Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in *A. thaliana*

Alessandra Maria Sullivan

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Christine Queitsch, Chair

Philip Green

Celeste A. Berg

Program authorized to offer degree:

Genome Sciences

University of Washington

Abstract

Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana

Alessandra Maria Sullivan

Chair of the Supervisory Committee:
Christine Queitsch
Associate Professor of Genome Sciences
Department of Genome Sciences

Our understanding of gene regulation in plants is constrained by our limited knowledge of plant *cis*-regulatory DNA and its dynamics. One way in which *cis*-regulatory elements can be delineated is by their characteristic hypersensitivity to the endonuclease DNase I. In this dissertation I present the development and application of a DNase I-based assay for *A. thaliana.* I describe the development of a DNase I-seq method that can be used to map genome-wide chromatin accessibility with nucleotide resolution and cell-type specificity. I then use this method to map DNase I hypersensitive sites (DHSs) in *A. thaliana* seedlings and use genomic footprinting to delineate ~700,000 sites of *in vivo* transcription factor (TF) occupancy. I reveal general properties of DHSs in *A. thaliana* are novel, including evidence that highly significant GWAS variants are enriched within DHSs and that widespread TF binding within exons may have shaped codon usage patterns. I also show that the architecture of *A. thaliana* TF regulatory networks is strikingly similar to that of

animals in spite of diverged regulatory repertoires. I then explore chromatin dynamics in response to environmental stimuli by mapping the chromatin landscape during heat shock and photomorphogenesis, disclosing thousands of environmentally-sensitive elements and the TFs that bind them. Next, I continue exploring chromatin dynamics, this time looking at developmentally dynamic DHSs during the maturation of seed-coat epidermal cells during the transition from growth to mucilage secretion. I show that differentially expressed genes are associated with dynamic DHSs and implicate new TFs and candidate genes involved in seed coat epidermal differentiation. Finally, I investigate the natural variation in chromatin accessibility through the examination of chromatin landscapes of 5 diverse *A. thaliana* ecotypes. I show that variable DHSs are more polymorphic than static DHSs across the accessions. I also show that deletions account for 15% of variable DHSs, suggesting they are a powerful force in shaping diverse patterns of gene regulation in the ecotypes. In the appendices I present supplemental figures and methods for Chapter 2, as well as a project summary and general information on phenotypic robustness, which is affected by *cis*-regulatory elements.

Acknowledgments

This dissertation would not have been possible without the help of many many people. First I would like to thank my family, especially Grandpa Wally, who encouraged and inspired me to be curious about living things. And Dad, who exposed me to science through his career in physics - I vividly remember meeting my first transgenic mouse at Lawrence Berkeley Lab when I was a kid. I would also like to thank my supportive partner, Shawn, who was so helpful with analyzing DNase I data, that he is an author on my first paper!

I could not have chosen a better community to join for my PhD training than the Genome Sciences Department - thanks! I would like to thank my advisor, Christine Queitsch, for the opportunity to lead big projects and for her enthusiasm for science and my ideas. I would like to thank members of the Queitsch lab for their friendship, feedback, collaborations, and patience during lab meetings: Jen Lachowiec, Karla Schultz, Max Press, Michael Dorrity, Cris Alexandre, Tzitziki Lemus, Pauline Rival, Kerry Bubb, Alex Mason, Beth Morton, James Urton, Janne Lempe, and Soledad Undurraga. Kerry Bubb in particular has made my DNase I work both possible and enjoyable. I would also like to thank the members of the Stamatoyannopoulos lab, especially John Stamatoyannopoulos, Richard Sandstrom, Bob Thurman, Shane Neph, and Andrew Stergachis, for their help with figures and data analysis.

Last but not least, I would like to thank my committee, John Stamatoyannopoulos, Celeste Berg, Phil Green, Christine Queitsch, and Eric Klavins, for their time, feedback, and encouragement.

Table of contents

Chapter 1. Introduction to gene regulation in plants

Why study gene regulation in plants?

As sessile organisms, plants are subject to influences from their environment. The course of plant development can be altered by light, temperature, humidity and the presence of pathogens, all of which bring about large changes in gene expression. Gene expression is largely controlled at the level of transcription where *cis*-acting DNA elements and the *trans*-acting factors that bind them ultimately coordinate the activity of RNA polymerases. Careful orchestration of gene activity is required for appropriate physiological responses to the environment and explains how a single genome can give rise to many diverse cell types (Davidson and Britten, 1971).

Despite the important role plants play in human survival, their regulatory repertoires remain poorly understood. Large consortia are dedicated to the annotation of DNA elements, including *cis*-regulatory DNA elements and transcription factor regulatory networks, in animals such as *Drosophila*, *C. elegans* and humans (ENCODE, 2012; Gerstein et al., 2010; Levine, 2010a; Roy et al., 2010); however, no similar large-scale, coordinated effort exists for *A. thaliana* or any other plant.

Identification of *cis*-regulatory elements with DNase I mapping

Accessible regulatory DNA elements such as promoters (Wu et al., 1979a), enhancers (Banerji et al., 1983) , insulators (Chung et al., 1997), silencers (Antoniou et al., 1988), and locus control regions (Talbot et al., 1989) can be detected by their characteristic hypersensitivity to the endonuclease DNase I (Wu et al., 1979a; Wu et al., 1979b). Within

DNase I hypersensitive sites, "footprints" mark protected regions of DNA caused by DNA-binding proteins (Galas and Schmitz, 1978; Wu, 1984). DNase I sensitivity assays were initially carried out with Southern blots by probing individual, treated or control sequences with radiolabeled probes (Wu et al., 1979a; Wu et al., 1979b). Since then, diverse technologies have emerged (Zhang et al., 2014) and been largely supplanted by the widely used DNase I-seq technology (ENCODE, 2012; Hesselberth et al., 2009; Thomas and Elgin, 1988; Thomas et al., 2011; Thurman et al., 2012b; Zhang et al., 2012a), which relies on next generation sequencing of low molecular weight DNA released from nuclei treated with DNase I (Hesselberth et al., 2009). Improvements to sequencing and library preparation have enhanced the information content of DNase I-seq, for example regulatory regions and nucleosome occupancy can be captured with the DNase I-based technique DNase-FLASH (Vierstra et al., 2014). A new way of assaying accessible chromatin using the transposition based assay ATAC-seq (Buenrostro et al., 2013) has also emerged; although its usage by the community is still limited, it may prove a viable alternative to DNase I-seq.

DNaseI hypersensitivity mapping (Thurman et al., 2012b) and genomic footprinting (Hesselberth et al., 2009; Neph et al., 2012c) have been extensively employed to delineate *cis*-regulatory DNA and transcription factor (TF) occupancy at nucleotide resolution in higher organisms. Human disease- and trait-associated variation is concentrated in regulatory DNA, and localizes within transcription factor recognition sequences and the regulatory pathways they define (Maurano et al., 2012b). Global mapping of transcription factor footprints provides a uniquely powerful foundation for construction of extensive TF regulatory networks encompassing hundreds of TFs active within a given cell type, as well

as comparative analysis of regulatory network dynamics between different cell states (Neph et al., 2012b).

DNase I mapping in plants has historically lagged behind the animal field. One explanation is that the plant cell walls and high chloroplast numbers interfere with high quality nuclei isolation. From my own experiences, I found isolating nuclei suitable for DNase I-seq to be challenging; only about 1 in 10 DNase I-seq samples produced signal. It was not until I employed a nuclear enrichment strategy (INTACT method (Deal and Henikoff, 2010)), in which the nuclear envelope is tagged with biotin and captured with streptavidin beads (**Figure 1**), that I was able to consistently generate high quality DNase I-seq libraries. An additional explanation is that plants were not among the founding members (human cell lines, SV40, and *Drosophila*) of the gene regulation field. However, as will be discussed below, one can take advantage of the knowledge of gene regulation in metazoans towards predicting and interpreting gene regulatory mechanisms in plants.

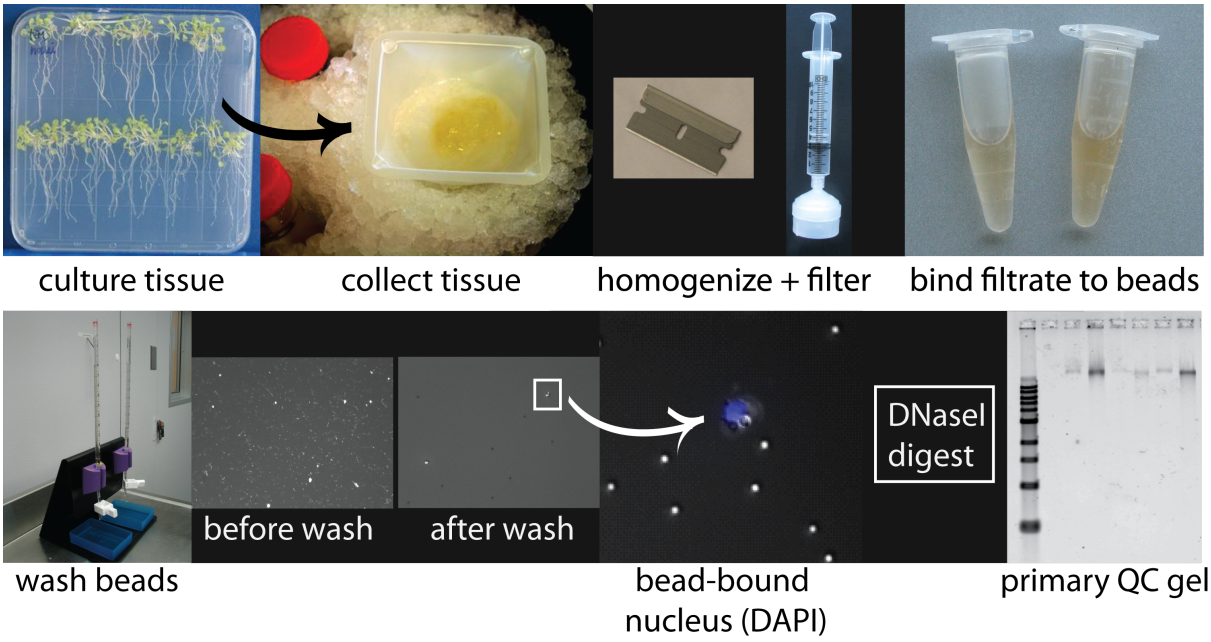


Figure 1. Schematic of tissue culture and INTACT nuclei isolation in roots. Seedlings were grown vertically on sterile media. Roots were harvested and homogenized by cutting carefully with a razor blade. The resulting homogenate was filtered and combined with magnetic streptavidin beads, which bind the biotin labeled nuclei. Beads (and nuclei) were captured using a magnet and washed; bead-bound nuclei were identified with DAPI staining. The resulting nuclei and beads were digested with DNase I and run on a gel to verify DNA quality (QC = quality control).

How does gene regulation compare in plants and animals?

The last common ancestor to plants and animals is hypothesized to have been single celled and to have lived 1.6 billion years ago (Wang et al., 1999). Despite this deep divergence, plants and animals share many aspects of gene regulation including core chromatin and transcription machinery, regulatory motifs, and regulatory phenomena. Plants and animals have conserved histones, nucleosome structure, histone modifications, high mobility group proteins, and core transcriptional machinery (e.g. Pol-II) (Bianchi and Agresti, 2005; Singh, 1998; Spiker, 1985; The_Arabidopsis_Genome_Initiative, 2000). Plants and animals also share homologs for chromatin remodeling components including SWI/SNF2-like components (Alvarez-Venegas, 2010; The_Arabidopsis_Genome_Initiative, 2000), Polycomb and Trithorax group proteins (Alvarez-Venegas, 2010), and TF families. In *A. thaliana* 55% of transcription factors are from TF families common to both plant and non-plant eukaryotes (Riechmann et al., 2000). Because many core transcriptional components are conserved, core promoter motifs such as the TATA-box, initiator element and kozak motifs are also shared between plants and animals (Yamamoto et al., 2007).

Plants and animals also exhibit similar gene regulatory phenomena: both have sequential waves of master regulators, which initiate specific patterns of expression in space and time during development (Meyerowitz, 2002). Finally, both animals and plants rely on complex combinatorial interactions between core transcriptional machinery and transcription factors, which modulate--and can be modulated by--chromatin architecture (Adrian et al., 2010; Kumar and Wigge, 2010; Michaels, 2009; Singh, 1998).

While many core chromatin and transcriptional components are conserved between plants and animals, the components are often slightly different in function or have multiplied and diverged in plants (The_Arabidopsis_Genome_Initiative, 2000). Plant RNA polymerases are a good example of this diversification. Eukaryotes have three RNA polymerases (Pol I, II, and III) that are responsible for synthesizing RNA in the nucleus. Plants have additional polymerases (Pol IV and Pol V), which originated from duplications of Pol II (Luo and Hall, 2007; Luo et al., 2007). In *A. thaliana* and maize, Pol IV and Pol V are involved in silencing of transposable elements through RNA-directed DNA methylation pathways (Erhard et al., 2013; Zhong et al., 2012). For more information on these plant polymerases, see Appendix A. Transcription factor families present in both animals and plants have also expanded (The_Arabidopsis_Genome_Initiative, 2000). For example, there is a single Heat Shock Factor (HSF) TF in *Drosophila* and yeast, and 21 in *A. thaliana*. Core promoter regulatory elements and composition also differ between animals and plants. Plant promoters do not have CpG islands, but do commonly have pyrimidine patches (Y-Patch) and other general sequence enrichments (e.g. GA, CA elements), which are not present in animals (Morton et al., 2014; Yamamoto et al., 2007).

Both plants and animals exhibit complex promoters (Morton et al., 2014) and distal elements capable of looping to interact with promoter machinery (Cao et al., 2014; Mendes et al., 2013), however, the *cis*-regulatory elements themselves have been independently recruited to their tasks in plants and animals (Meyerowitz, 2002). There are regulatory elements that are well characterized in animals, but are undiscovered or scarcely documented in plants. These include insulators, which block promoters from interacting with enhancers or the advance of heterochromatinization (e.g. CTCF) (Burgess-Beusse et

al., PNAS 2002); silencers, which can act in an orientation- and distance-independent manner to block transcription (Ogbourne and Antalis Biochem J. 1998); and long-range (>10kb) enhancers, which increase promoter output through complex looping interactions also in an orientation- and distance-independent fashion (Dekker et al., 2013; Levine, 2010b; Levine and Tjian, 2003). In fact, the only long-range enhancer element identified in plants to date is the *tb1* enhancer, which is ~60kb away from *tb1* (Clark et al., 2006); this enhancer is particularly important because changes in *tb1* regulation are responsible for morphological changes during maize domestication.

By mapping and analyzing the dynamics of regulatory elements and TF networks in *A. thaliana* I have helped fill the knowledge gap between what we know about plant and animal chromatin and transcriptional regulation. By identifying the location, dynamics, and in some cases the *trans*-acting factors of regulatory elements, I opened the door to further in-depth studies of individual regulatory elements. I found complex patterns of chromatin, which warrant further study and have the potential to shed light on further similarities and differences between plant and animal gene regulation.

Dissertation objectives

To address the lack of *cis*-regulatory information in the model plant *A. thaliana*, in collaboration with members of the Queitsch, Nemhauser, and Stamatoyannopoulos labs:

- Developed methods for cell-type specific, genome-wide DNase I mapping of *cis*-regulatory elements in *A. thaliana*
- Generated DNase I-seq data for
    - Root tissue

    - Seedlings under different environmental treatments (light, heat)
    - Cell types (seed coat cells, root hair, and root non-hair cells)
    - Diverse *A. thaliana* ecotypes
- Identified environmentally and developmentally activated regulatory elements
- Generated TF regulatory networks based on DNase I footprinting information
- Generated novel *de novo* motif models based on DNase I footprinting information
- Identified known master regulators and predicted candidate genes based on their local regulatory landscape
- Identified conserved regulatory elements among *A. thaliana* ecotypes

Chapter 2. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*[12]

Acknowledgements

In this acknowledgements section, and those that follow, I detail my own contributions as well as the significant contributions of collaborators.

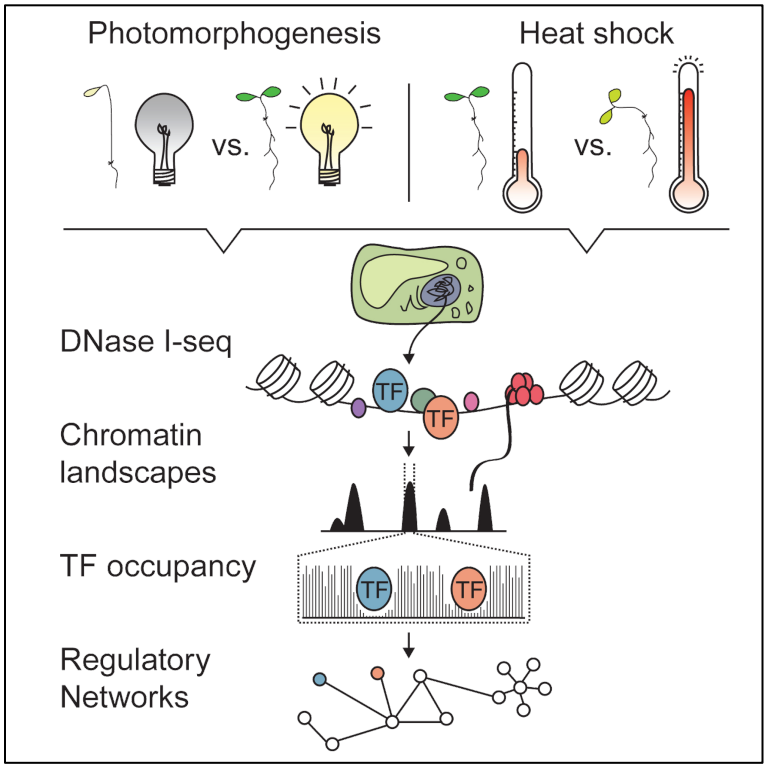
I developed DNase I methods for *A. thaliana*; prepared the root, root hair and root non-hair samples; analyzed and interpreted post-processed data such as hotspots, DHSs, footprints, motif instances, and networks; created all of the figures with the exception of Figure 2, which was created by Andrew Stergachis; and wrote the manuscript with help from Christine Queitsch and John Stamatoyannopoulos. Jennifer Nemhauser and Andrej Arsovski chose conditions for light and heat treatments; Andrej Arsovski and Agnieszka Thomson DNase I-treated the light- and heat-treated samples. Members of John Stamatoyannopoulos' Informatics Group and sequencing unit prepared and sequenced the samples, performed alignments, and called hotspots, peaks, footprints, and networks. Kerry Bubb identified dynamic DHSs, performed the methylation analysis, and assisted me with other analyses. Motif PWMs were provided by Tim Hughes and Mathew Weirauch. Benjamin Vernot performed the nucleotide diversity analysis.

