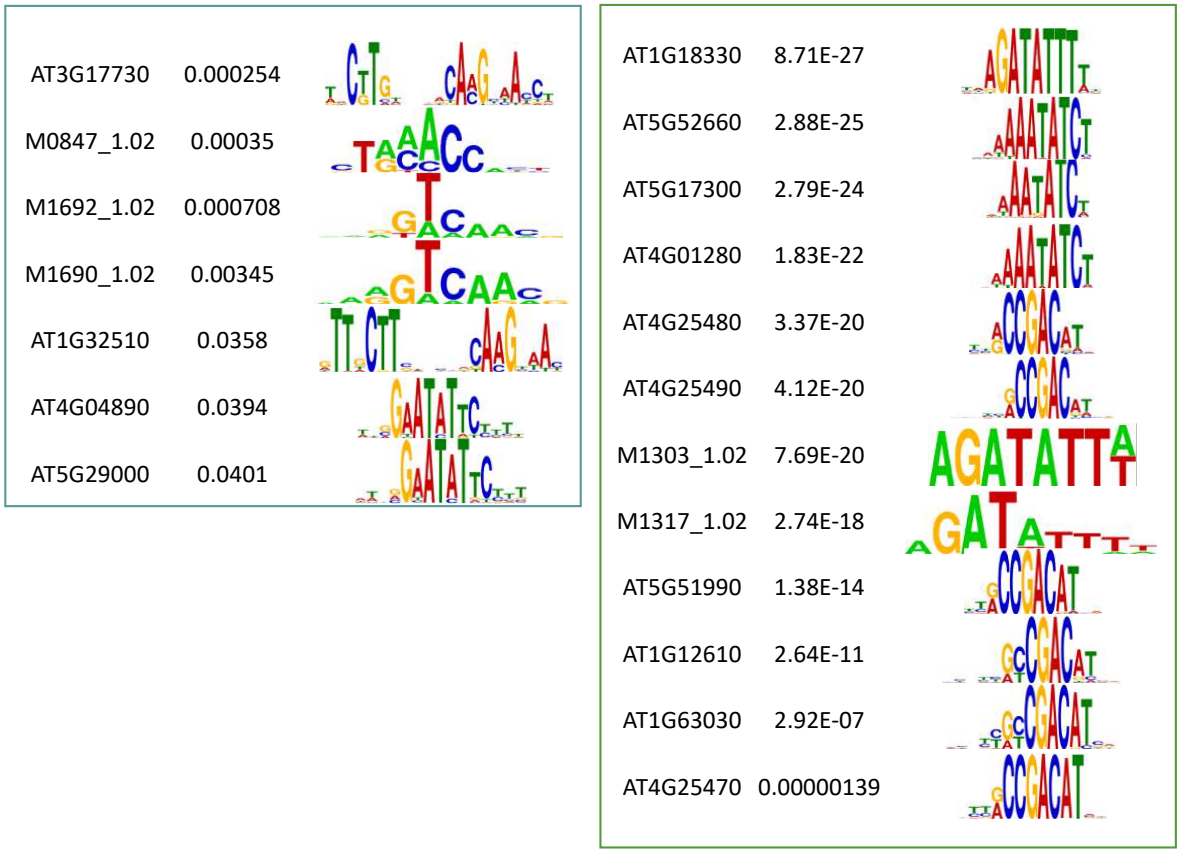


Figure S6.13 Motifs uniquely enriched in guard cell and mesophyll THS sequences. (A) Table of motif sequence binding TFs and CisBP IDs, their E-value, and positional weight matrices for motifs uniquely identified in guard cell (left, teal) or mesophyll (right, green) THS sequences.

Table 6.1

Sample	Primer Used	Primer Sequence
All samples	ATAC-Primer 1	AATGATACGGCGACCACCGAGATCTACACTCGT CGGCAGCGTCAGATGTG
MUTE::NTF Rep 1	ATAC-Primer 2.2	CAAGCAGAAGACGGCATACGAGAT**CTAGTACG** GTCTCGTGGGCTCGGAGATGT
MUTE::NTF Rep 2	ATAC-Primer 2.3	CAAGCAGAAGACGGCATACGAGAT**TTCTGCCT** GTCTCGTGGGCTCGGAGATGT
MUTE::NTF Rep 3	ATAC-Primer 2.4	CAAGCAGAAGACGGCATACGAGAT**GCTCAGGA** GTCTCGTGGGCTCGGAGATGT
FAMA::NTF Rep 1	ATAC-Primer 2.5	CAAGCAGAAGACGGCATACGAGAT**AGGAGTCC** GTCTCGTGGGCTCGGAGATGT
FAMA::NTF Rep 2	ATAC-Primer 2.6	CAAGCAGAAGACGGCATACGAGAT**CATGCCTA** GTCTCGTGGGCTCGGAGATGT
FAMA::NTF Rep 3	ATAC-Primer 2.7	CAAGCAGAAGACGGCATACGAGAT**GTAGAGAG** GTCTCGTGGGCTCGGAGATGT
MYB60::NTF Rep 1	ATAC-Primer 2.8	CAAGCAGAAGACGGCATACGAGAT**CCTCTCTG** GTCTCGTGGGCTCGGAGATGT
MYB60::NTF Rep 2	ATAC-Primer 2.9	CAAGCAGAAGACGGCATACGAGAT**AGCGTAGC** GTCTCGTGGGCTCGGAGATGT
MYB60::NTF Rep 3	ATAC-Primer 2.10	CAAGCAGAAGACGGCATACGAGAT**CAGCCTCG** GTCTCGTGGGCTCGGAGATGT

Table 6.1 Primers used for ATAC-seq library preparation. The unique index specific to each primer is underlined.