---

[1] This chapter was published in Cell Reports on September 24, 2014.
Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*
Alessandra M. Sullivan*[1], Andrej A. Arsovski*[2], Janne Lempe*[1], Kerry L. Bubb*[1], Matthew T. Weirauch[3], Peter J. Sabo[1], Richard Sandstrom[1], Robert E. Thurman[1], Shane Neph[1], Alex P. Reynolds[1], Andrew B. Stergachis[1], Benjamin Vernot[1], Audra K. Johnson[1], Eric Haugen[1], Shawn T. Sullivan[1], Agnieszka Thompson[1], Fidencio V. Neri III[1], Molly Weaver[1], Morgan Diegel[1], Sanie Mnaimneh[4], Ally Yang[4], Timothy R. Hughes[4,5], Jennifer L. Nemhauser[2], Christine Queitsch[1,#], John A. Stamatoyannopoulos[1#]

[2] Supplementary information for this chapter is located in Appendix B

Graphical abstract



Summary

Our understanding of gene regulation in plants is constrained by our limited knowledge of plant *cis*-regulatory DNA and its dynamics. We mapped DNaseI hypersensitive sites (DHSs) in *A. thaliana* seedlings and used genomic footprinting to delineate ~700,000 sites of *in vivo* transcription factor (TF) occupancy at nucleotide resolution. We show that variation associated with 72 diverse quantitative phenotypes localizes within DHSs. TF footprints encode an extensive *cis*-regulatory lexicon subject to recent evolutionary pressures, and widespread TF binding within exons may have shaped codon usage patterns. The architecture of *A. thaliana* TF regulatory networks is strikingly similar to that of animals in spite of diverged regulatory repertoires. We analyzed

regulatory landscape dynamics during heat shock and photomorphogenesis, disclosing thousands of environmentally-sensitive elements, and enabling mapping of key TF regulatory circuits underlying these fundamental responses. Our results provide an extensive resource for the study of *A. thaliana* gene regulation and functional biology.

Highlights

- *A. thaliana* regulatory DNA, TF footprints, and conserved *cis*-regulatory lexicon
- TF binding in protein-coding exons may have shaped *A. thaliana* codon usage
- *A. thaliana* TF network architecture is strikingly similar to human
- Light- and heat-cued regulatory DNA dynamics and TF network remodeling

Introduction

As sessile organisms, plants are shaped by their environment and respond acutely to cues such as light and temperature. Such responses result in significant alterations in gene expression; however, the *cis*-regulatory elements and transcription factor regulatory networks controlling these changes remain largely undefined.

DNaseI hypersensitive sites (DHSs) (Wu et al., 1979b) are the *sine qua non* of regulatory DNA in eukaryotic genomes, and DNaseI hypersensitivity mapping (Thurman et al., 2012a) and genomic footprinting (Neph et al., 2012c) have been extensively employed to delineate *cis*-regulatory DNA and TF occupancy at nucleotide resolution in higher organisms. Such maps have provided wide-ranging insights into genome function, evolution, and the genetic basis of common phenotypes (Maurano et al., 2012a). Global mapping of transcription factor footprints provides a powerful foundation for construction of extensive regulatory networks encompassing hundreds of TFs and comparative analysis of regulatory network dynamics (Neph et al., 2012b).

Here we apply these powerful approaches to delineate the regulatory DNA landscape of the reference plant *A. thaliana* at unprecedented resolution; to analyze the

relationship between regulatory DNA and phenotype-associated variation; to define the major features of the *A. thaliana* TF lexicon and regulatory network architecture; and to map the regulatory circuitry underlying responses to temperature and light, the most important environmental cues shaping plant growth and development. All raw and processed data are available at plantregulome.org.

Results

Distribution and features of *A. thaliana* DHSs

To extract nuclei from *A. thaliana* tissues, we created an INTACT (Deal and Henikoff, 2010) line constitutively expressing a nuclear pore biotin tag, and developed a protocol for gentle mechanical disruption of plant tissue to release nuclei, which were isolated using streptavidin affinity reagents (**Supplemetal Methods**). We then adapted previous DNase-seq protocols (John et al., 2011) to create, map, and sequence DNaseI fragment libraries, following which we modified DHS and footprint detection algorithms for use in the smaller *A. thaliana* genome with appropriate false discovery rate (FDR) thresholds (**Supplemetal Methods**).

We first performed standard-depth DNase-seq on whole plant seedlings, and confirmed that the resulting maps visualized DHSs present in more specialized subsamples such as root tissue or root epidermal cell types (e.g., root hair cells, root non-hair cells), while also revealing quantitative differences in accessibility at individual elements (**Figure 1A, Table S1**).

We then performed deep DNaseI-seq (>260 million uniquely mapped genomic reads) on high-quality seedling samples, and defined 34,288 DHS at a stringent false-discovery rate (FDR) threshold (FDR 1%), which covered 4% of the *A. thaliana* genome. The DHS distribution across the *A. thaliana* genome reflected its high gene density, with 37% of DHSs localizing within ~400 bp upstream of transcription start sites (TSSs) (**Figure 1B**). DHSs were relatively enriched in intergenic regions and 5'UTRs (**Figure 1B; Table S2**). Although *A. thaliana* transposons and introns were generally depleted for DHSs (**Figure 1B**)**,** intronic DHSs were more likely to reside within genes encoding transcriptional regulators such as the *PHYTOCHROME INTERACTING FACTOR*s *1* and *4* (PIFs), and *ELONGATED HYPOCOTYL 5* (HY5) ($p < 2.2x10^{-16}$; binomial distribution test), consistent with known intron-dependent regulation of some plant TFs (Sieburth and Meyerowitz, 1997) (**Table S3**).

A. thaliana DHSs were depleted for DNA methylation (Lister et al., 2008) ($p<1.0X10^{-50}$; binomial distribution test), but contained altered ratios of cytosine methylation contexts (CG, CHG, CHH; H indicates a non-G base); specifically, the proportion of me-C in the asymmetric CHH context increased ($p<2.2x10^{-16}$; chi-square) (**Figure S1A; Table S4**). All DHSs, regardless of whether they coincided with transposable elements or repeats, had similar ratios of cytosine methylation contexts. In plants, asymmetric methylation is maintained by constant *de novo* methylation and silences repeated and foreign DNA, including transposons (Law and Jacobsen, 2010). The increased presence of me-C in the plastic CHH context in DHSs is consistent with involvement of these regions in dynamic gene regulation.

As previously reported in other organisms (Hesselberth et al., 2009; Neph et al., 2012c) TF occupancy mapped using orthogonal approaches such as ChIP-seq localized within DHSs. For example, reproducible *A. thaliana* PIF3 ChIP-seq peaks (Zhang et al., 2013) were far more likely to co-localize with DHSs than non-reproducible ones (**Figure S1B**), and PIF3 motifs were highly enriched in reproducible ChIP-seq peaks that co-localized with DHSs (**Figure S1C, D**).

Footprinting the *A. thaliana* genome

Dense mapping of DNaseI cleavages enables genome-wide mapping of TF footprints (Hesselberth et al., 2009; Neph et al., 2012c). We defined 697,899 footprints at 1% FDR in a deeply sequenced seedling sample. Distinct footprints were readily apparent in whole seedling data, and could be resolved to specific TF recognition sequences defined by classical footprinting assays (occupancy of the photomorphogenesis master regulator HY5 within the *RBCS1A* promoter (Chattopadhyay et al., 1998) (**Figure S2E**)). More complicated relationships were also discernable. For example, footprints in the promoter of the cell cycle control gene *RETINOBLASTOMA-RELATED (RBR)* coincided with binding sites for the cell cycle control TFs E2F and E2L1(Gutierrez, 2009)(**Figure 1C**). *RBR* is the plant homolog of the human retinoblastoma gene, the protein product of which targets and inactivates E2F transcription factors (Gutierrez, 2009). The footprint data suggest a feedback loop in which *RBR* expression is regulated by its E2F targets (Vandepoele et al., 2005).

Within TF recognition sequences, per-nucleotide DNA accessibility is heterogeneous and tracks the topology of the protein-DNA interface (Hesselberth et al., 2009). This feature was evident for plant-specific TFs such as ATERF-1 (**Figure S1E**). Two of the originally-

defined MADS box factors, the *A. thaliana* homeotic factor AGAMOUS (AG) and the human cell cycle regulator Serum Response Factor (SRF) share similar DNase I cleavage profiles, suggesting that DNA accessibility patterns recapitulate DNA binding domain conservation (**Figure S1F**).

Expanding the *A. thaliana cis*-regulatory lexicon

TF footprints reflect occupancy of recognition elements by their cognate TFs. Footprints can be systematically mined to derive the *cis*-regulatory lexicon for an organism (Neph et al., 2012c). We performed *de novo* motif discovery on the 697,899 well-defined (FDR 1%) seedling footprints. We identified a total of 636 distinct 8-16 bp motifs, each of which was detected in at least 1,466 footprints (median 4,799 footprints) (**Figure 1D; Table S5**). These 636 motifs accounted for more than 89% of the seedling footprints, and encompassed 80/82 of previously-defined plant TF recognition sequences (Bryne et al., 2007; Matys et al., 2006).

To validate the footprint-derived motifs, we compared them with 382 motif models derived from protein-binding microarray (PBM) analysis of 334 cloned *A. thaliana* TFs (Franco-Zorrilla et al., 2014; Weirauch, (in press)), (**Table S6)**. Of the experimentally-defined motif models, 96% (366/382) were close matches to at least one of our footprint-derived motifs (**Figure 1D**).

To distinguish novel footprint-derived recognition sequences from known motifs, we subtracted all motif models that resembled any known TRANSFAC or JASPAR or PBM-motif using liberal matching criteria (Gupta et al., 2007) (**Methods**), which yielded 112 novel motif models (**Table S5**).

To validate these 112 motifs, we analyzed recent evolutionary selection within these elements using nucleotide diversity data for 80 *A. thaliana* accessions (Cao et al., 2011a). Similar to known motifs, the novel 112 footprint-derived motifs showed significantly reduced nucleotide diversity relative to neutrally-evolving sequences (**Figure 1E**), compatible with recent evolutionary selection. These results indicate that TF footprints collectively define an evolving functional compartment of the *A. thaliana* genome (**Figure 1E**).

A. thaliana GWAS variants are enriched in DHSs

In human, common disease- and phenotypic trait-associated variation mapped in genome-wide association studies (GWAS) localizes in DHSs (Maurano et al., 2012a). We sought to determine whether *A. thaliana* single nucleotide variants (SNVs) associated with diverse phenotypes (Atwell et al., 2010a) showed similar properties. *A. thaliana* GWAS studies encompass far smaller sample sizes than human studies (<200 strains in *A. thaliana* vs. thousands of humans), and are complicated by extensive population structure (Clark et al., 2007), leading to many false positive associations for complex traits. Despite these limitations, we found that a significantly greater fraction of trait-associated SNVs resided within DHSs; moreover, this fraction increased with increasing SNV significance. This pattern holds for the majority (72 of 107) of GWAS phenotypes (**Table S7**). Flowering time is among the GWAS phenotypes exhibiting an enrichment of strongly-associated SNVs in DHSs (**Figure 1F**; $P < 0.0015$, K-S test). DHS stratification of phenotype-associated variants may thus highlight the most promising variants for functional studies.

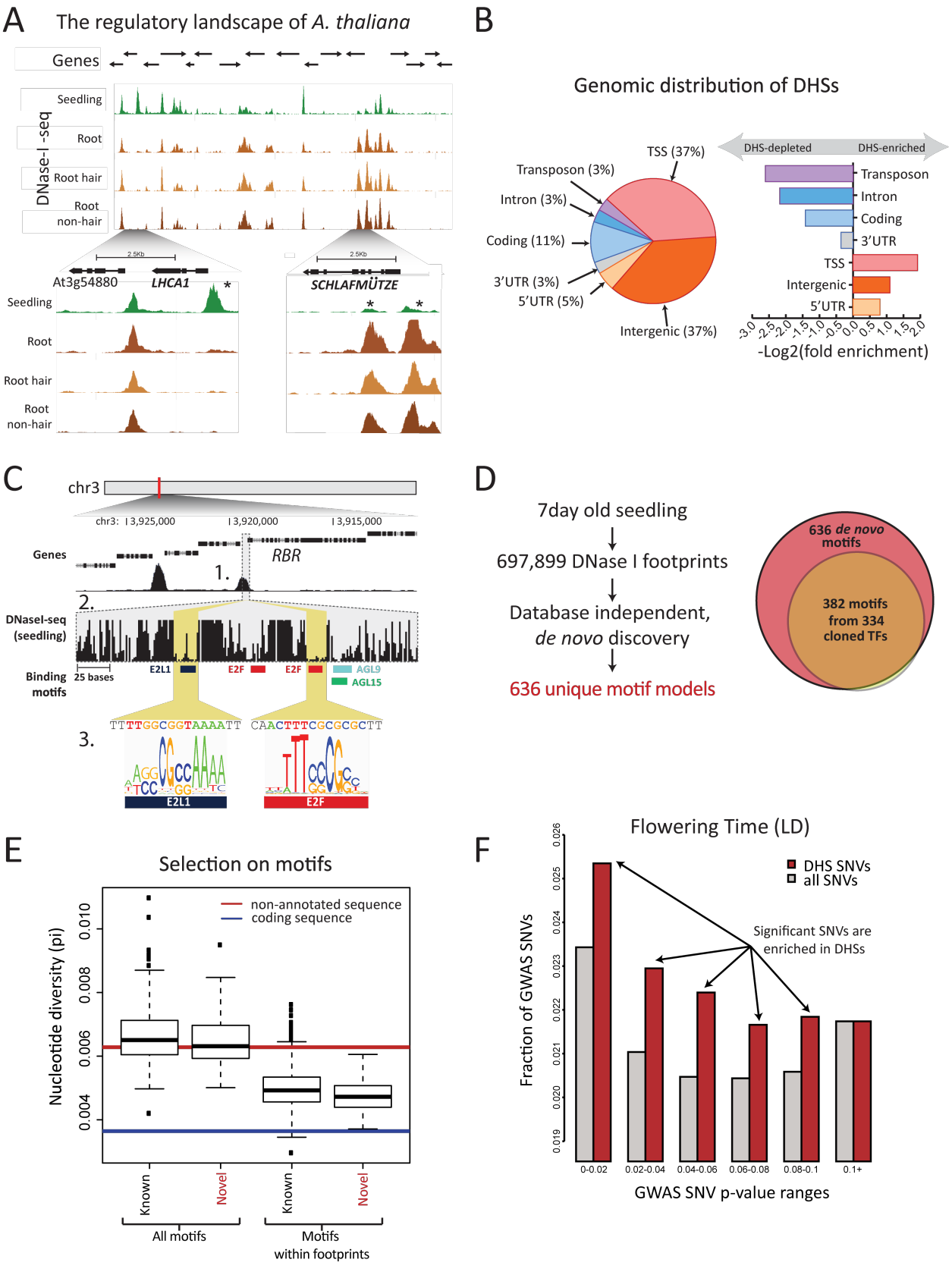

A
The regulatory landscape of A. thaliana
Genes
DNase-I -seq
Seedling
Root
Root hair
Root non-hair
2.5Kb
At3g54880
LHCA1
SCHLAFMÜTZE
B
Genomic distribution of DHSs
TSS (37%)
Transposon (3%)
Intron (3%)
Coding (11%)
3'UTR (3%)
5'UTR (5%)
Intergenic (37%)
DHS-depleted
DHS-enriched
Transposon
Intron
Coding
3'UTR
TSS
Intergenic
5'UTR
-Log2(fold enrichment)
C
chr3
chr3: 13,925,000
13,920,000
13,915,000
Genes
RBR
1.
2.
DNaseI-seq (seedling)
Binding motifs
25 bases
E2L1
E2F
E2F
AGL9
AGL15
3.
TTTTGGCGGTAAAATT
CAACTTTCGCGCGCTT
E2L1
E2F
D
7day old seedling
697,899 DNase I footprints
Database independent, de novo discovery
636 unique motif models
636 de novo motifs
382 motifs from 334 cloned TFs
E
Selection on motifs
non-annotated sequence
coding sequence
Nucleotide diversity (pi)
0.004
0.006
0.008
0.010
Known
Novel
Known
Novel
All motifs
Motifs within footprints
F
Flowering Time (LD)
DHS SNVs
all SNVs
Significant SNVs are enriched in DHSs
Fraction of GWAS SNVs
0.019
0.020
0.021
0.022
0.023
0.024
0.025
0.026
0-0.02
0.02-0.04
0.04-0.06
0.06-0.08
0.08-0.1
0.1+
GWAS SNV p-value ranges

Figure 1. Nucleotide-resolution mapping of regulatory DNA enabled discovery of TF footprints and *de novo* TF motifs.

A, DNase I hypersensitivity (read-depth normalized density tracks) in whole seedling, seedling root tissue, and two root epidermal cell types within a representative 100 kb region of chromosome 3. Tissue-specific DNase I hypersensitive sites (DHSs) resided near the light- and flowering time-associated genes *LHCA1* and *SCHLAFMÜTZE* (marked by asterisk *). **B,** DHSs disproportionately resided in intergenic, TSS, and 5'UTR elements. **C,** Representative example of footprints with TF motifs. Shown is a chromosome 3 region with tracks denoting (1) a DHS in the *RBR* promoter, (2) per-base cleavage in the *RBR* DHS with bars indicating TF motifs, and (3) footprints containing the motifs E2L1 and E2F. **D,** *De novo* motif discovery yielded 636 motifs. We validated these by comparing them to 382 protein binding microarray-derived motif models (**Table S6)** considering only the best *de novo* motif match. Three hundred and sixty six (96%) of the 382 protein binding microarray-derived motif models matched at least one *de novo* motif. **E**, Nucleotide diversity was similar for *de novo* motifs that match known and novel motifs. Red line is nucleotide diversity of non-annotated, presumably neutrally evolving DNA. Blue line is nucleotide diversity (pi) of coding regions, which mostly evolve under purifying selection. Motifs in footprints showed substantially reduced diversity compared with all matches genome-wide. **F,** Highly significant GWAS SNVs were significantly enriched in DHSs (*p value 0.00153; K-S test) for the GWAS phenotype flowering time (LD).

TF occupancy of *A. thaliana* protein-coding exons may modulate codon choice

In mammalian genomes, transcription factor occupancy within protein-coding exons may modulate codon choice and protein evolution (Stergachis et al., 2013a). The generality of this phenomenon across kingdoms is unknown. Overall, 14% of *A. thaliana* footprints localized within protein-coding exons (**Figure 2A-B**). The specific codons that were preferentially contained in TF footprints differed substantially between *A. thaliana* and human (e.g., Lys AAA) (**Figure 2C-E**). Changes in TF binding preferences between *A. thaliana* and human correlated with directional codon biases (r=0.61) (**Figure 2E, F**). For example, the leucine-encoding CTG was preferentially bound in human compared to *A. thaliana*; this codon is far more frequently utilized in human coding regions (**Figure 2E**). In human, stop codon trinucleotides (TAA, TAG, TGA) are significantly depleted from DNaseI footprints genome-wide (Stergachis et al., 2013a). We found an analogous situation (**Figure 2G**), indicating that the *A. thaliana* TF repertoire has likewise been depleted of DNA binding domains capable of recognizing stop codons.

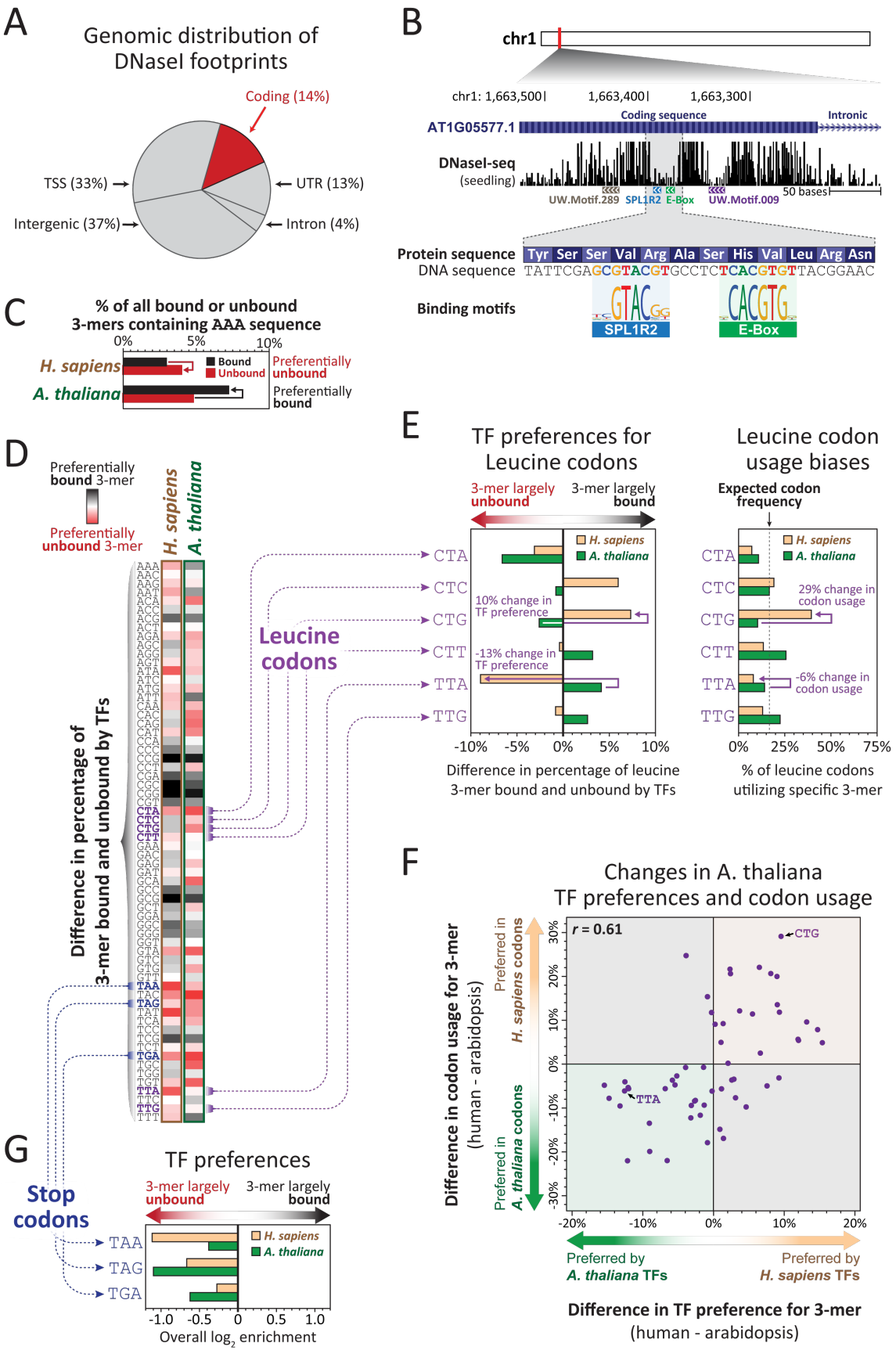

A
Genomic distribution of DNaseI footprints
Coding (14%)
TSS (33%)
UTR (13%)
Intergenic (37%)
Intron (4%)
B
chr1
chr1: 1,663,500
1,663,400
1,663,300
Coding sequence
Intronic
AT1G05577.1
DNaseI-seq
(seedling)
50 bases
UW.Motif.289
SPL1R2
E-Box
UW.Motif.009
Protein sequence
Tyr Ser Ser Val Arg Ala Ser His Val Leu Arg Asn
DNA sequence
TATTCGAGCGTACGTGCCTCTCACGTGTTACGGAAC
Binding motifs
SPL1R2
E-Box
C
% of all bound or unbound 3-mers containing AAA sequence
0%
5%
10%
H. sapiens
A. thaliana
Bound
Unbound
Preferentially unbound
Preferentially bound
D
Preferentially bound 3-mer
Preferentially unbound 3-mer
H. sapiens
A. thaliana
Difference in percentage of 3-mer bound and unbound by TFs
Leucine codons
CTA
CTC
CTG
CTT
TTA
TTG
E
TF preferences for Leucine codons
3-mer largely unbound
3-mer largely bound
H. sapiens
A. thaliana
10% change in TF preference
-13% change in TF preference
-10%
-5%
0%
5%
10%
Difference in percentage of leucine 3-mer bound and unbound by TFs
Leucine codon usage biases
Expected codon frequency
29% change in codon usage
-6% change in codon usage
0%
25%
50%
75%
% of leucine codons utilizing specific 3-mer
F
Changes in A. thaliana TF preferences and codon usage
r = 0.61
CTG
TTA
Difference in codon usage for 3-mer (human - arabidopsis)
Preferred in H. sapiens codons
Preferred in A. thaliana codons
30%
20%
10%
0%
-10%
-20%
-30%
-20%
-10%
0%
10%
20%
Preferred by A. thaliana TFs
Preferred by H. sapiens TFs
Difference in TF preference for 3-mer (human - arabidopsis)
G
Stop codons
TF preferences
3-mer largely unbound
3-mer largely bound
H. sapiens
A. thaliana
TAA
TAG
TGA
-1.0
-0.5
0
0.5
1.0
Overall log2 enrichment

Figure 2. TF codon binding preferences correlated with organismal codon bias.
A, Fourteen percent of footprints resided in coding regions. **B,** A representative example of footprints in a coding region (AT1G05577) on chromosome 1. **C,** Example of a 3-mer (AAA, encoding lysine) that was preferentially bound by TFs (i.e. resided within footprints, red) in *A. thaliana* but not in human. **D,** TF binding preferences in *A. thaliana* and human were calculated for all codons, revealing both similarities and differences in TF binding preferences between *A. thaliana* and human. **E,** TF binding preferences for leucine codons (left) correlated with differences in *A. thaliana* and human codon usage (right). **F,** Usage differences for all codons were strongly correlated with differences in TF binding preferences. **G,** Stop codons were consistently unbound in *A. thaliana* (green) and human (light-brown).

Regulatory DNA landscape dynamics during photomorphogenesis

In seedlings, light triggers photomorphogensis, a fundamental and irreversible re-shaping of plant form and metabolism to optimize photosynthesis. Underlying this transition to photoautotrophy is a wave of transcriptional reprogramming and alteration in gross chromatin compaction (van Zanten et al., 2012), during which expression levels of nearly one third of all *A. thaliana* transcripts are altered (Ma et al., 2001).

To analyze the regulatory DNA landscape of photomorphogenesis, we exposed dark-grown seedlings to 0, 0.5, 3 or 24 hours of light (LD conditions) and performed DHS mapping and genomic footprinting at each time point. We identified 734 photodynamic DHSs across these conditions (**Figure 3; Figure S2A-D, Table S8**), which clustered into five distinct DHS accessibility patterns (**Figure 3B**). Many DHSs within the five clusters resided in proximity to genes previously implicated in the light response (**Figure 3C**, panels 3-5, 9). For example, a DHS overlying the promoter of *HY5 HOMOLOG* (*HYH*), a key regulator of the light response (Brown and Jenkins, 2008), reached peak activity after 24 hours of light (cluster V) (**Figure 3C**, panel 5). Numerous photodynamic DHSs were localized near genes with previously uncharacterized roles in photomorphogenesis (**Figure 3C** panels 1, 2, 6-8, 10; **Figure S2D** panels 1-5**; Table S8**). For example, several members of the *SAUR* gene family including *SAUR24* (**Figure 3C**, panel 1), contained dark-activated DHS in their promoters. *SAUR* genes play critical roles in cell expansion and transport of the plant hormone auxin but their role in the dark is unknown (Spartz et al., 2012).

Gene ontology (GO) analysis of the genes proximal to photodynamic DHSs highlighted specific biological processes associated with dark- and light-activated DHSs (**Table S9**). Genes enriched for light stimulus and response to UV were associated with 3hr

and 24hr light-activated DHSs (Cluster IV), and genes enriched for response to auxin and shade avoidance were associated with dark-activated DHSs (Cluster I).

To explore the TFs mediating photodynamic DHSs, we compared the densities of recognition sequences for known TFs among DHS clusters (**Figure 3D** and **Table S8**). The light-activated DHS clusters III, IV, and V contained a high density of recognition sequences for the photomorphogenesis master regulator HY5 (Jiao et al., 2007) relative to dark-activated and 30min light-activated DHS clusters (**Figure 3D**). By contrast, the light-activated DHS clusters III and IV were densely populated with PIF1 and PIF3 motifs relative to the other clusters (**Figure 3D**). Members of the PIF gene family control seedling growth (Leivar and Quail, 2011), and the quadruple *pif1 pif3 pif4 pif5* mutant displays a constitutive photomorphogenic phenotype (Shin et al., 2009). The dark-activated DHS cluster I contained a high density of motifs for A. THALIANA RESPONSE REGULATOR 10 (ARR10), which regulates the cytokinin response (Mason et al., 2005)(**Figure 3D**). The cytokinin class of plant hormones is implicated in cell division, shoot initiation and growth, leaf senescence, and photomorphogenic development (Mok and Mok, 1994).

To identify which TFs distinguish photodynamic DHSs from the rest of the *cis*-regulatory landscape, we analyzed enrichment of TF recognition sequences (including novel motifs) within each cluster of photodynamic DHSs relative to all seedling DHSs, the vast majority of which were static (**Figure S2F, Table S8**). For the dark-specific cluster I, this analysis yielded a striking enrichment for novel motifs in addition to several ARR factors, including ARR10, and three homeobox factors (**Figure S2F**). Despite the fact that TFs are often lowly expressed and thus difficult to evaluate, all nine *A. thaliana* TFs with dark-enriched motifs are expressed in the dark, and six of nine, including *ARR1* and *ARR2*,

show comparable or higher expression in dark relative to light conditions (Diurnal expression browser: http://diurnal.mocklerlab.org) (Michael et al., 2008). The density of TF recognition sequences encoded within light-activated DHSs did not differ greatly from all seedling DHSs (**Figure S2F**). Indeed, when compared to the genome, recognition sequences for light-related TFs (including HY5 and PIFs) are enriched within DHSs (**Figure S2G**). This enrichment of light-related TF recognition motifs in all DHSs and the prevalence of static DHSs suggests that plant chromatin is poised for light.

Taken together, our results indicate that photodynamic DHSs are programmed in an exposure time-dependent fashion, which is achieved by a specific temporally-coordinated set of TFs (**Figure 3**).

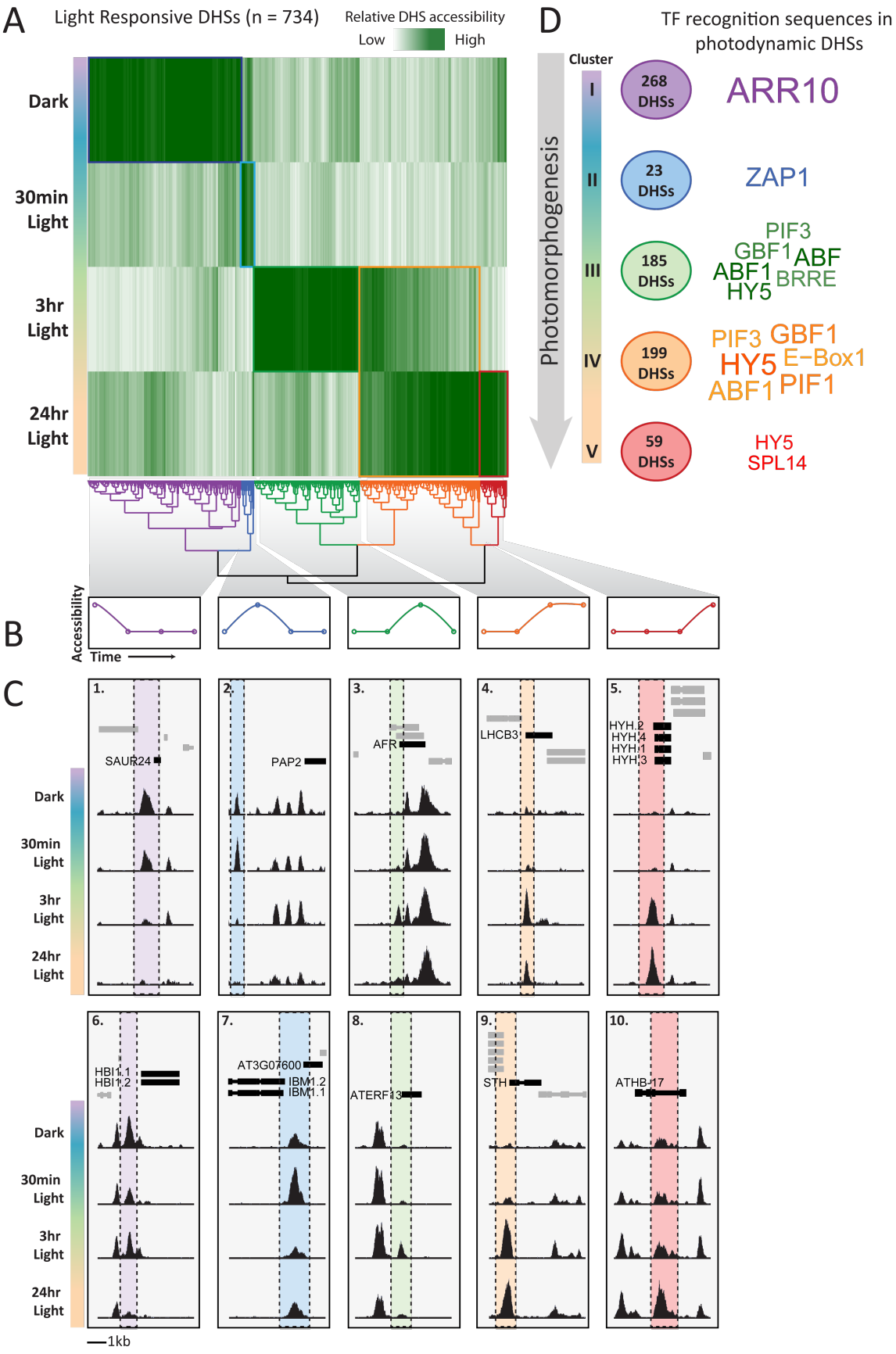

A
Light Responsive DHSs (n = 734)
Relative DHS accessibility
Low
High
Dark
30min Light
3hr Light
24hr Light
B
Accessibility
Time
C
1.
SAUR24
2.
PAP2
3.
AFR
4.
LHCB3
5.
HYH.2
HYH.4
HYH.1
HYH.3
6.
HBI1.1
HBI1.2
7.
AT3G07600
IBM1.2
IBM1.1
8.
ATERF13
9.
STH
10.
ATHB-17
1kb
D
TF recognition sequences in photodynamic DHSs
Photomorphogenesis
Cluster
I
II
III
IV
V
268 DHSs
ARR10
23 DHSs
ZAP1
185 DHSs
PIF3
GBF1
ABF
ABF1
BRRE
HY5
199 DHSs
PIF3
GBF1
HY5
E-Box1
ABF1
PIF1
59 DHSs
HY5
SPL14

Figure 3. Dynamic chromatin changes during photomorphogenesis.
A, DHSs identified for each light treatment were clustered, yielding five condition-specific DHS clusters: (I, purple) DNase I-accessible in dark, (II, blue) accessible in 30 minutes light, (III, green) accessible in 3hr light, (IV, orange) accessible in 3hr and 24hr light, (V, red) accessible in 24hr light. **B,** Characteristic patterns of DNase I-accessibility. **C,** Representative examples of DHSs (1-10) from each condition-specific DHS cluster were located near known and novel photomorphogenesis genes. Each window is 5kb; vertical ranges vary, but are consistent for a given DHS example, highlighting fold change rather than absolute differences. **D**, TRANSFAC motif densities relative to background within each cluster are represented as a word cloud.