LITERATURE CITED

Adrian, J., J. Chang, C. E. Ballenger, B. O. Bargmann, J. Alassimone, K. A. Davies, O. S. Lau, J. L. Matos, C. Hachez, A. Lanctot, A. Vaten, K. D. Birnbaum and D. C. Bergmann (2015). "Transcriptome dynamics of the stomatal lineage: birth, amplification, and termination of a self-renewing population." Dev Cell **33**(1): 107-118.

Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Bajic, M., K. A. Maher and R. B. Deal (2018). "Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq." Methods Mol Biol **1675**: 183-201.

Barton, M. K. (2010). "Twenty years on: the inner workings of the shoot apical meristem, a developmental dynamo." Dev Biol **341**(1): 95-113.

Bergmann, D. C. and F. D. Sack (2007). "Stomatal development." Annu Rev Plant Biol **58**: 163-181.

Besnard, F., T. Vernoux and O. Hamant (2011). "Organogenesis from stem cells in planta: multiple feedback loops integrating molecular and mechanical signals." Cell Mol Life Sci **68**(17): 2885-2906.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang and W. J. Greenleaf (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nat Methods **10**(12): 1213-1218.

Cominelli, E., M. Galbiati, A. Vavasseur, L. Conti, T. Sala, M. Vuylsteke, N. Leonhardt, S. L. Dellaporta and C. Tonelli (2005). "A guard-cell-specific MYB transcription factor regulates stomatal movements and plant drought tolerance." Curr Biol **15**(13): 1196-1200.

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

Deal, R. B. and S. Henikoff (2011). "The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana." Nat Protoc **6**(1): 56-68.

Dong, J., C. A. MacAlister and D. C. Bergmann (2009). "BASL controls asymmetric cell division in Arabidopsis." Cell **137**(7): 1320-1330.

Du, Z., X. Zhou, Y. Ling, Z. Zhang and Z. Su (2010). "agriGO: a GO analysis toolkit for the agricultural community." Nucleic Acids Res **38**(Web Server issue): W64-70.

Endo, M., H. Shimizu, M. A. Nohales, T. Araki and S. A. Kay (2014). "Tissue-specific clocks in Arabidopsis show asymmetric coupling." Nature **515**(7527): 419-422.

Frerichs, A., J. Engelhorn, J. Altmuller, J. Gutierrez-Marcos and W. Werr (2019). "Specific chromatin changes mark lateral organ founder cells in the Arabidopsis thaliana inflorescence meristem." J Exp Bot.

Geisler, M., J. Nadeau and F. D. Sack (2000). "Oriented Asymmetric Divisions That Generate the Stomatal Spacing Pattern in Arabidopsis Are Disrupted by the *too many mouths* Mutation." The Plant Cell **12**: 2075-2086.

Grant, C. E., T. L. Bailey and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Han, S. K., X. Qi, K. Sugihara, J. H. Dang, T. A. Endo, K. L. Miller, E. D. Kim, T. Miura and K. U. Torii (2018). "MUTE Directly Orchestrates Cell-State Switch and the Single Symmetric Division to Create Stomata." Dev Cell **45**(3): 303-315 e305.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Mol Cell **38**(4): 576-589.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Lee, L. R. and D. C. Bergmann (2019). "The plant stomatal lineage at a glance." J Cell Sci **132**(8).

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Lu, Z., B. T. Hofmeister, C. Vollmers, R. M. DuBois and R. J. Schmitz (2017). "Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes." Nucleic Acids Res **45**(6): e41.

Maher, K. A., M. Bajic, K. Kajala, M. Reynoso, G. Pauluzzi, D. A. West, K. Zumstein, M. Woodhouse, K. Bubb, M. W. Dorrity, C. Queitsch, J. Bailey-Serres, N. Sinha, S. M. Brady and R. B. Deal (2018). "Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules." Plant Cell **30**(1): 15-36.

McLeay, R. C. and T. L. Bailey (2010). "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data." Bioinformatics **11**(165).

Nadeau, J. A. and F. D. Sack (2003). "Stomatal development: cross talk puts mouths in place." Trends in Plant Science **8**(6): 294-299.

O'Malley, R. C., S. C. Huang, L. Song, M. G. Lewsey, A. Bartlett, J. R. Nery, M. Galli, A. Gallavotti and J. R. Ecker (2016). "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape." Cell **165**(5): 1280-1292.

Ohashi-Ito, K. and D. C. Bergmann (2006). "Arabidopsis FAMA controls the final proliferation/differentiation switch during stomatal development." Plant Cell **18**(10): 2493-2505.

Pillitteri, L. J., N. L. Bogenschutz and K. U. Torii (2008). "The bHLH protein, MUTE, controls differentiation of stomata and the hydathode pore in Arabidopsis." Plant Cell Physiol **49**(6): 934-943.

Pillitteri, L. J. and J. Dong (2013). "Stomatal development in Arabidopsis." Arabidopsis Book **11**: e0162.

Pillitteri, L. J., D. B. Sloan, N. L. Bogenschutz and K. U. Torii (2007). "Termination of asymmetric cell division and differentiation of stomata." Nature **445**(7127): 501-505.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Ramirez, F., F. Dundar, S. Diehl, B. A. Gruning and T. Manke (2014). "deepTools: a flexible platform for exploring deep-sequencing data." Nucleic Acids Res **42**(Web Server issue): W187-191.

Salmon-Divon, M., H. Dvinge, K. Tammoja and P. Bertone (2010). "PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci." BMC Bioinformatics **11**: 415.

Sijacic, P., M. Bajic, E. C. McKinney, R. B. Meagher and R. B. Deal (2018). "Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks." Plant J **94**(2): 215-231.

Stempor, P. and J. Ahringer (2016). "SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics." Wellcome Open Res **1**: 14.

Supek, F., M. Bosnjak, N. Skunca and T. Smuc (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms." PLoS One **6**(7): e21800.

Szklarczyk, D., J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen and C. von Mering (2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." Nucleic Acids Res **45**(D1): D362-D368.

Thorvaldsdottir, H., J. T. Robinson and J. P. Mesirov (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Brief Bioinform **14**(2): 178-192.

Tian, T., Y. Liu, H. Yan, Q. You, X. Yi, Z. Du, W. Xu and Z. Su (2017). "agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update." Nucleic Acids Res **45**(W1): W122-W129.

Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J. C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F. Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker and T. R. Hughes (2014). "Determination and inference of eukaryotic transcription factor sequence specificity." Cell **158**(6): 1431-1443.

Wickham, H. (2009). "ggplot2: Elegant Graphics for Data Analysis." Journal of Statistical Software **35**.

You, Y., A. Sawikowska, M. Neumann, D. Pose, G. Capovilla, T. Langenecker, R. A. Neher, P. Krajewski and M. Schmid (2017). "Temporal dynamics of gene expression and histone marks at the Arabidopsis shoot meristem during flowering." Nat Commun **8**: 15120.

Zhao, L. and F. D. Sack (1998). "Ultrastructure of Stomatal Development in Arabidopsis (Brassicaceae) Leaves." American Journal of Botany **86**(7).

CHAPTER 7: DISCUSSION – IMPLICATIONS AND FUTURE DIRECTIONS

The goal of my dissertation research was to understand the interplay between transcription factors and gene expression that leads to multitudes of unique transcriptional regulatory networks from differential utilization of the same DNA information. Working towards this goal, I analyzed chromatin accessibility and transcriptional output in five different plant species, seven different cell types, and in response to an environmental stress. To achieve this level of comparison, I first worked with Kelsey Maher to establish the use of INTACT-ATAC-seq in *Arabidopsis thaliana* roots (Chapter 2). We then worked alongside several other researchers to analyze chromatin accessibility across four different plant species: *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, and *Solanum lycopersicum* (Chapter 3). Additionally, we combined ATAC-seq and RNA-seq analyses to build gene regulatory networks (GRNs) in the epidermal hair cells of the root (Chapter 3). To define chromatin accessibility changes between two cell types that represent the start and end points of cell differentiation, I worked with Paja Sijacic to build GRNs in stem cells of the Arabidopsis shoot apical meristem and the mesophyll cells of the leaf (Chapter 4). This work was utilized again in ongoing work that is outlined in Chapter 6, where I defined sites of chromatin accessibility in three cell states of the stomatal lineage that can be used with complementary RNA-seq experiments to build GRNs in the future. Finally, I worked with collaborators from three different labs to build upon work started in Chapter 3. In this culmination of my dissertation research, we were able to combine RNA-seq results with chromatin accessibility from the same nuclei or cells to define submergence stress-induced GRNs shared across four different species: *Medicago truncatula*, *Oryza sativa*, *Solanum lycopersicum*, and *Solanum pennellii* (Chapter 5). The sum of this work established the use of INTACT-ATAC-seq in plants, the development of the analysis pipelines for ATAC-seq data, and the incorporation of chromatin accessibility data with RNA-seq results to build GRNs in different plants, cell types, and in response to an environmental stress. In this chapter, I discuss the implications of this work, remaining questions, and future directions.

Gene Regulatory Networks connect transcription factors and regulated gene targets

Plant regulatory elements are proximal to genes

Currently, there are few chromatin accessibility profiles in plants that were not generated within this work. We generated and made publicly available 61 ATAC-seq and 645 RNA-seq datasets. In addition to providing chromatin accessibility profiles for two cell types of the *Arabidopsis* root, five cell types of the *Arabidopsis* shoot, and the root tips of four species under control and submergence stress conditions, we also sequenced ATAC-seq libraries generated using genomic DNA stripped of proteins, revealing the sequence biases of the Tn5 transposase.

For all five plant species analyzed throughout this research, chromatin accessibility was observed predominantly within the 3 kb upstream region of the transcription start site. This was a surprising discovery because the genomes of the five plant species are organized differently, particularly with respect to their intergenic distances. Arabidopsis is known to have few distal chromatin conformation interactions, and this may be true of other plants (Rowley et al. 2017). This suggests that enhancers in plants are primarily found in promoters, which is different from enhancers in animals that are found primarily in introns or distal intergenic space (Zhang et al. 2014).