Empirical TF-to-TF networks in *A. thaliana*

Genomic footprinting enables systematic analysis of TF occupancy within regulatory DNA of transcription factor genes, providing a direct and empirical approach for mapping cross-regulatory interactions (edges) between TF genes (nodes) (**Figure 4A**). Systematically applying this approach to all TFs with defined recognition sequences enables the construction of TF-to-TF networks, and analysis of their organization, dynamics, and architectural features (Neph et al., 2012b). This approach recapitulates validated connections, provides visualizable and interpretable information, is agnostic with respect to positive or negative interactions, and accounts for redundant recognition motifs among TFs.

To create large-scale TF regulatory networks for *A. thaliana* we iterated the approach of Neph et al., 2012a over 251 TFs with defined recognition sequences (from TRANSFAC, JASPAR, and PBM data). For conservative assignment of regulatory DNA to specific genes, we considered only proximal regulatory DNA (footprints occurring within 500bp upstream of the TSS and extending over the gene body). This resulted in a network comprising 7,662 edges connecting 251 TF nodes (average of 31 edges per node). Networks are available at plantregulome.org.

Architecture of the *A. thaliana* TF network

Biological networks are comprised of simple 3-node network motifs that are universal, and finite in number (n=13)(Milo et al., 2004). The relative frequency of each 3-node motif can be used to compare the topology of diverse biological networks. Analysis of network architecture is agnostic to the connection sign; *i.e.* any connection may be negative

or positive, or both, depending on conditions. The central parameter is the connection direction ($TF_A \rightarrow TF_B$).

To analyze the architecture of the *A. thaliana* TF network, we computed the frequencies and relative enrichments of all thirteen 3-node network motifs within the network. We then compared these frequencies to other biological networks, including the human TF network and the *C. elegans* neuronal network. In spite of its highly diverged *cis*- and *trans*-regulatory repertoire, this analysis revealed a highly similar topology for the *A. thaliana* TF network (**Figure 4B**).

Light-induced changes in TF networks

Across the tested light conditions, total TF network size ranged from 1,340 regulatory edges in the dark to 1,930 edges after 3 hours of light.

Autoregulatory loops are a well-established mechanism for re-enforcing and fine-tuning gene expression patterns during developmental transitions (Crews and Pearson, 2009). As photomorphogenesis is one of the major developmental transitions in plant life, we reasoned that it should result in re-wiring of autoregulatory loops.

We detected the appearance and disappearance of several known or posited autoregulatory loops for key photomorphogenetic factors in response to light. For example, an *EPR1* auto-regulatory loop appeared with increasing exposure to light (**Figure 4C**). *EPR1*, which is regulated by HY5 in the light, represses expression of its endogenous copy when overexpressed (Kuno et al., 2003; Li et al., 2011). We also observed the disappearance of a *MYC2* auto-regulatory feedback loop upon 24hrs of light (**Figure 4C**), consistent with a negative feedback loop (Dombrecht et al., 2007). MYC2, a master

regulator of growth that integrates light cues and plant hormone signaling pathways, acts as a negative regulator of photomorphogensis (Yadav et al., 2005). By contrast, ABI5, a general integrator of light and abscisic acid signaling (Chen et al., 2008a), appeared to be constitutively auto-regulated (**Figure 4C**).

We examined first-degree connections between *HY5* and several TFs implicated in photomorphogenesis (**Figure 4D, Table S10**). *HY5* exhibited many connections that remain stable across all conditions (**Figure 4D, grey lines**), in addition to dynamic connections involving known or putative photomorphogenetic factors (**Figure 4D, black lines**). For example, after 30 minutes of light, we detected *HY5* occupancy at *EIN3* and *ZAT10*, which are involved, respectively, in greening (Zhong et al., 2009) and photo-oxidative and other abiotic stresses (Rossel et al., 2007). We also identified a connection across all conditions between HY5 and *ABI5* that recapitulates genetic and biochemical evidence (Chen et al., 2008b).

We further analyzed 35 TFs with light-related GO-annotations, revealing a highly interconnected sub-network (**Figure 4E**), comprising 108 regulatory edges, significantly more than expected from randomly-selected network TFs ($p=0.006$). These 35 TFs were far more interconnected under light vs. dark conditions (red vs. blue lines, **Figure 4E, F**).

Together, these results enable the *de novo* identification of photomorphogenesis regulators that are dynamically rewired in response to light, and previously unappreciated relationships among light-regulated TFs.

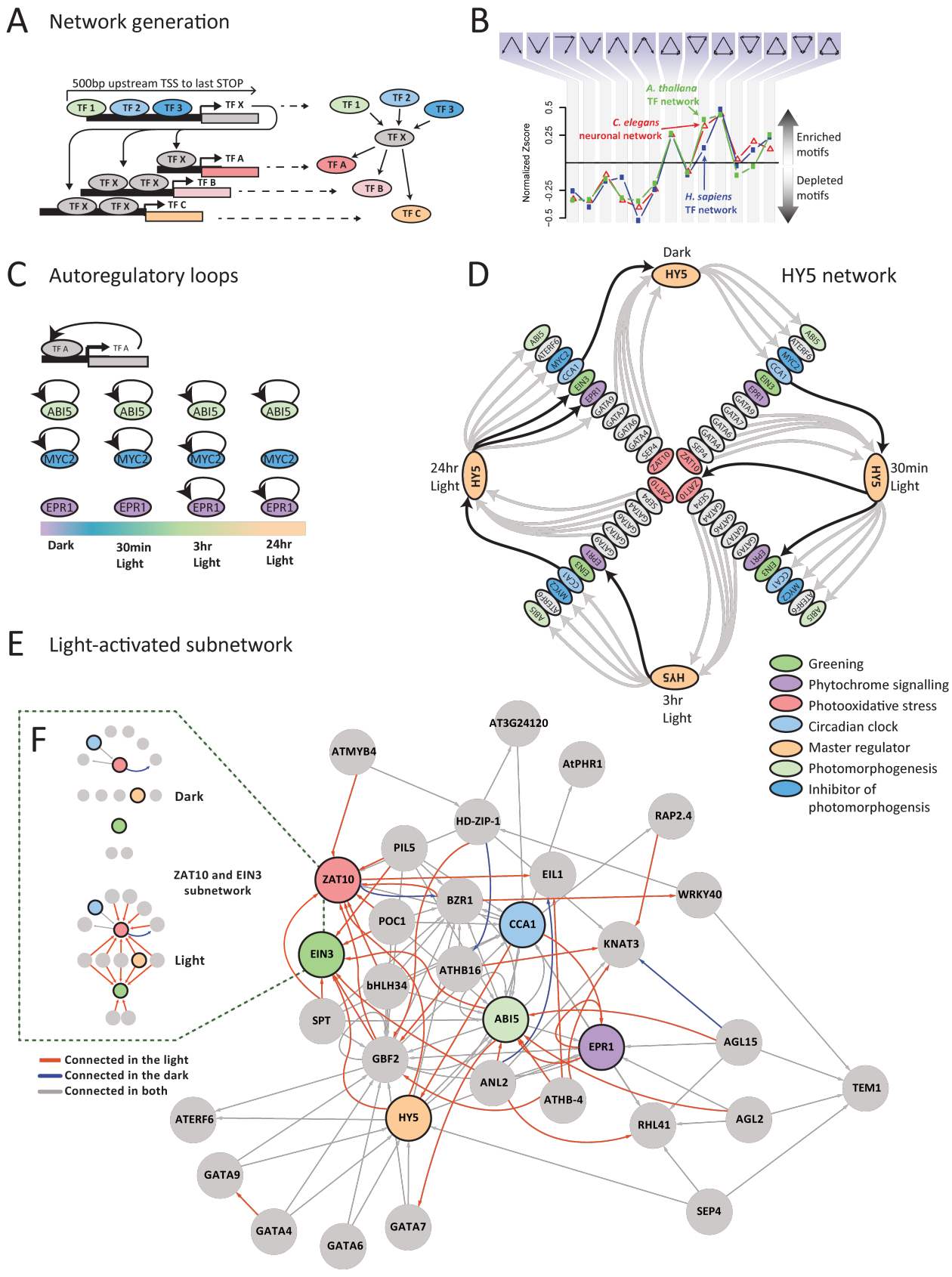

A Network generation
500bp upstream TSS to last STOP
TF 1
TF 2
TF 3
TF X
TF A
TF B
TF C
B
A. thaliana TF network
C. elegans neuronal network
H. sapiens TF network
Normalized Zscore
Enriched motifs
Depleted motifs
C Autoregulatory loops
ABI5
MYC2
EPR1
Dark
30min Light
3hr Light
24hr Light
D
HY5 network
Dark
24hr Light
30min Light
3hr Light
E Light-activated subnetwork
Greening
Phytochrome signalling
Photooxidative stress
Circadian clock
Master regulator
Photomorphogenesis
Inhibitor of photomorphogenesis
F
Dark
Light
ZAT10 and EIN3 subnetwork
Connected in the light
Connected in the dark
Connected in both
AT3G24120
ATMYB4
AtPHR1
HD-ZIP-1
RAP2.4
PIL5
ZAT10
EIL1
WRKY40
BZR1
POC1
CCA1
KNAT3
EIN3
ATHB16
bHLH34
SPT
ABI5
EPR1
AGL15
GBF2
ANL2
ATHB-4
TEM1
ATERF6
HY5
RHL41
AGL2
GATA9
SEP4
GATA4
GATA7
GATA6

Figure 4. TF networks were re-wired during photomorphogenesis.
A, To build networks, an edge was created when a footprinted motif of a source TF overlapped a target TF gene, including 500bp upstream of the target TF's transcription start site (500bp was chosen based on the observation that DHSs are highly enriched near the TSS). Target TF X in grey, TFs with regulatory input into TF X in shades of green and blue, output TFs regulated by TF X in shades of orange and red. **B,** Network motif topology in *A. thaliana* was similar to the previously described *C. elegans* neuronal network and human TF network (Milo et al., 2004; Neph et al., 2012b). **C,** Representative examples of auto-regulatory loops and their dynamics in response to light. **D,** HY5 regulation with respect to selected light-related TFs (oval color denotes functional annotation) changed in response to light (black edges), however, most edges stayed constant (grey). Converging arrows represent multiple motifs found within a single footprint, which occurs when related factors can occupy the same motif within a footprint, or when a single footprint spans multiple unrelated motifs. In both cases we consider the underlying DNA to have high-regulatory potential. **E,** Upon light exposure**,** light-related factors increased in connectivity (grey ovals, otherwise colored ovals denote functional annotation, see D). Regulatory connections observed in any of the light conditions but not in the dark are in red, regulatory connections observed in the dark and not in any light condition are in blue. Many regulatory connections were constitutive across all four conditions (grey). **F,** TFs related to greening (EIN3, green) and photo-oxidative stress (ZAT10, red) became highly connected in response to light and shared five regulators, including HY5 (orange).

TF-specific photomodulated connectivity patterns

To gain further insight into the connectivity patterns of highly connected photomodulated TFs (EIN3, MYC2, ABI5, HY5), we tallied all connections for each TF in the network (input and output edges, auto-regulatory and bi-directional loops) and differentiated common vs. dynamic edges (**Figure 5A**). This analysis revealed marked differences between TFs with similar baseline connectivity (**Figure 5A**). For example, EIN3 transitioned from a relatively unconnected dark state to a stable, highly connected light state driven chiefly by increased input edges (**Figure 5B** column 1). By contrast, MYC2 showed the reverse pattern (**Figure 5B** column 2). *ABI5* underwent a progressive increase in input edges with light exposure (**Figure 5B** column 3), whereas the high connectivity of *HY5* derived chiefly from common edges (**Figure 5B** column 4).

To determine how the connectivity of a given factor was propagated to network neighbors, we compared each factor's first degree regulatory relationships (**Figure 5C,** orange edges) and, in turn, the first degree regulatory relationships shared among its first degree neighbors (**Figure 5C,** grey edges), and then sorted all TFs by their overall degree of connectedness in the entire network (**Figure 5C**). Qualitative differences again emerged between highly connected TFs. For example, following 30min of light, EIN3 and its first degree neighbors became more interconnected (**Figure 5D**, column 1). By contrast, MYC2 showed initially increased connectivity with first-degree neighbors, yet almost all these interactions were lost in response to 24hrs light (**Figure 5D**, column 2). Our detailed network analysis revealed striking differences among major light-related TFs with regard to their connectivity patterns across conditions and network neighborhood.

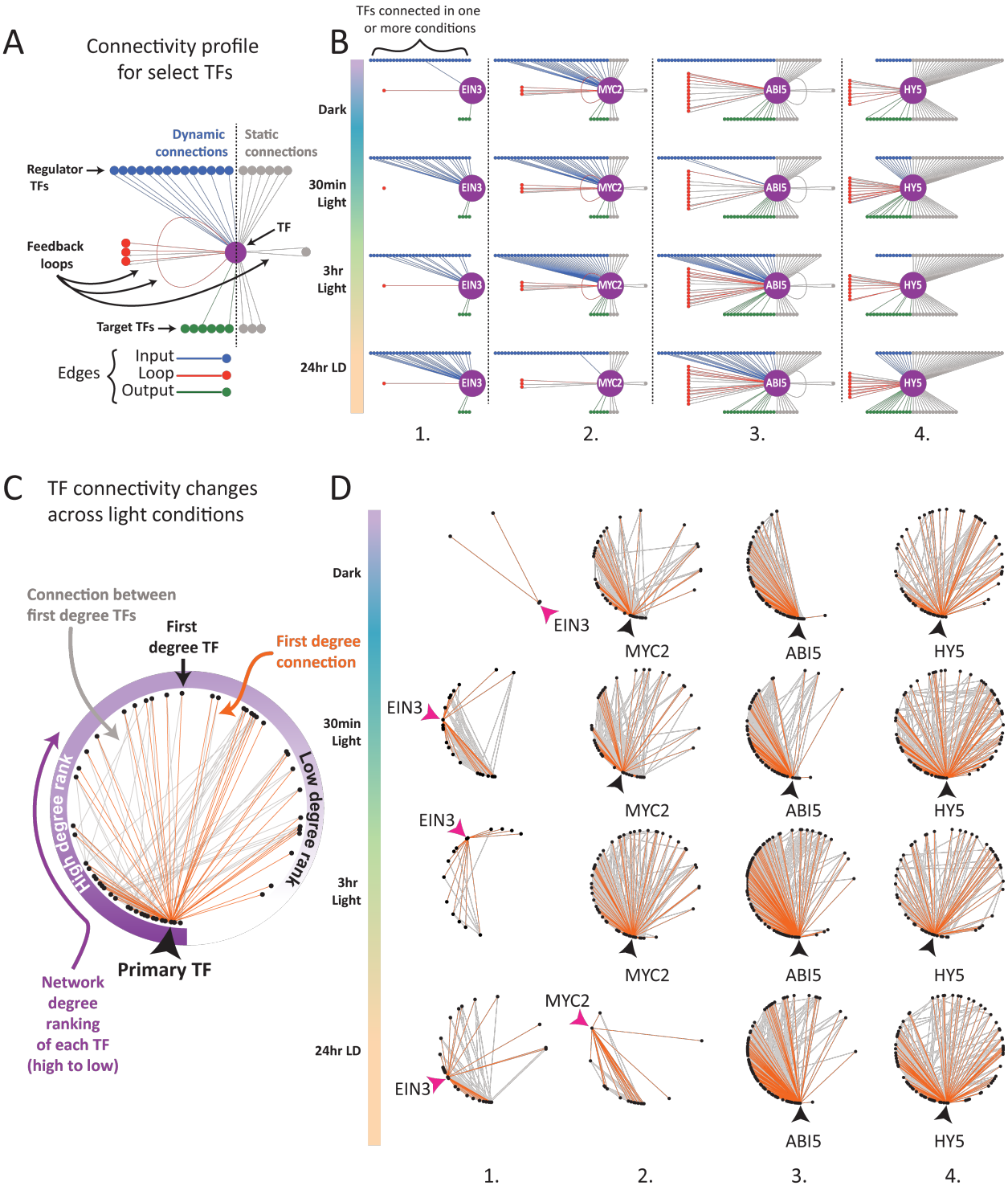

A
Connectivity profile for select TFs
Dynamic connections
Static connections
Regulator TFs
TF
Feedback loops
Target TFs
Edges
Input
Loop
Output
B
TFs connected in one or more conditions
Dark
30min Light
3hr Light
24hr LD
EIN3
MYC2
ABI5
HY5
1.
2.
3.
4.
C
TF connectivity changes across light conditions
Connection between first degree TFs
First degree TF
First degree connection
High degree rank
Low degree rank
Primary TF
Network degree ranking of each TF (high to low)
D
Dark
30min Light
3hr Light
24hr LD
EIN3
MYC2
ABI5
HY5
1.
2.
3.
4.

Figure 5. TF network re-wiring followed complex patterns during photomorphogenesis.
A, For representative TFs, connectivity was dissected into input edges (regulators, top), output edges (targets, bottom), bi-directional loops (condition-specific, left; common across conditions, right), and auto-regulatory loops, revealing regulatory differences despite similarities in overall interaction degree (e.g., ABI5 and HY5). **B,** Connectivity changed across light conditions for key TFs: EIN3, MYC2, ABI5, and HY5. **C,** First degree neighborhoods of the same TFs visualize how a TF's connectivity percolated through the larger network. In each circular network, TFs were sorted by number of regulatory connections, such that connectivity increases clock-wise around the circle starting from the bottom. For each TF, only edges between TF and first degree neighbors (orange) and the edges between first degree neighbors (grey) are shown. Factors with large condition-specific changes in connectivity (EIN3, MYC2 at 24hr LD) are demarcated by pink arrows. **D,** Neighborhood connectivity changed across light conditions for key TFs: EIN3, MYC2, ABI5, and HY5.

Impact of heat shock on the regulatory DNA landscape

Exposure to heat triggers a conserved response involving the rapid up-regulation of heat shock proteins (HSPs), accompanied by down-regulation of many normally active genes and the inhibition of most protein translation (Lindquist, 1986). Although heat shock has been studied for decades in other organisms little is known about the impact of heat shock on plant chromatin and transcriptional regulatory pathways outside of studies of the heat shock transcription factors (HSFs)(Scharf et al., 2012) and the nucleosome variant H2A.Z in the response to ambient temperature change (Kumar and Wigge, 2010).

To define the effects of heat on the chromatin landscape and TF regulatory network, we mapped DHSs and TF footprints in control and heat-treated seven-day-old seedlings (**Methods**). We focused on the most extreme (top and bottom 2.5%) heat-activated or heat-repressed DHSs (**Figure 6A, Figure S3A, B, Table S11**). This approach identified equal numbers of strongly heat-activated and heat-repressed DHSs (n= 990, 1,980 total) but fold differences varied. The genomic distributions of heat-activated vs. -repressed DHSs also differed markedly, with the former localizing in distal intergenic regions, and the latter localizing primarily in gene-proximal regions (TSS, 5'UTR, 3'UTR, intron, coding)(**Figure 6B, Table S11**).

A subset of genes with extreme heat-induced DNaseI accessibility

A small fraction (14.6%, n=145) of heat-activated DHSs displayed extreme accessibility (**Figure 6A**), and were concentrated within the bodies of 63 genes encoding canonical heat shock proteins, their co-chaperones, and several heat stress-related TFs, in addition to novel heat shock-responsive genes (**Figure 6A-D; Table S11**). Above a z-score

of 12 (dotted line), all extremely accessible genes were unique to heat shock (**Figure 6D**, red dots).

The most highly accessible gene was *HSP101*, which is crucial for acquired thermotolerance (Queitsch et al., 2000), followed by the genes encoding the heat-inducible HSP90.1 chaperone and HOP3, an important co-chaperone of HSP90 (Krishna and Gloor, 2001) (**Figure 6A, D**). This extreme DNase I accessibility was unique to specific members of gene families, presumably reflecting their functional specialization in the acute heat shock response.

Genes with extreme accessibility displayed 'poised' promoters (Keene et al., 1981) in control conditions (**Figure 6C**) and tended to be highly expressed upon heat shock (**Figure S3C**). They were even more significantly enriched for the GO term "response to heat" (p-value 4.11×10^{-71}) than the genes associated with heat-activated DHSs (p-value 6.45×10^{-13}) (**Figure S3D**). In contrast, heat-repressed DHSs were enriched near growth, transport, and metabolic genes **(Figure S3D),** consistent with down-regulation of these energy-requiring cellular functions.

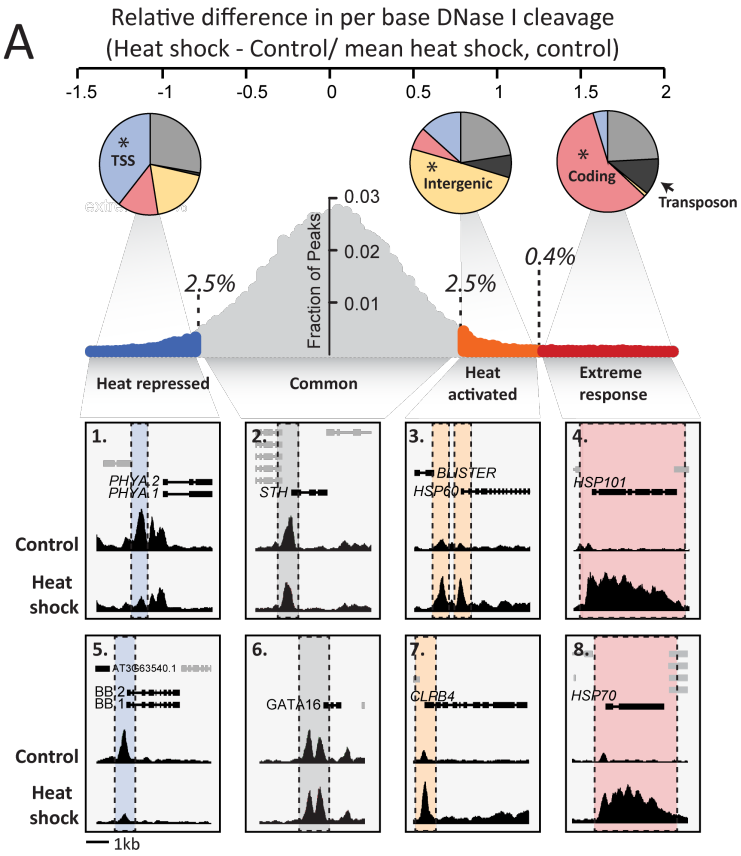
Taken together, our results identify genes displaying a unique chromatin signature with poised promoters in control conditions and extreme accessibility in response to heat shock. Given the outsize importance of heat tolerance in today's agriculture, these genes may be of direct relevance for genetic engineering.

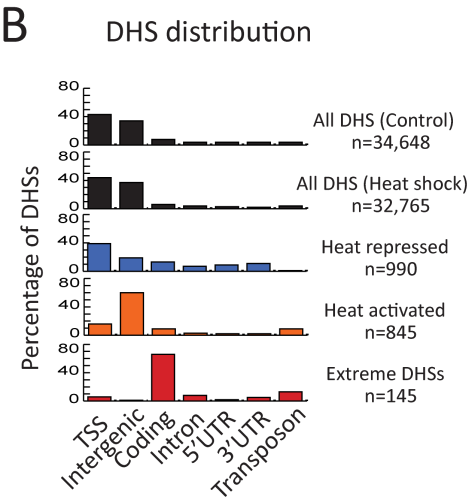
TF drivers of heat-activated DHSs

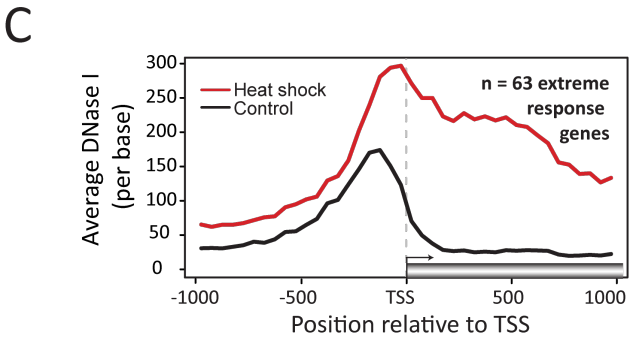
To identify TF drivers of the regulatory changes accompanying heat shock, we analyzed the TF recognition site repertoire of heat-responsive DHSs. The Heat Shock

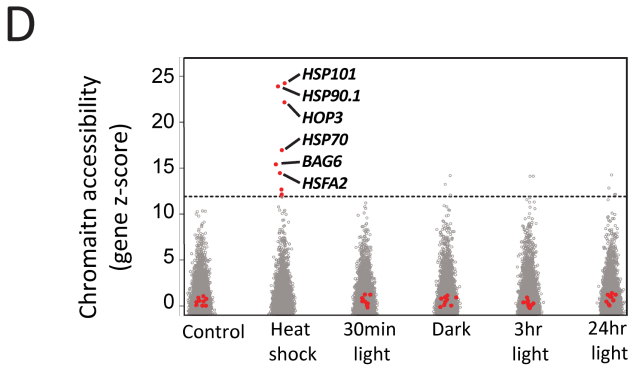
Element (HSE, AGAAnnTTCT) was highly enriched in heat-activated DHSs and the promoters of extreme-accessibility genes **(Figure 6E; Figure S3E; Table S11)**. We also detected complex, heat-associated footprinting patterns at partial HSEs. For example, the promoter of *ZAT10*, a factor involved in response to abiotic stresses that is activated by HSFA2 in a heat shock-specific manner (Schramm et al., 2006), contains three adjacent partial HSEs, all of which showed footprints in control conditions (**Figure 6F)**. In response to heat, these footprints merged into a single larger footprint, suggesting either that the resident HSF (presumably HSF2A) greatly increases occupancy time, or that it partners with additional regulatory factors.

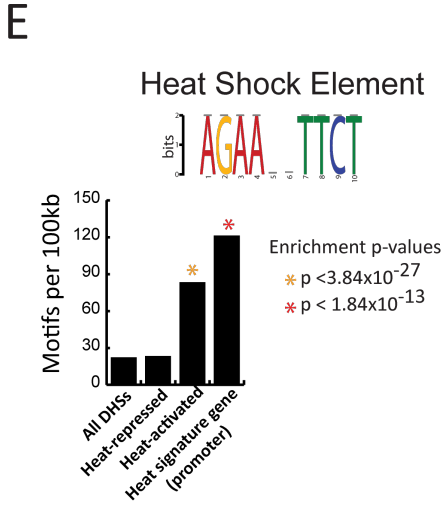
Unexpectedly, MADS box motifs, such as the recognition sequence for AGL9 (SEP3), were also enriched in heat-activated DHSs **(Figure S3E; Table S11)**. Both the HSE and MADS box motifs (CArG-box) were selective for heat-activated elements, present among the top 25 over-represented footprint-derived motifs in heat-activated DHSs, but absent from the top 25 footprint-derived over-represented motifs in heat-repressed DHSs. Similar to static DHSs during photomorphogensis, static DHSs during heat shock did not contain vastly different TF recognition sequence compositions from all seedling DHSs; the few signficant differences were of small magnitude (**Figure S3C**). Light-related TF motifs (HY5 and PIFs), which were generally pervasive in *cis*-regulatory DNA (**Figure S2G**), were among the motifs significantly enriched in static DHSs and depleted from heat-dynamic DHSs. Collectively, our findings re-affirm the importance of HSFs in driving the heat shock response and suggest a prominent role for MADS-box factors in remodeling the chromatin and regulatory landscape in response to heat shock.


A
Relative difference in per base DNase I cleavage
(Heat shock - Control/ mean heat shock, control)
-1.5
-1
-0.5
0
0.5
1
1.5
2
TSS
Intergenic
Coding
Transposon
0.03
0.02
0.01
Fraction of Peaks
2.5%
2.5%
0.4%
Heat repressed
Common
Heat activated
Extreme response
1.
PHYA 2
PHYA 1
2.
STH
3.
BLISTER
HSP60
4.
HSP101
Control
Heat shock
5.
AT3G63540.1
BB 2
BB 1
6.
GATA16
7.
CLPB4
8.
HSP70
Control
Heat shock
1kb


B
DHS distribution
Percentage of DHSs
All DHS (Control)
n=34,648
All DHS (Heat shock)
n=32,765
Heat repressed
n=990
Heat activated
n=845
Extreme DHSs
n=145
TSS
Intergenic
Coding
Intron
5'UTR
3'UTR
Transposon


C
Average DNase I (per base)
Heat shock
Control
n = 63 extreme response genes
Position relative to TSS
-1000
-500
TSS
500
1000


D
Chromaitn accessibility (gene z-score)
HSP101
HSP90.1
HOP3
HSP70
BAG6
HSFA2
Control
Heat shock
30min light
Dark
3hr light
24hr light


E
Heat Shock Element
bits
Motifs per 100kb
Enrichment p-values
p <3.84x10-27
p < 1.84x10-13
All DHSs
Heat-repressed
Heat-activated
Heat signature gene (promoter)

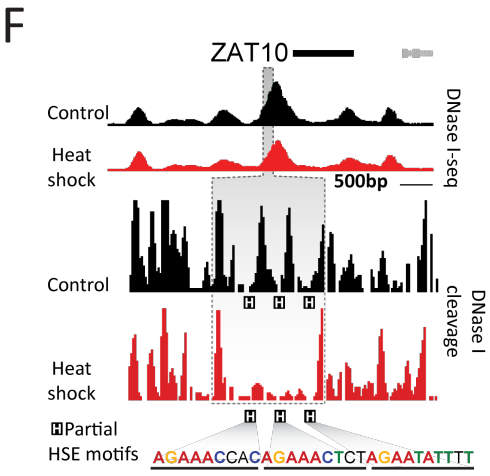

F
ZAT10
Control
Heat shock
DNase I-seq
500bp
Control
Heat shock
DNase I cleavage
Partial HSE motifs
AGAAACCACAGAAACTCTAGAATATTTT

Figure 6. Dynamic chromatin changes in response to heat shock.
A, Relative difference in DHSs between control and heatshock (**Supplemental Methods**) The most heat shock-responsive DHSs were designated as heat-activated (top 2.5%) and heat-repressed (bottom 2.5%) DHSs. The heat-activated DHSs (orange) showed a long tail (red). The most extreme heat-activated DHSs (red) encompassed extremely accessible heat-induced genes. Pie charts reflect the genomic distributions of DHS type (e.g., heat activated), asterisks (*) denote the genomic feature, in which the greatest proportion of DHSs resided relative to other genomic features. **B**, Genomic DHS distributions were similar for control and heat shock (black), however, among heat-repressed DHSs (blue), heat-activated DHSs (orange), and DHSs in promoters of extremely accessible, heat-induced genes (red) distributions varied markedly. **C,** The promoter regions of extremely accessible genes were poised for activation. Aggregated DNase I-accessibility across these genes and their upstream regions is shown for control (black) and after heat shock (red). The left plot side (left of grey dotted line) shows average per-base DNase I-accessibility 1000 bp upstream of the TSS of extremely accessible, heat-induced genes. The right side (right of grey dotted line, TSS) shows normalized DNase I-accessibility over the first 1000bp of coding regions; the average length of extremely accessible genes was 1651 bp. **D,** Extremely accessible heat-induced genes (red dots), were unique to heat shock. Labeled genes (*HSP101, HSP90.1, HOP3, HSP70, BAG6, HSFA2*) were more accessible in heat shock that similarly sized genes in any other condition examined. **E.** HSE was the most highly enriched motif within heat shock-activated DHSs and within DHSs marking extremely accessible gene promoters relative to all DHSs (p-values from Bonferroni-corrected hypergeometric tests performed on motif counts). **F,** Differential footprinting in the promoter of ZAT10, which is involved in tolerance to abiotic stresses (Mittler et al., 2006). The differentially footprinted region coincided with 3 partial HSE motifs.

Rewiring and involution of the TF network in response to heat shock

Heat shock resulted in substantial rewiring of the TF network, with a net loss (9%) of network edges (**Figure 7A**). Loss of network edges was widely distributed across TF nodes rather than being confined to a few highly connected TFs (**Figure 7A**). These results indicate a central role for the TF network in mediating the repressive component of the heat shock response. RNA polymerase II is known to depart from actively transcribed genes in response to heat shock, yet the mechanisms driving this dynamic are unknown (Teves and Henikoff, 2011). Our findings suggest that the loss of TF occupancy at target promoters may contribute to Pol II departure and global down-regulation of transcription.

HSF-centric sub-networks

A. thaliana encodes 21 HSFs, which fall into different classes based on domain structure. There is no single master regulator of heat shock-responsive genes; rather, three major HSFs (HSFA1A, HSFA1B and HSFA2) together regulate the early heat shock response (Nover et al., 2001). To gain insight into HSF regulation in response to heat shock, we analyzed HSF-centric sub-networks comprising all edges connecting the 21 *A. thaliana* HSFs in control vs. heat-treated conditions. We detected only subtle changes in *HSFA1A* and *HSFA1B* regulation upon heat shock, consistent with their constitutive expression (**Figure 7B**). By contrast, the heat-inducible HSFs such as *HSFA2*, *HSFA7A*, *HSFA7B*, *HSFB1*, and *HSFB2B* (Kilian et al., 2007) are increasingly regulated upon heat shock (**Figure 7B**).