Within the different ATAC-seq datasets generated, all the accessible sites identified within an organism showed some level of accessibility in the different cell states analyzed. For example, THSs identified in leaf tissue had some accessibility in roots. This observation is different from the accessibility profiles observed in animals, where accessible sites detected in one cell type can have no accessibility in another cell type (Buenrostro et al. 2018). Widespread accessibility in plants may explain how virtually any differentiated cell in a plant is able to undergo dedifferentiation to form a callus that can be differentiated into a complete plant. This can be further evaluated by performing ATAC-seq on the dedifferentiated cells of a callus to determine whether the regulatory regions identified in cell-type-specific chromatin profiles all become more accessible when the cell is not differentiated.

Further work needs to be done to confirm whether accessible sites found proximal to genes can actually function as enhancers. Initially, the list of potential sites can be evaluated computationally by

overlapping promoter-localized accessible sites with publicly available chromatin immunoprecipitation sequencing (ChIP-seq) results for mono-methylation of histone H3 lysine 4 and acetylation of histone H3 lysine 27, epigenetic marks known to be found at enhancers (Heintzman et al. 2009). Functionally, tagmented DNA can be used as the input for Self-Transcribing Active Regulatory Region sequencing (STARR-seq) to measures the enhancer activity of millions of DNA fragments in parallel (Arnold et al. 2013). Very recently, STARR-seq was used in rice to identify thousands of potential enhancers (Sun et al. 2019), but it still needs to be used in other plant species to characterize plant enhancers.

Constructing Gene Regulatory Networks

Gene regulatory networks are composed of many nodes connecting TFs and their putative target genes (Figure 7.1A). To build connections between a TF and its regulated genes it is important to determine where the TF binds DNA and which genes are being affected by this binding. Sites where a transcription factor binds DNA can be visualized experimentally using chromatin immunoprecipitation experiments, such as ChIP-seq or ChIP-qPCR. However, it is not straightforward to perform these experiments if no antibody for the TF of interest exists. As an alternative, thousands of TF binding sites can be predicted from overrepresented sequence motifs found in transposase hypersensitive sites (THSs) identified using ATAC-seq. The variability in the sequence that a TF binds is represented by a positional weight matrix (PWM) that has been experimentally determined by aggregating thousands of sites where the TF binds. The majority of PWMs are determined from in vitro experiments, which was the case for the datasets we utilized (Weirauch et al. 2014, O'Malley et al. 2016), and may not reflect the exact motif that a TF binds in vivo.

Even if specific sites where a TF binds DNA are known, it is still difficult to determine which genes are being regulated by these TF-DNA interactions. As mentioned before, our work demonstrates that the majority of the cis regulatory elements in plants are found proximal to genes, as was previously reported in Arabidopsis (Rowley et al. 2017) and rice (Sun et al. 2019). Therefore, a connection between a TF and a regulated gene can be built by identifying the nearest gene to a TF-binding sequence.

Alternatively, the same logic can be applied to build a connection between differentially regulated genes and nearby TF-binding sites (Lowe et al. 2019). Mechanistically, if a group of genes is differentially regulated between two data points, such as between two cell types during cell differentiation, then it is possible to identify overrepresented TF-biding sequence motifs that regulate these genes in plants simply by analyzing proximal chromatin accessible sites (Figure 7.1B). This is especially true if the accessibility in these proximal DNA sites also changed between the two data points.

We built several GRNs by connecting differentially regulated genes with overrepresented sequence motifs found in proximal THSs. Throughout this work, several common features were made apparent in the composition of GRNs. First, GRNs tend to have "apex" TFs that regulate the expression of other TFs, but are not themselves subject to regulation by other TFs in the network. Second, many TFs, and other non-TF target genes, are regulated by multiple TFs. Third, this implies that combinatorial regulation is common in plants despite their relatively simplified regulatory structure.

Chromatin and transcription readouts from the same starting material construct more specific GRNs

The availability of comparable, high quality RNA-seq data for use with ATAC-seq results is an important consideration for establishing accurate conclusions. Using the same starting materials for the two analyses allows for confident comparisons between chromatin and transcription. The rationale is that the biological and technical variations among replicates will be reflected in both levels of analysis, allowing for the identification of regulated connections that are replicated among the samples (Cusanovich et al. 2015).

By using matched ATAC-seq and RNA-seq datasets from the same set of input nuclei, the submergence response GRNs built for the four plants represent connections that are specific to submergence stress. Secondary circuits, such as connections between TFs and target genes that are specific to root development for example, are not reflected in the GRN. Because the GRNs in all four plant species are constructed using the same pieces, TFs and regulated genes, the resulting GRNs reflect an evolutionarily conserved network. However, the circuitry of connections between the TFs and the

regulated genes differs among the plant species, even though the pieces are the same. This demonstrates how a network that has existed for over 123 million years has evolved to be utilized differently.