The re-wiring of TF networks in response to heat shock also offered novel information about less well-understood HSFs, some of which function in development rather than in the canonical heat shock response (Kotak et al., 2007). For example, HSFA9

regulates heat shock gene expression during seed development, yet is not induced during heat shock; *HSFA9* showed no alteration in network connectivity. HSFA8 is induced only during heat shock recovery; consistent with this role, *HSFA8* showed numerous regulatory edges in the control state that were lost in heat shock (**Figure 7B**, right).

Finally, we examined HSF feedback loops representing regulation of HSF genes by other HSFs or HSF-regulated TFs (**Figure 7C**). We observed the formation of several novel feedback loops upon heat shock, including a loop between the stress responsive activator ATERF-1 and the heat activated *HSFB1* (Fujimoto et al., 2000)(**Figure 7C**). HSFs thus form a densely connected sub-network linked by shared first-order connections with non-HSF TFs.

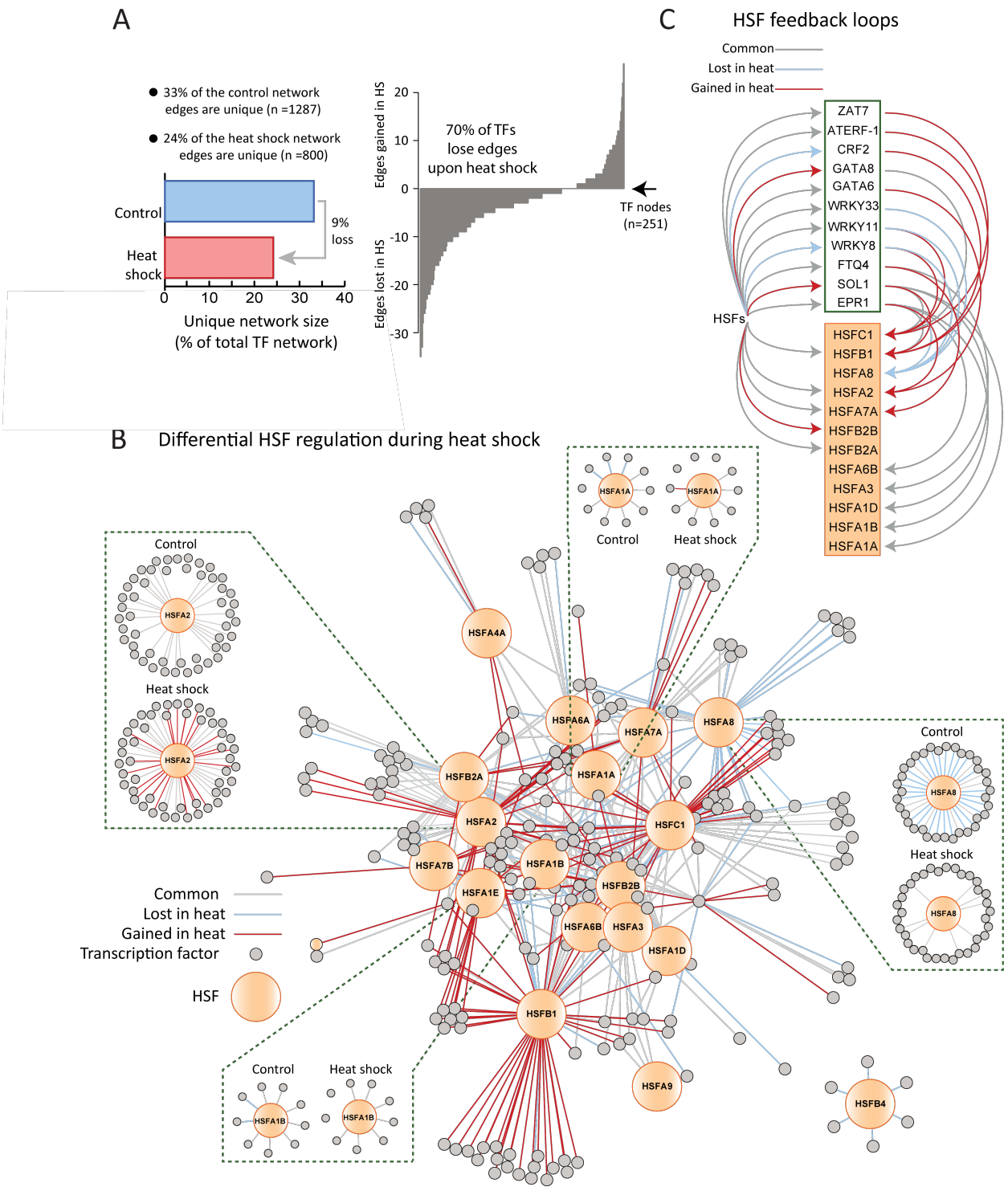

A
33% of the control network edges are unique (n =1287)
24% of the heat shock network edges are unique (n =800)
Control
Heat shock
9% loss
0
10
20
30
40
Unique network size (% of total TF network)
Edges gained in HS
Edges lost in HS
20
10
0
-10
-20
-30
70% of TFs lose edges upon heat shock
TF nodes (n=251)
C
HSF feedback loops
Common
Lost in heat
Gained in heat
HSFs
ZAT7
ATERF-1
CRF2
GATA8
GATA6
WRKY33
WRKY11
WRKY8
FTQ4
SOL1
EPR1
HSFC1
HSFB1
HSFA8
HSFA2
HSFA7A
HSFB2B
HSFB2A
HSFA6B
HSFA3
HSFA1D
HSFA1B
HSFA1A
B
Differential HSF regulation during heat shock
Control
Heat shock
HSFA2
HSFA1A
HSFA4A
HSPA6A
HSFA7A
HSFA8
HSFB2A
HSFC1
HSFA7B
HSFA1B
HSFA1E
HSFB2B
HSFA6B
HSFA3
HSFA1D
HSFB1
HSFA9
HSFB4
Common
Lost in heat
Gained in heat
Transcription factor
HSF

Figure 7. Network connectivity was lost and HSF networks were re-wired in response to heat shock.

A, Left, the number of unique edges in the TF network decreased in heat shock (red) compared to control (blue). Network edges were defined as in **Figure 4A**. Right, TFs tended to lose network edges upon heat shock. **B,** HSF family members (orange circles) were differentially regulated (input edges only). HSFA2, for example, gained many regulators in response to heat shock (see detail left, heat shock-specific edges, red). In contrast, HSFA8 lost many regulators in response to heat shock (see detail right, control-specific edges, grey). This arrangement revealed that HSFB4 was regulated by a small neighborhood of TFs that did not regulate any other HSFs. **C,** HSF feedback loops were generated, using the generic HSE motif for all HSF family members. The indicated TFs (green box) were regulated by one or more HSF and in turn regulated at least one specific HSF.

Discussion

We used genomic footprinting and transcription factor network analysis to map the *cis*- and *trans*-regulatory landscapes of *A. thaliana* whole seedlings, and landscape dynamics in response to light and heat, the two most important environmental cues modulating plant growth and development.

First, the availability of a map of *A. thaliana* DHSs enables integration with other DNA sequence-based annotations such as DNA methylation or genetic variation. For example, SNV's most strongly associated with diverse *A. thaliana* quantitative traits were concentrated in the ~4% of the genome marked by DHSs. Second, using genomic footprinting, we defined a core regulatory compartment in the *A. thaliana* genome comprising nearly 700,000 short sequence elements occupied *in vivo* by TFs. From these elements, we derived a *cis*-regulatory lexicon for *A. thaliana*, including many novel motifs that show evidence of recent purifying selection. We also find evidence that, similar to humans, TF binding shapes codon bias in *A. thaliana*. Finally, we leveraged genomic footprinting data to construct large-scale TF-to-TF regulatory networks.

In spite of the vast divergence between plants and humans, the architecture of *A. thaliana* transcription factor network was strikingly similar to that of the human TF network and other complex information processing systems. This conservation of non-rate limited information processing between plants and animals is striking because plant development is so exquisitely sensitive to environmental cues. In contrast to the multicellular *A. thaliana*, other sessile organisms with acute environmental responses (the unicellular eukaryote yeast, bacteria) show rate-limited, sensory networks for quick responses to transient signals. As the ancestors of plants and animals were unicellular, our

study suggests convergent multicellularity rather than environmental responsiveness as a major driver of optimal network topology.

At the *cis*-regulatory level, photomorphogenesis is characterized by a progression of distinct regulatory DNA compartments that respond to specific light exposures, each of which can be linked to the sequential actions of discrete TFs. We speculate that the prevalence of light-related TF recognition sequences in all seedling DHSs and the prevalence of static DHSs during photomorphogenesis reflect that plant chromatin is poised for the light response. At the TF network level, these changes are accompanied by substantial rewiring between groups of light- and dark-responsive TFs. This rewiring is particularly well demonstrated for autoregulatory feedback loops.

At the *cis*-regulatory level, the hallmark of heat shock is the sharp partitioning of dynamic DHSs into gene-proximal (repressed) and gene-distal (activated) compartments. We also identified novel heat-responsive genes that develop extreme DNaseI accessibility over their promoters and gene bodies upon heat shock. A majority of the DHS landscape remained surprisingly static given the globally repressive nature of heat shock. Because chromatin remodeling is energy intensive, persistence of DHSs may facilitate rapid recovery from heat shock. At the TF network level, the large decrease in network edges suggests a role for departing TFs in mediating transcriptional repression during heat shock.

DHS and footprint data can be applied to improve interpretation of GWAS and quantitative trait locus studies by pinpointing potentially functional non-coding variants. The data can also guide the selection of mutant lines from insertion collections or the selection of DNA elements that can be targeted to perturb specific pathways. Finally, our results constitute a reference against which other *A. thaliana* accessions may be compared,

as much of the vast phenotypic variation among diverse *A. thaliana* accessions is thought to arise from non-coding regulatory regions (Gan et al., 2011).

Methods

Plant Materials. The *UBQ10* INTACT line, in which the *UBQ10* promoter is fused with NTF, is available from ABRC (CS68649). **Treatments.** Seven-day-old seedlings were used for root samples. *Light treatments.* Seven day old dark-grown seedlings were exposed to light for 0mins, 30mins, 3hrs, and 24hrs respectively. *Heat treatments.* Seven-day-old seedlings were heat shocked at 45°C for 30 mins; control plants remained in LD conditions. **Sample preparation.** Nuclei were purified and treated with 45u DNase I for 3 minutes at 25°C. Size fractionation and sequencing of double-cut DNA fragments was done as described (Hesselberth et al., 2009; Neph et al., 2012c). RNA was extracted from 100-200mg tissue, treated with DNase I, and subjected to ribosomal subtraction before library prep and sequencing. RNA expression differences were determined with Cufflinks and Cuffdiff 2.0.2 (Trapnell et al., 2012). Short read archives are in GEO accession GSE53322. For detailed protocols see http://plantregulome.org/protocols. **Mapping DNase I hypersensitivity**. Uniquely-mapping sequencing reads (36 bp) were mapped to TAIR9. The 5' ends of reads were used to calculate per-base DNase I cleavage. DNase I sensitive regions (hotspots), and DHSs (150bp peaks) were identified as in John et al 2011, with minor modifications. Read depth was normalized by subsampling reads. Footprints were computed as described previously (Neph et al., 2012c). **General features of the chromatin landscape.** Binomial distribution tests determined the probability of DHS and/or footprint overlaps with genomic features, including introns and methylated cytosines. Chi-square tests determined

whether ratios of cytosine methylation contexts (CG:CHG:CHH) within DHSs deviated significantly from ratios in regions outside of DHSs. Dark-grown seedling data was used to determine DHS and footprint overlaps with previously published PIF3 ChIP-seq data (Zhang et al., 2013). ***De novo* motif discovery.** 636 *de novo* motifs were discovered by clustering sequences found within footprints from a deeply sequenced 7-day-old seedling sample. *De novo* motifs were validated in two ways. First, by comparing them to 382 protein binding microarray-derived motif models (**See Supplemental Methods**), and second, by comparing estimates of π for neutrally evolving DNA, DNA under purifying selection, and known and novel motifs within and outside of 1% FDR footprints. **GWAS**. For each GWAS phenotype, we used a non-parametric, one-sample Kolmogorov-Smirnov test to determine if low p-value SNVs were more likely to occur in DHSs. **TF trinucleotide preferences and codon usage bias.** Trinucleotide frequencies within footprinted and non-footprinted regions were tabulated for coding and non-coding portions of the human and *A. thaliana* genomes. Codon usage was determined from consensus CDS gene annotations (Pruitt et al., 2009) in human, and coding sequences listed in the TAIR10_GFF3_genes.gff file in *A. thaliana*. **Dynamic DHSs.** In the light series, we defined dynamic DHSs as the top 2% most variable DHSs across conditions. In heat shock, we defined dynamic DHSs as the 5% of DHSs with the greatest relative difference between conditions. **Motif enrichment.** Hypergeometric tests were used to test if motifs frequencies differed in DHS subsets. **Networks.** Methods are as in Neph and co-authors, 2012a, except the region scanned for TF motifs within footprints included 500bp upstream of the TSS and the entire gene model. Potential TF binding sites were determined using FIMO (Bailey et al., 2009), version 4.6.1, with a maximum p-value threshold of 10^{-4} and

defaults for other parameters. **Light activated subnetworks.** We simulated the random selection of 35 light related TF's edges from the network without replacement 1000 times to test the significance of finding 108 regulatory edges among 35 TFs. **First degree TF neighborhoods.** Using Cytoscape (Shannon et al., 2003) all first degree nodes and adjacent edges were selected and all other nodes were removed. **DNase I-accessibility gene outliers.** DNase I-accessibility gene outliers were identified by calculating z-scores from DNase I cleavages overlapping TAIR10 genes. **HSF regulation during heat shock.** HSF-centric regulatory networks were constructed by scanning HSF gene regulatory regions for TF motifs within footprints. The generic HSF motif (Megraw and Hatzigeorgiou, 2010) was used to represent any of the 21 possible HSFs when calculating HSF feedback edges. For details see **Supplemental Methods**.

Acknowledgements

This work was supported by grants from the National Science Foundation (MCB1243627) (JS, CQ, JN), and Graduate Research Fellowship (DGE-0718124) (AMS), and EMBO long-term and HFSP long-term fellowships (JL). We thank Roger Deal and Steven Henikoff for sharing INTACT lines and experimental expertise, members of the Stamatoyannopoulos and Queitsch labs for useful discussions, Stanley Fields for helpful comments on the manuscript. The authors have no conflicts of interest.

Accession numbers

The GEO accession number for short read data reported in thie paper is GSE53322.

Cover artwork by A.M. Sullivan

Chapter 3. Mapping and dynamics of cell-type-specific regulatory DNA during seed coat cell maturation[3]

Acknowledgements

I performed the subsampling of DNase I-seq reads to normalize the data, called hotspots and DHSs, and determined dynamic DHSs. I also performed the expression analysis, motif enrichments, GO enrichments, made the figures, interpreted my analysis in terms of known seed coat biology, and am the author of this chapter. Andrej Arsovski conceived of the experiments, and together with Agnieszka Thomson, DNase I-treated the seed coat samples. Members of John Stamatoyannopoulos' Informatics Group and sequencing unit sequenced the samples and performed alignments.

Introduction

Spatial and temporal regulation of gene expression is critical for development and specialization of tissues. Previous analyses of regulatory DNA and its dynamics in *A. thaliana* focused on identifying regulatory modules involved in the response to environmental cues such as light and heat (**Chapter 2**) (Sullivan et al., 2014). In this chapter we focus on the chromatin landscape during development by following the regulatory DNA landscape of seed coat epidermal cells during their transition from a non-mucous-secreting state to a mucous-secreting state.

The angiosperm seed is composed of three parts: the diploid zygotic embryo, the triploid zygotic endosperm, and the diploid maternal seed coat. The seed coat, which is the

[3] This chapter is part of a manuscript in preparation.

focus of this chapter, differentiates from the integuments of the ovule after fertilization has occurred. In many species, seed coat cells produce and store polysaccharide-rich mucilage (a trait known as myxospermy). When wetted, this mucilage expands and extrudes from mucous secreting cells, forming a jelly layer around the seed (**Figure 1A, B**) (Western et al., 2000; Windsor et al., 2000). In Arabidopsis, mucilage is composed primarily of pectin with lesser amounts of cellulose and xyloglucan (Haughn and Western, 2012). The function of this jelly layer depends on the species and environmental context (Garcia-Fayos et al., 2010; Garwood, 1985; Gutterman and Shem-Tov, 1997; Yang et al., 2011; Yang et al., 2010), but in general it is thought to protect the seedling and facilitate dispersal and germination.

In *A. thaliana,* seed coat cell differentiation and maturation is well characterized morphologically (**Figure 1A**) (Western et al., 2000; Windsor et al., 2000). The seed coat is derived from the outer integuments (2 cell layers) and inner integuments (3 cell layer) of the mature ovule (Schneitz et al., 1995). Mucous secreting cells occupy only the outermost layer of cells of the seed coat (epidermis). The differentiation and maturation sequence from ovule integument to mucous secreting cell as discussed in Western *et al.,* 2000 and Windsor *et al.,* 2000 is summarized below (**Figure 1A**).

In the mature ovule the outermost layer of cells are simple in structure with half of the cell volume occupied by the vacuole. During the first four days after fertilization the vacuole swells causing the cells to grow; starch granules appear, but mucilage is not yet evident (**Figure 1A**). By seven days after fertilization the vacuole shrinks and the cytoplasm forms a column in the middle of the cell that is full of vesicles and golgi stacks. At this time, mucilage is also being secreted into the extracellular space between the primary

cell wall and the cytoplasm (this space is also known as the apoplast) (**Figure 1A**). By ten days after fertilization, mucilage production is completed, the vacuole is restricted under the cytoplasm, starch granules are shrinking, and a secondary cell wall is being deposited around the cytoplasm forming a solid volcano-like structure called a columella (**Figure 1A**). Once differentiation is complete, dehydration shrinks the stored mucilage causing the primary cell wall to drape over the newly formed columella, creating a polygonal donut pattern that can be seen on the exterior of the dry seed.

Approximately 40 genes involved in seed coat cell differentiation and maturation have been characterized in *A. thaliana* (North et al., 2014). These genes fall roughly into 3 temporally sequenced categories: epidermal cell differentiation, mucilage synthesis and secretion, and secondary cell wall synthesis (**Table 1**). It is worth noting that genes controlling specification of ovule integument impact seed coat cell differentiation because seed coat cells are derived from ovule integument. Many of the genes identified as required for differentiation are transcription factors (**Table 1**).

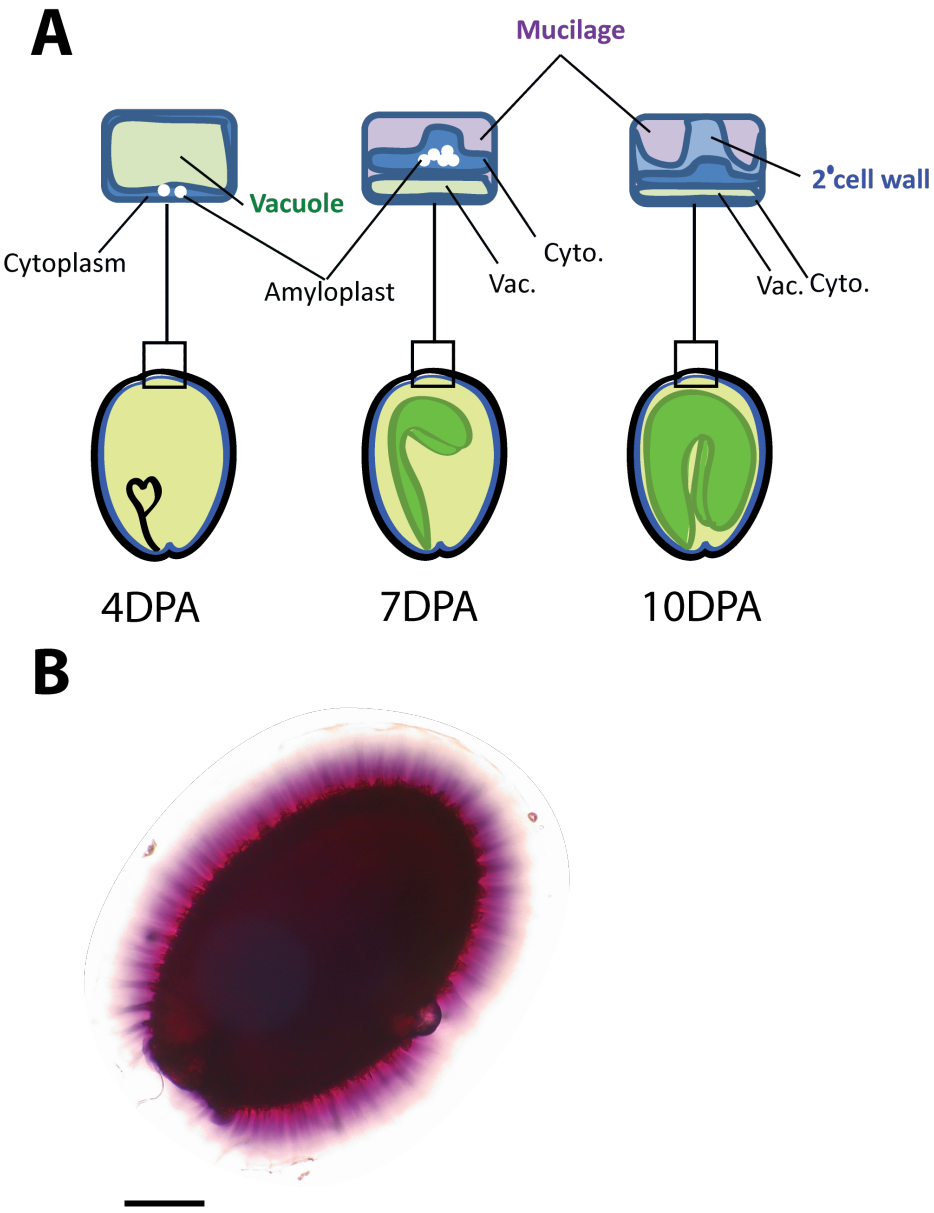


Figure 1. Seed coat maturation. A, The stages of maturation of seed coat epidermal cells and their corresponding embryo stages. "DPA" stands for days post-anthesis; anthesis means flower opening, which happens around the time of fertilization. Top: magnified view of a seed coat epidermis cell in cross section; the vacuole is green; the cytoplasm is dark blue; amyloplasts (starch granules) are in white; the columella, which is made of secondary cell wall, is in light blue. Bottom: shows the corresponding embryo stages for 4, 7, and 10

DPA samples, which are heart embryo (black heart, left), linear cotyledon embryo (green, middle), and mature embryo (green, right). **B,** *A. thaliana* seed (ecotype Col-0) that was imbibed with water and stained with 0.01% Ruthenium red. Mucilage stains purple/pink in color. Seed coat epidermal cells cannot be seen in this image, though the ray like pattern in the mucilage is indicative of the mucilage extrusion points. Scale (black bar) is approximately 100 micrometers. (*Special thanks to Nate Peters for assistance with microscopy*).

Name	ATG	Gene type	Function	Citation
ATS	AT5G42630	KANDI TF	ovule integument	Leon-Kloosterziel et al., 1994
AP2	AT4G36920	AP2/ERF TF	ovule integument	Ohto et al., 2009
BGAL6	AT5G63800	beta-galactosidase 6	pectin modification	Dean et al., 2007
BXL1	AT5G49360	B-d-xylosidase	pectin modification	Arsovski et al., 2009
CESA2	AT4G39350	cellulose synthase	secondary cell wall cellulose	Mendu et al., 2011
CESA9	AT2G21770	cellulose synthase	secondary cell wall cellulose	Stork et al., 2010
CESA5	AT5G09870	cellulose synthase	mucilage, secondary cell wall	Sullivan et al., 2011
DCR	AT5G23940	acyl transferase	cuticular ridges	Hima Rani et al., 2010
DF-1	AT1G76880	putative MYB	mucilage extrusion	Vasilevski et al.,, 2012
DP1	AT4G11180	dirigent protein	lignan synthesis	Esfandiari et al., 2012
ECHIDNA	AT1G09330	DUF846 domain	secretion	Genre et al., 2013
EXO70A1	AT5G03540	exocyst complex	secretion	Kulich et al., 2010
FEI2	AT2G35620	receptor-like kinase	pectin adhesion; cell wall	Harpaz-Saad et al., 2011
FLY1	AT4G28370	RING E3 ubiquitin ligase	secretion; extrusion	Voiniciuc et al., 2013
GAUT11	AT1G18580	galacturonosyltransferase	pectin biosynthesis	Caffall et al., 2009
GATL5	AT1G02720	glycosyltransferase	mucilage production	Kong et al., 2013
GA3OX4	AT1G80330	enzyme in GA pathway	mucilage production	Young-Cheon et al., 2005
GL2	AT1G79840	homeodomain TF	differentiation	Koornneef et al., 1981
KNAT7	AT1G70510	Knox TF	mucilage, cell morphology	Bhargava et al., 2013
LUG	AT4G32551	WD repeat	mucilage extrusion	Walker et al., 2011
LUH	AT2G32700	WD repeat	mucilage extrusion	Huang et al., 2011
MOR1	AT2G35630	MAP215 family	cell structure	McFarlane et al., 2008
MUM4	AT1G53500	rhamnose synthase	mucilage production	Usadel et al, 2004; Western et. al. 2004
MYB5	AT3G13540	MYB TF	mucilage, cell shape	Gonzalez et al, 2009; Li et al., 2009
MYB61	AT1G09540	MYB TF	mucilage, extrusion	Penfield et al., 2001
MYB75	AT1G56650	MYB TF	mucilage, cell morphology	Bhargava et al.,, 2013
NARS1	AT3G15510	NAM/ATAF/NAC TFs	differentiation, cell wall	Kunieda et al., 2008
NARS2	AT1G52880	NAM/ATAF/NAC TFs	differentiation, cell wall	Kunieda et al., 2008
PER36	AT3G50990	class III peroxidase	cell wall, mucilage release	Kunieda et a. 2013
PMEI6	AT2G47670	pectin methyl-transferase inhibitor	pectin modification, extrusion	Saez-Aguayo et al., 2013
RWS3	AT5G63840	glucosidase II	cellulose secretion	Burn et al.,, 2002
SBT1.7	AT5G67360	subtilisin-like serine protease	pectin modification, extrusion	Rautengarten et al., 2008
SEC8	AT3G10380	exocyst complex component	secretion	Kulich et al., 2010
SOS5	AT3G46550	fasciclin-like arabinogalactan	pectin adhesion	Harpaz-Saad et al., 2012
TT1	AT1G34790	WIP zinc finger protein	pigmentation; differentiation	Sagasser et al., 2002
TTG2	AT2G37260	WRKY TF	mucilage production	Johnson et al., 2002
TT2	AT5G35550	MYB TF	pigmentation, differentiation	Gonzalez et al., 2009
TT8	AT4G09820	bHLH TF	pigmentation, differentiation	Gonzalez et al., 2009
TTG1	AT5G24520	WD repeat	pigmentation, differentiation	Galaway et al., 1994
YIP4a	AT2G18840	YPT/RAB GTPase Interacting	secretion	Gendre et al., 2013
YIP4b	AT4G30260	YPT/RAB GTPase Interacting	secretion	Gendre et al., 2013
ROH1	AT1G63930	DUF793 domain	secretion	Kulich et al., 2010

Table 1. Known genes involved in seed coat differentiation and maturation. Gene name, ATG, type and a brief description of the function are listed.

In contrast to the many TFs identified as regulators of seed coat cell fate, few TFs promoting polysaccharide synthesis have been identified. Additionally, individual regulatory elements and their activation during seed coat epidermis differentiation and maturation are largely unknown; exceptions include the promoter of *DP1*, which specifically drives seed coat epidermal expression (Esfandiari et al., 2013), and the L1 box in the *CESA5* promoter, which interacts with GL2 (a seed coat epidermis differentiation factor) in yeast (Tominaga-Wada et al., 2009). To address this paucity of regulatory information, we employed cell-type specific chromatin accessibility profiling (DNase I-seq) to identify regulatory elements, their dynamics, and their constituent TF motifs in seed coat epidermis before and after the initiation of mucilage secretion.

Results

The regulatory DNA landscape of maturing seed coat epidermal cells

To capture the regulatory landscape of seed coat epidermal cells, we employed cell-type-specific nuclear capture (INTACT) (Deal and Henikoff, 2010) followed by DNase I-seq (Sullivan et al., 2014). We used an existing transgenic plant line (Deal and Henikoff, 2010) in which the *GL2* promoter controls the targeting of biotin to the nuclear envelope. We used this line because histochemical staining of seeds harboring the *GL2*promoter:GUS reporter construct suggested that *GL2* expression in the seed is limited to the seed coat epidermis (Windsor et al., 2000). However, very low levels of *GL2* expression likely exist elsewhere in the seed structure (Belmonte et al., 2013). We sampled whole siliques (fruits), which encase 40-60 seeds, at 4 and 7 days post-anthesis (DPA), to capture the regulatory

landscape before and after mucilage production begins in the seed coat epidermis. Anthesis describes the point at which flowers open and fertilization has just occurred.

We created five DNase I-seq libraries, including replicates for each time point (**Table 2**). I combined reads from two 7DPA samples and two 4DPA samples to generate samples of higher read depth. I read-depth normalized the pooled 7DPA and 4DPA samples, and the additional non-pooled 4DPA sample by subsampling to 6,650,000 million reads (ChrC, ChrM, and centromeric reads were excluded) before further analysis (**Table 3**). I called Hotspots and DHSs for each sample using established methods (Sullivan et al., 2014). I identified a total of 27,262 DHSs in the 7DPA sample; 28,095 in the pooled 4DPA sample, and 29,856 in the non-pooled sample at a 1% FDR threshold.

DNase I-seq library	Number of reads (Chr1-5 only)
4DPA-DS20201	158,052,096
4DPA-DS20132	3,582,146
4DPA-DS20131	3,246,245
7DPA-DS21306	6,228,499
7DPA-DS20134	1,227,315

Table 2. DNase I-seq libraries. Sample number and number of reads mapping to chromosomes 1-5 (ChrC/ChrM reads have been filtered out) are listed.

Sample	Filtered reads	Quality	DHSs (1% FDR)
4DPA (pooled)	6,650,000	0.643	28,095
4DPA (DS20201)	6,650,000	0.594	29,856
7DPA (pooled)	6,650,000	0.476	27,262

Table 3. Seed coat samples after data processing. Sample name, number of reads (Chr1-5 only, no centromeric regions), signal quality score (proportion of reads in 1% FDR hotspots), and the number of 1% FDR DHSs.

2,111 DHSs are dynamic between 4DPA and 7DPA

Next we sought to compare the regulatory landscape of the 4DPA and 7DPA samples to understand how the regulatory landscape changes over developmental time. To do this we calculated the relative difference in quality-normalized DNase I cleavage in 36,075 DHSs, which represent a merged set of DHSs from all 3 samples (methods). We compared each of the 4DPA samples to the single, pooled 7DPA sample and used only the DHSs that are reproducibly different in the 7DPA sample (**Figure 2A, B**). This yielded 540 7DPA-deactivated and 1571 7DPA-activated DHSs with a relative difference cut off of -1 and 1 respectively; 70% of the dynamic DHSs identified in each 4DPA-7DPA comparison were reproducible.

We found dynamic DHSs near known seed coat development genes. For example, we found 7DPA-activated DHSs near *MYB61*, which is required for mucilage production, and *PER36*, which is required for proper mucilage release (**Figure 2C**). We also identified dynamic DHSs near genes that are not yet associated with seed coat development. For example, the meristem identity TF, *LMI2*, was near a DHS that is deactivated during seed coat cell maturation (**Figure 2C**). We also identified DHSs that were static during development, such as the one in the promoter of *TTG1* which is required for seed coat determination (**Figure 2C**). Furthermore, the landscape of seed coat cells looks very different from the regulatory landscape of root or root non-hair cells (**Figure 2C**).

The majority of developmentally dynamic DHSs are intergenic

Next we asked if the distribution of dynamic DHSs was different than the distribution of all DHSs in the genome by tabulating the number of DHSs occurring in various genomic contexts (e.g. intragenic). Similar to whole seedling DHSs (Sullivan et al.,

2014), seed coat epidermal DHSs (dynamic and static), are enriched intergenically and near transcription start sites (TSSs) (400 bp upstream of the TSS), and are depleted intragenically and in transposable elements (TEs). However, dynamic seed coat DHSs are more likely to be inter- and intragenic, and 7DPA-activated DHSs were more likely to be in TEs (**Figure 2D**).

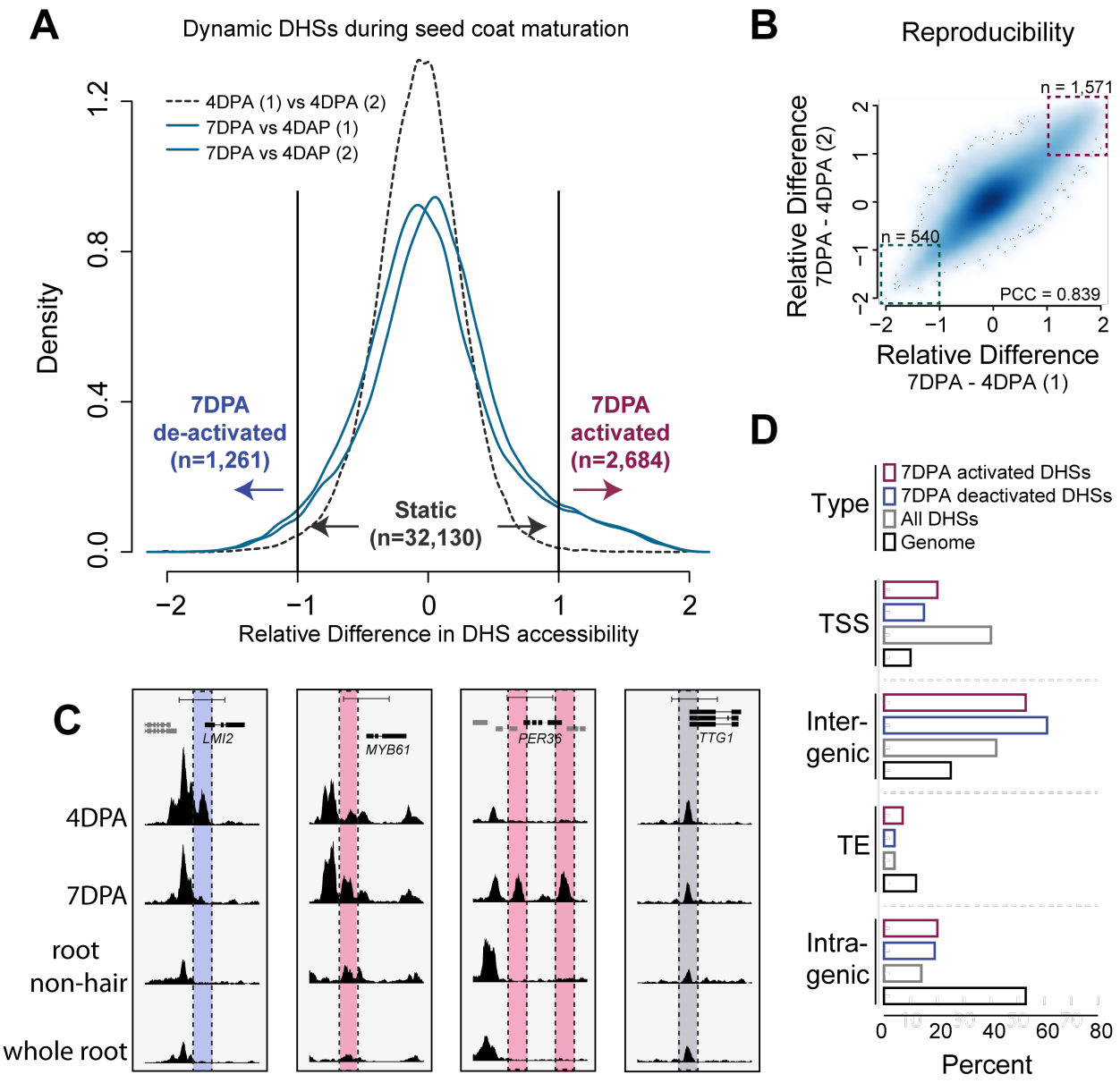

A
Dynamic DHSs during seed coat maturation
4DPA (1) vs 4DPA (2)
7DPA vs 4DAP (1)
7DPA vs 4DAP (2)
Density
0.0
0.4
0.8
1.2
7DPA
de-activated
(n=1,261)
7DPA
activated
(n=2,684)
Static
(n=32,130)
−2
−1
0
1
2
Relative Difference in DHS accessibility
B
Reproducibility
n = 1,571
n = 540
PCC = 0.839
Relative Difference
7DPA - 4DPA (2)
Relative Difference
7DPA - 4DPA (1)
C
LMI2
MYB61
PER36
TTG1
4DPA
7DPA
root
non-hair
whole root
D
Type
7DPA activated DHSs
7DPA deactivated DHSs
All DHSs
Genome
TSS
Inter-
genic
TE
Intra-
genic
0
20
40
60
80
Percent