It is still not known whether the submergence response GRNs that we described are present in other plant species, but the distantly related species used in our work make a strong argument that the submergence response that we categorized is probably present in the majority of angiosperms. The four motifs that orchestrate the submergence response are particularly interesting because they may act combinatorially to drive expression of nearby genes. This is of interest for both understanding how the orientation of these motifs relative to each other can regulate expression, as well as for developing submergence-specific regulatory elements that can be used to improve agriculture by either driving the expression of specific genes only during submergence or by being used as sensors to drive the expression of a visual reporter to identify plants or tissues undergoing submergence stress.

Protein-DNA interactions may leave footprints in chromatin accessibility data

Throughout my work, the farthest I could interrogate TF-binding sites using ATAC-seq data was to determine whether they are accessible or not. Originally, it appeared possible to use ATAC-seq results to determine if a specific TF was actually bound to DNA. This can be digitally visualized by processing paired-end ATAC-seq data to visualize only the specific cut sites in each read, which corresponds to the +4 base on the plus strand and the -5 base on the minus strand which represents the 9-bp offset where the Tn5 dimer cleaves the DNA strands (Karabacak Calviello et al. 2019). In the processed data, regions of high cut coverage with a drop in cutting within that region represent a footprint where a protein, such as a TF, is occupying DNA and preventing the Tn5 from cutting the DNA at that specific site, leaving a TF footprint in the accessible site (Figure 7.2A). However, the TF footprints observed in nuclei were also observed in the genomic DNA samples, which are supposed to be devoid of proteins and tertiary structure.

Detecting occupancy for many TFs in one chromatin accessibility assay has confounding factors associated with it that have been reported, such as DNA sequence biases of the enzyme and TF-specific

variations in binding kinetics (Koohy et al. 2013, He et al. 2014, Raj et al. 2014, Rusk 2014, Sung et al. 2014, Madrigal 2015). This was indeed the case with the 50+ positional weight matrices (PWMs) I analyzed throughout this dissertation research. Regardless of which cell type or condition was being analyzed, some PWMs had clear footprints whereas others had less obvious footprints, or what appeared to be super-hypersensitive sites with even more cutting at the mapped PWM sites. It is possible that the distinction among these different categories of footprints results from Tn5 sequence bias at these sites. Alternatively, these footprints may arise from differences in residence time by the TFs that bind these sequences (Sung et al. 2014), or perhaps the gDNA samples are not completely devoid of proteins.

To address the question of strongly interacting proteins on gDNA, I propose additional experiments that can be done using the datasets generated here, along with any ATAC-seq genomic DNA samples that are publicly available (Figure 7.2B). The proposed analyses consist of looking at the same set of mapped PWMs in the genomic DNA ATAC-seq samples isolated from the five different plant species studied in this dissertation research. Additionally, the same PWMs need to be visualized in ATAC-seq gDNA datasets that utilized a different isolation protocol, such as protease-treated DNA, or are from a very different background, such as animal gDNA. Even though gDNA still showed sequencing bias for Tn5 that may be caused by TFs that are still interacting with DNA, this finding is novel and represents potential future experiments that could categorize which TFs are still present on gDNA and what properties, such as long occupancy times, are responsible for their strong DNA interactions.

Technical improvements for INTACT-ATAC-seq

Improving the specificity of nuclei isolated using INTACT

The Isolation of Nuclei TAgged in specific Cell Types (INTACT) technique allows for the isolation of nuclei from specific cell types, providing chromatin accessibility and RNA readouts that are less heterogeneous, and more informative, compared to similar results obtained from whole tissue (Deal et al. 2010, Deal et al. 2011). However, even within a specific cell type there is heterogeneity among the different cells that make up that cell type. This has been observed in several different experiments

performed on single cells (Pott et al. 2015, Cao et al. 2018, Tasic et al. 2018, Rodriguez-Villalon et al. 2019, Shulse et al. 2019) . The variation likely arises, at least in part, from the different microenvironments the cells occupy due to being located at different places in absolute space. This can result in some cells of the same cell type, such as the hair cells of the root epidermis, being exposed to different environmental conditions, such as water availability in the immediate vicinity of the root hair cells.

Further optimization of nuclei isolated using INTACT can be achieved by exchanging the ubiquitously-expressed ACTIN2 promoter that drives the biotin ligase BirA gene with another promoter that is specific to some environmental stress. This would allow for the study of stress responding nuclei, isolated using streptavidin-coated magnetic beads, compared to the nuclei of the same cell type that are not responding to environmental stress, isolated by anti-GFP-coated magnetic beads. Additionally, nuclei isolated through INTACT can be used for single-cell combinatorial indexing methods (Cusanovich et al. 2015) to parse out the heterogeneous variation among cells of a specific type.

Improving sequencing efficiency and specificity

The three biggest improvements for performing ATAC-seq in isolated nuclei are: 1) reducing organellar reads in sequencing libraries, 2) accounting for biological variation within the nuclei of a given cell type, and 3) evaluating sequencing data efficiently and accurately.

Fewer organellar reads can be obtained from ATAC-seq libraries if the amplified libraries are treated with recombinant Cas9 protein complexed with a library of guide RNAs that target mitochondrial and chloroplast genomes (Gu et al. 2016, Montefiori et al. 2017). Alternatively, the addition of detergents to nuclei before tagmentation and the use of phosphate-buffered saline (PBS) in the transposition reaction have been shown to reduce the amount of contaminating organellar reads in the sequenced libraries (Corces et al. 2017). These improvements to ATAC-seq proved effective in animal cells but they have not yet been used in plants.

For the majority of our experiments we used between 25,000 to 50,000 nuclei as the input for the transposition reaction. However, as few as 500 plant nuclei have been used as input for the transposition reaction (Lu et al. 2017). Successful chromatin profiling has also been done by performing ATAC-seq on single cells (Pott et al. 2015, Buenrostro et al. 2018, Lareau et al. 2019, Ludwig et al. 2019, Shema et al. 2019). Most recently, the most powerful comparisons between chromatin accessibility and transcriptional response have come from single cell experiments done using jointly isolated DNA and RNA from the same cell (Cao et al. 2018). The covariance between chromatin accessibility and transcription across a large population of single cells increases the causal relations between nearby and distal regulatory elements and the regulated genes (Cao et al. 2018). This approach represents the gold standard for establishing a connection between chromatin organization and gene expression.

Improving motif discovery and mapping

We utilized two databases of publicly reported PWMs for the identification of TFs that bind overrepresented motifs in our work (Weirauch et al. 2014, O'Malley et al. 2016). These databases were curated by characterizing TF binding either in fragmented genomic sequences or DNA microarrays. More high-throughput methods need to be performed in plants, and their results need to be deposited to motif enrichment analyzers as references. Better algorithms still need to be developed for mapping PWMs genome-wide. In fact, a recent publication followed up on results from Chapter 4 and found that more interconnected maps can be built using ATAC-seq data if the PWM mapping program FIMO (Grant et al. 2011) that we utilized is replaced with more robust mapping algorithms such as cluster-buster, matrix-scan, and Motif Occurrence Detection Suite (Kulkarni et al. 2019). These represent technical limitations to building complete GRNs.

Final statements

Understanding how gene expression is regulated involves identifying TFs that bind cis elements and connecting these cis elements with their target genes. Combining ATAC-seq and RNA-seq

experiments, particularly in nuclei of a specific cell type, allows us to build inferred gene regulatory networks specific to a cell type or a specific response. This work is able to create maps of gene expression regulation that can identify new transcription factors and regulated genes important for cell differentiation or stress response. The conclusions from this work can be used to identify regulatory elements specific to different cell types or submergence stress that can be used to drive expression of reporters or genes of interest within those cells or conditions. Additionally, the approach outlined within this work can be applied to other organisms and to study disease.

FIGURES

Fig 7.1

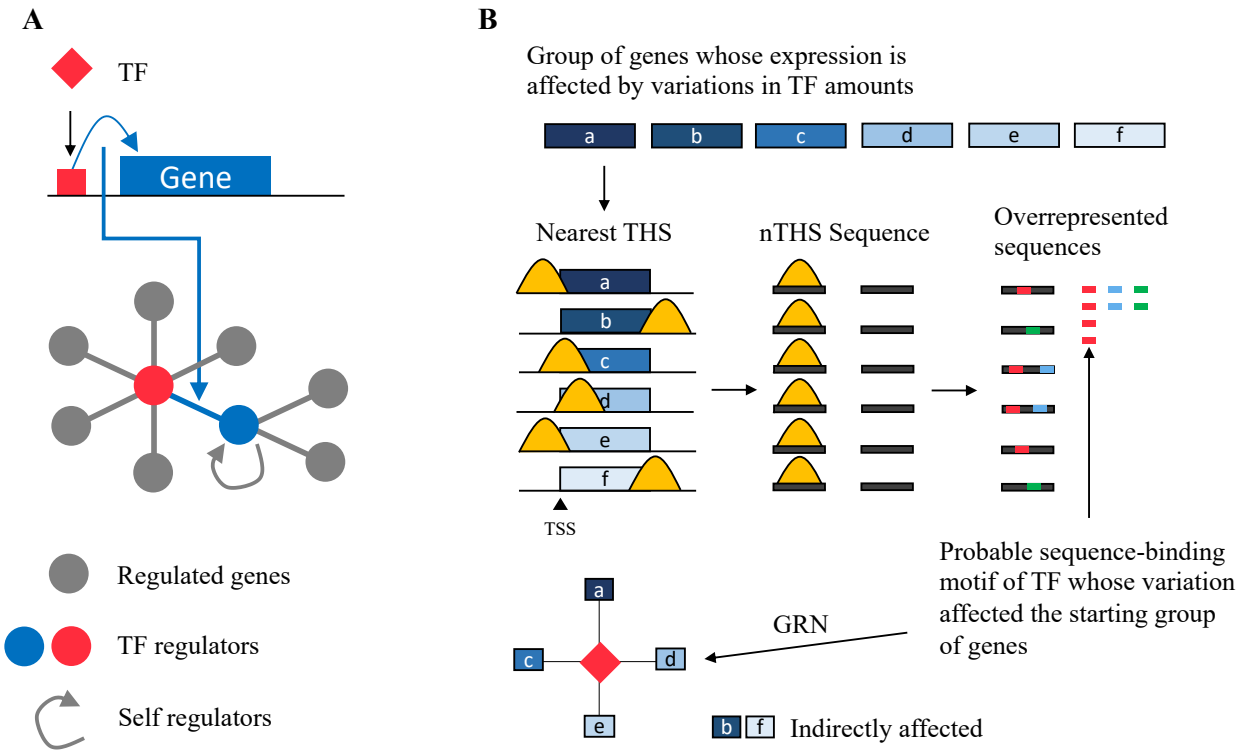


Figure 7.1. Combining ATAC-seq data with RNA-seq results from TF mutant or overexpression lines. (A) The construction of a Gene Regulatory Network (GRN) consists of connections between regulators, such as TFs, and the target genes they regulate. The blue line denotes a connection between a regulated gene its regulating TF whose binding site is proximal to the gene. The small GRN that is shown below the schematic for the connection depicts the connection along with other genes that TF regulates. Additionally, the GRN depicts the regulated gene of the red TF as a TF itself and shows two other genes that are regulated by this TF, including itself. **(B)** RNA-seq results from lines where a TF is knocked out or overexpressed can be used to identify genes that are regulated by that TF. ATAC-seq data can be combined with the group of genes that are regulated by the TF to build GRNs. This is done by finding nearest THSs for each gene in the group of affected genes. Sequences from these nearest THSs (nTHSs) can be evaluated to find overrepresented motifs that the TF may be binding. The presence of the

overrepresented motif can then be used as the binding site for the regulator to make the connection between TF and the regulated gene.

Fig 7.2

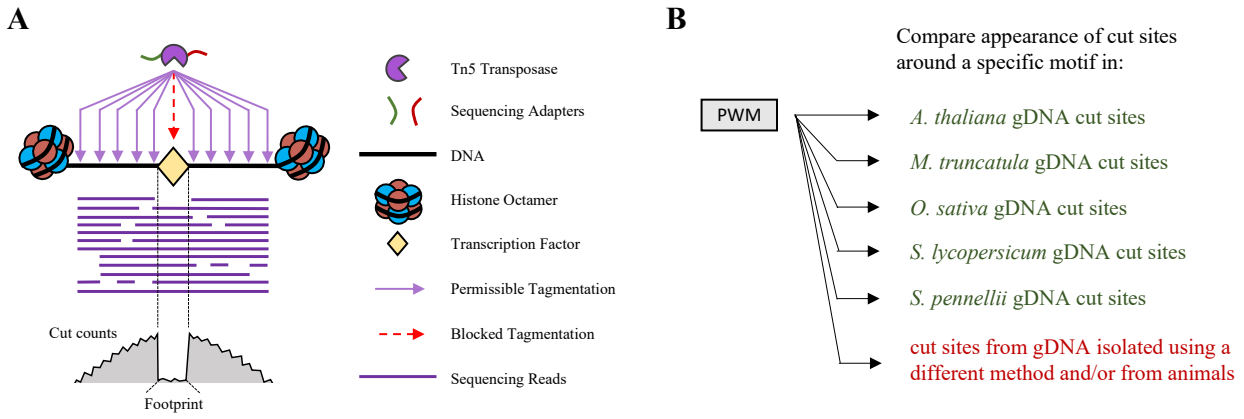


Figure 7.2. Transcription Factor footprints and sequence-specific bias. (A) Overview of TF-footprints detected in ATAC-seq data. The Tn5 transposase is able to cut DNA and ligate sequencing adapters into accessible chromatin regions. However, the specific sites where DNA is occupied by TFs prevent Tn5 from cutting specifically at those sites. Sequenced reads can be processed to show the exact sites where Tn5 cut the DNA. The lack of cutting where DNA was occupied by a TF in a region of accessible chromatin appears as a drop in cut counts around that region, which is referred to as a footprint. **(B)** Proposed analyses to determine whether genomic DNA (gDNA), also referred to as naked DNA, is free of any proteins interacting with the DNA. In addition to performing cut site visualization for several different PWMs in all the gDNA ATAC-seq datasets generated for this dissertation research, the same visualization should be performed on publicly available gDNA ATAC-seq datasets if they utilized a different gDNA isolation protocol or are from a significantly different species.

LITERATURE CITED

Arnold, C. D., D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath and A. Stark (2013). "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq." Science **339**(6123): 1074.

Buenrostro, J. D., M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang and W. J. Greenleaf (2018). "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation." Cell **173**(6): 1535-1548 e1516.

Cao, J., D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell and J. Shendure (2018). "Joint profiling of chromatin accessibility and gene expression in thousands of single cells." Science **361**(6409): 1380-1385.

Corces, M. R., A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf and H. Y. Chang (2017). "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues." Nat Methods **14**(10): 959-962.

Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell and J. Shendure (2015). "Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing." Science **348**(6237): 910-914.

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

Deal, R. B. and S. Henikoff (2011). "The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana." Nat Protoc **6**(1): 56-68.

Grant, C. E., T. L. Bailey and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Gu, W., E. D. Crawford, B. D. O'Donovan, M. R. Wilson, E. D. Chow, H. Retallack and J. L. DeRisi (2016). "Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove

unwanted high-abundance species in sequencing libraries and molecular counting applications." Genome Biol **17**: 41.

He, H. H., C. A. Meyer, S. S. Hu, M. W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu and M. Brown (2014). "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification." Nat Methods **11**(1): 73-78.

Heintzman, N. D., G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis and B. Ren (2009). "Histone modifications at human enhancers reflect global cell-type-specific gene expression." Nature **459**(7243): 108-112.

Karabacak Calviello, A., A. Hirsekorn, R. Wurmus, D. Yusuf and U. Ohler (2019). "Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling." Genome Biol **20**(1): 42.

Koohy, H., T. A. Down and T. J. Hubbard (2013). "Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme." PLoS One **8**(7): e69853.

Kulkarni, S. R., D. M. Jones and K. Vandepoele (2019). "Enhanced maps of transcription factor binding sites improve regulatory networks learned from accessible chromatin data." Plant Physiol.

Lareau, C. A., F. M. Duarte, J. G. Chew, V. K. Kartha, Z. D. Burkett, A. S. Kohlway, D. Pokholok, M. J. Aryee, F. J. Steemers, R. Lebofsky and J. D. Buenrostro (2019). "Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility." Nat Biotechnol **37**(8): 916-924.

Lowe, E. K., C. Cuomo, D. Voronov and M. I. Arnone (2019). "Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity." Methods Cell Biol **151**: 115-126.

Lu, Z., B. T. Hofmeister, C. Vollmers, R. M. DuBois and R. J. Schmitz (2017). "Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes." Nucleic Acids Res **45**(6): e41.

Ludwig, L. S., C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack, T. Law, C. Rodman, J. H. Chen, G. M. Boland, N. Hacohen, O. Rozenblatt-Rosen, M. J. Aryee, J. D. Buenrostro, A. Regev and V. G. Sankaran (2019). "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics." Cell **176**(6): 1325-1339 e1322.

Madrigal, P. (2015). "On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions." Front Bioeng Biotechnol **3**: 144.

Montefiori, L., L. Hernandez, Z. Zhang, Y. Gilad, C. Ober, G. Crawford, M. Nobrega and N. Jo Sakabe (2017). "Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9." Sci Rep **7**(1): 2451.

O'Malley, R. C., S. C. Huang, L. Song, M. G. Lewsey, A. Bartlett, J. R. Nery, M. Galli, A. Gallavotti and J. R. Ecker (2016). "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape." Cell **165**(5): 1280-1292.

Pott, S. and J. D. Lieb (2015). "Single-cell ATAC-seq: strength in numbers." Genome Biol **16**: 172.

Raj, A. and G. McVicker (2014). "The genome shows its sensitive side." Nat Methods **11**(1): 39-40.

Rodriguez-Villalon, A. and S. M. Brady (2019). "Single cell RNA sequencing and its promise in reconstructing plant vascular cell lineages." Curr Opin Plant Biol **48**: 47-56.

Rowley, M. J., M. H. Nichols, X. Lyu, M. Ando-Kuri, I. S. M. Rivera, K. Hermetz, P. Wang, Y. Ruan and V. G. Corces (2017). "Evolutionarily Conserved Principles Predict 3D Chromatin Organization." Mol Cell **67**(5): 837-852 e837.

Rusk, N. (2014). "Transcription factors without footprints." Nature Methods **11**(10): 988-989.

Shema, E., B. E. Bernstein and J. D. Buenrostro (2019). "Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution." Nat Genet **51**(1): 19-25.

Shulse, C. N., B. J. Cole, D. Ciobanu, J. Lin, Y. Yoshinaga, M. Gouran, G. M. Turco, Y. Zhu, R. C. O'Malley, S. M. Brady and D. E. Dickel (2019). "High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types." Cell Rep **27**(7): 2241-2247 e2244.

Sun, J., N. He, L. Niu, Y. Huang, W. Shen, Y. Zhang, L. Li and C. Hou (2019). "Global Quantitative Mapping of Enhancers in Rice by STARR-seq." Genomics Proteomics Bioinformatics **17**(2): 140-153.

Sung, M. H., M. J. Guertin, S. Baek and G. L. Hager (2014). "DNase footprint signatures are dictated by factor dynamics and DNA sequence." Mol Cell **56**(2): 275-285.

Tasic, B., Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren, M. N. Economo, S. Viswanathan, O. Penn, T. Bakken, V. Menon, J. Miller, O. Fong, K. E. Hirokawa, K. Lathia, C. Rimorin, M. Tieu, R. Larsen, T. Casper, E. Barkan, M. Kroll, S. Parry, N. V. Shapovalova, D. Hirschstein, J. Pendergraft, H. A. Sullivan, T. K. Kim, A. Szafer, N. Dee, P. Groblewski, I. Wickersham, A. Cetin, J. A. Harris, B. P. Levi, S. M. Sunkin, L. Madisen, T. L. Daigle, L. Looger, A. Bernard, J. Phillips, E. Lein, M. Hawrylycz, K. Svoboda, A. R. Jones, C. Koch and H. Zeng (2018). "Shared and distinct transcriptomic cell types across neocortical areas." Nature **563**(7729): 72-78.

Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J. C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F. Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker and T. R. Hughes (2014). "Determination and inference of eukaryotic transcription factor sequence specificity." Cell **158**(6): 1431-1443.

Zhang, W., T. Zhang, Y. Wu and J. Jiang (2014). "Open chromatin in plant genomes." Cytogenet Genome Res **143**(1-3): 18-27.