Figure 2. The chromatin landscape of maturing seed coat epidermal cells. A, Distribution of relative differences ((7DPA-4DPA)/mean(7DPA,4DPA)) in per-base DNase I cleavages in DHSs among replicate 4DPA samples (dashed grey line), and each 4DPA sample and the 7DPA sample (blue lines). A cut off of 1 and -1 (vertical lines) was used to determine if a DHS was "dynamic". **B,** Correlation of the relative differences for each 4DPA sample when compared with the 7DPA sample; the Pearson Correlation Coefficient (PCC) was 0.839. Reproducible dynamic DHSs are indicated by the dashed boxes. **C,** Examples showing a 7DPA-deactivated DHS, two examples of 7DPA-activated DHSs, and one example of a static DHS. The window for each example is 5kb. Root non-hair cell and whole root samples of the same read depth are shown for comparison. **D,** The numbers of DHSs and dynamic DHSs within various genomic contexts were tabulated; all DHSs are in grey, 7DPA-activated and 7DPA-deactivated DHSs are in red and blue respectively. The percent of base pairs in the 36-mer mappable genome belonging to each genomic context is in black.

Dynamic DHSs are enriched near differentially expressed genes

We used proximity to attempt to relate dynamic DHSs to the genes which they regulate. To do this, we considered both the closest gene and the closest differentially expressed gene to dynamic DHSs (**Figure 3A, B**). We took advantage of two published seed coat epidermis expression studies (Belmonte et al., 2013; Dean et al., 2011) and considered a gene to be differentially expressed if it exhibited a 2-fold change in expression between the developmental time points sampled.

Dean *et al.,* 2011 quantify expression differences in manually dissected seed coats at 3DPA and 7DPA in the ecotype Col-2; Dean *et al.,* identified 3,430 genes that changed expression by at least 2-fold. Belmonte *et al.,* 2013 quantify expression differences in many parts of the seed at many time points in the ecotype Ws-0 using laser capture micro dissection; the seed coat (chalazal SC and SC) and embryo proper (EP) data sets at the heart embryo and linear cotyledon stage (equivalent to ~4DPA and ~7DPA) were used in this analysis. A total of 3,735 genes changed expression by 2-fold in seed coat, and 1,851 changed expression in embryo proper. Both studies used microarrays to evaluate gene expression.

We used the embryo proper data set from the same time points (~4DPA and ~7DPA) from Belmonte *et al.,* 2013 as a control in further analyses. We used this control because embryos were present in the tissue collected for DNase I-seq, and these embryos are a potential source of signal contamination. If the differentially expressed genes in embryo behave like random genes, we will know our seed coat nuclei enrichment strategy is working as intended.

To test if differentially expressed genes are preferentially located near dynamic DHSs, the proximity of differentially expressed genes to their nearest dynamic DHSs was calculated (**Figure 3C, D**). Differentially expressed genes found in Dean *et al.,* 2011 seed coat and Belmonte *et al.,* 2013 seed coat are preferentially located within 4kb of dynamic DHSs (**Figure 3C, D**). Differentially expressed genes in Belmonte et al., 2013 embryo proper or randomly selected genes sets of similar sizes to the experimental data were not enriched within 4kb of dynamic DHSs, suggesting our nuclear enrichment strategy is working to capture seed-coat-specific nuclei (**Figure 3C,D**). The distribution of differentially expressed genes relative to the DHS is roughly symmetrical, though a slight bias towards genes being downstream of dynamic DHSs is clearly visible in the Dean et al., 2011 set of differentially expressed genes (**Figure 3C, D**).

The percent of all differentially expressed genes in each data set that mapped to a dynamic DHS within 5kb was 22-23% for seed coat (depending on the expression data set), and 17% for embryo; by chance we would expect 16% of genes to end up within 5kb of a DHS (**Figure 3E**). The proportion of all dynamic DHSs that are mapped to differentially expressed seed coat genes within 5kb is 41-42% (depending on the expression data set), much higher than the proportions of dynamic DHSs mapped to differentially expressed embryo genes or random gene sets of the same sizes as the experimental data sets (**Figure 3F**). The percent of genes or DHSs mapped to each other per base pair peaks at 250bp and then decreases with distance to the DHS (**Figure 3E, F**).

To understand the relationship between expression and DHS dynamics, we asked what fraction of the time the direction in DHS change matched the direction in gene

expression change for each expression set. We observed that DHSs and their nearest mapped differentially expressed gene agreed in direction (e.g. 7DPA-activated or 7DPA-deactivated) more often than one would expect by chance (**Figure 3G, H**). Our expectation for agreements (**Figure 3G, H,** dashed lines) were based on the proportion of dynamic DHSs that are 7DPA-deactivated (25.5%) or 7DPA-activated (74.4%), and the proportion of genes that are 7DPA-deactivated or 7DPA-activated in each of the expression data sets (Dean et al., 2011 deactivated (34.6%), activated (65.4%); Belmonte et al., 2013 seed coat deactivated (44.5%), activated (55.5%); Belmonte et al., 2013 embryo deactivated (60.6%), activated (39.4%)). The rate of agreement generally dropped with distance to the DHS, though agreement remained above expectation for genes greater than 8kb away from a dynamic DHS.

To determine how well our strategy of mapping dynamic DHSs to the nearest gene (**Figure 3A**) was capturing differentially expressed genes, we calculated the percentage of genes that were nearest to dynamic DHSs (**Figure 3A**) that were also differentially expressed (**Figure 3B**). In the Dean study, 24% of genes nearest to dynamic DHSs were also differentially expressed (**Figure 2I**). In the Belmonte study, 24% of genes nearest to dynamic DHSs were also differentially expressed in seed coat (**Figure 2I**). Only 7% of genes nearest to dynamic DHSs were also differentially expressed in the embryo (data not shown).

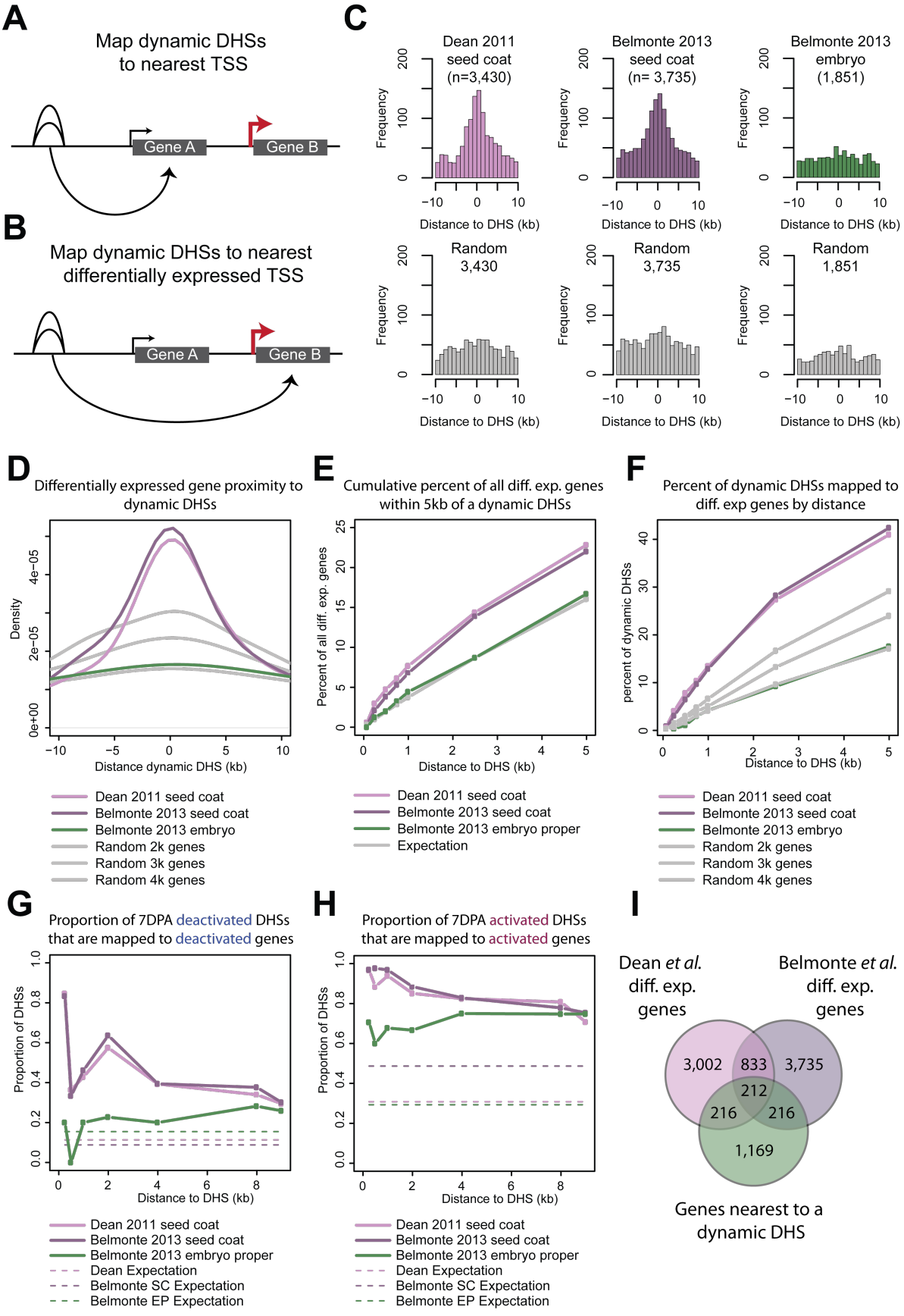

A
Map dynamic DHSs to nearest TSS
Gene A
Gene B
B
Map dynamic DHSs to nearest differentially expressed TSS
Gene A
Gene B
C
Dean 2011 seed coat (n=3,430)
Belmonte 2013 seed coat (n= 3,735)
Belmonte 2013 embryo (1,851)
Random 3,430
Random 3,735
Random 1,851
Frequency
Distance to DHS (kb)
D
Differentially expressed gene proximity to dynamic DHSs
Density
Distance dynamic DHS (kb)
E
Cumulative percent of all diff. exp. genes within 5kb of a dynamic DHSs
Percent of all diff. exp. genes
Distance to DHS (kb)
F
Percent of dynamic DHSs mapped to diff. exp genes by distance
percent of dynamic DHSs
Distance to DHS (kb)
Dean 2011 seed coat
Belmonte 2013 seed coat
Belmonte 2013 embryo
Belmonte 2013 embryo proper
Random 2k genes
Random 3k genes
Random 4k genes
Expectation
G
Proportion of 7DPA deactivated DHSs that are mapped to deactivated genes
H
Proportion of 7DPA activated DHSs that are mapped to activated genes
Proportion of DHSs
Distance to DHS (kb)
Dean Expectation
Belmonte SC Expectation
Belmonte EP Expectation
I
Dean et al. diff. exp. genes
Belmonte et al. diff. exp. genes
3,002
833
3,735
212
216
216
1,169
Genes nearest to a dynamic DHS

Figure 3. Mapping dynamic DHSs to nearby and differentially expressed genes. A, Mapping dynamic DHSs to the nearest TSS. **B,** Mapping dynamic DHSs to the nearest differentially expressed gene. **C,** Distribution of distances of differentially expressed gene TSSs to 2111 reproducible dynamic DHSs. Distance to DHS is the distance of the TSS relative to the DHS midpoint which is at zero on the x-axis; a negative distance means the gene is upstream of the DHS and a positive distance means the gene is downstream of the DHS. Differentially expressed genes from Dean *et al.,* 2011 seed coat are in light purple, Belmonte *et al.,* 2013 seed coat in dark purple, Belmonte *et al.,* 2013 embryo proper is in green, and randomly selected genes sets of the same sizes to the experimental sets are in grey. **D,** Density distributions of the data in Figure 3C plotted together for comparison. **E,** The cumulative percent of differentially expressed genes in each data set that are within 5kb of a dynamic DHS. The expectation line (grey) reflects the percent of the genome in range of a dynamic DHS (midpoint) at the noted distances (75bp, 250bp, 500bp, 1kb, 2.5kb, 5kb). **F,** The cumulative percent of dynamic DHSs that are mapped to differentially expressed genes in each experimental data set, or random sets of genes of the same size. The noted distances are the same as in Figure 3E. **G,** The proportion of 7DPA-deactivated DHSs that are mapped to deactivated genes in each data set. Dashed lines indicate the expected agreement based on the proportion of dynamic DHSs which are deactivated and the proportion of deactivated genes in each expression data set. **H,** The proportion of 7DPA activated DHSs that are mapped to activated genes in each data set. Dashed lines indicate the expected agreement based on the proportion of dynamic DHSs which are deactivated and the proportion of deactivated genes in each expression data set. **I,** The proportion of nearest genes to dynamic DHSs that are also differentially expressed in one or both seed coat studies.

Genes nearest to dynamic DHSs agree with seed coat biological processes

To test if the genes that are nearest to dynamic DHSs (**Figure 3A**) are involved in known seed coat epidermis biology, we asked if they were enriched for GO terms, KEGG pathways (Kanehisa and Goto, 2000), and INTERPRO domains (Hunter et al., 2012) using DAVID (Huang da et al., 2009) (**Figure 4A**). We found that the genes nearest to both deactivated and activated dynamic DHSs are enriched for transcription factors and hormone biosynthesis. Genes closest to deactivated DHSs were additionally enriched for hormone response. Genes closest to activated DHSs were additionally enriched in genes related to metabolism and genes whose cellular compartment is the cell wall, extracellular region, or endomembrane system.

A

		Genes nearest deactivated DHSs		Genes nearest activated DHSs		Genes nearest all dynamic DHSs		
Category	Term	count	p-value	count	p-value	count	pvalue	
GOTERM_BP_FAT	GO:0032870:cellular response to hormone stimulus					52	0.000239	HORMONE
GOTERM_BP_FAT	GO:0042446:hormone biosynthetic process					12	7.6E-05	
GOTERM_BP_FAT	GO:0042445:hormone metabolic process					17	3.56E-05	
GOTERM_BP_FAT	GO:0009755:hormone-mediated signaling					52	0.000239	
GOTERM_BP_FAT	GO:0010817:regulation of hormone levels					22	9.08E-05	
GOTERM_BP_FAT	GO:0009725:response to hormone stimulus	39	1.22E-05			100	1.07E-05	RESPONSE
GOTERM_BP_FAT	GO:0009719:response to endogenous stimulus	43	1.87E-06			105	1.52E-05	
GOTERM_BP_FAT	GO:0010033:response to organic substance	47	7.67E-06			118	0.000102	
GOTERM_MF_FAT	GO:0003677:DNA binding	92	2.05E-11			217	1.85E-05	TRANSCRIPTIONAL REGULATION
INTERPRO	IPR012287:Homeodomain-related	21	8.47E-07			50	2.07E-08	
INTERPRO	IPR015495:Myb transcription factor	14	2.12E-06			24	0.000109	
GOTERM_BP_FAT	GO:0051252:regulation of RNA metabolic process	47	1.13E-06			116	1.1E-05	
GOTERM_BP_FAT	GO:0045449:regulation of transcription	84	3.49E-11			213	9.74E-10	
GOTERM_BP_FAT	GO:0006355:regulation of transcription, DNA-dependent	47	9.63E-07			115	1.41E-05	
GOTERM_BP_FAT	GO:0006350:transcription	54	4.89E-07			139	1.12E-06	
INTERPRO	IPR001092:Basic helix-loop-helix dimerisation region bHLH	11	0.000457					
INTERPRO	IPR014778:Myb, DNA-binding	16	6.47E-05					
INTERPRO	IPR017930:Myb-type HTH DNA-binding domain	14	0.000182					
INTERPRO	IPR001005:SANT, DNA-binding	17	8.13E-06					
INTERPRO	IPR001356:Homeobox					19	4.65E-05	
INTERPRO	IPR017970:Homeobox, conserved site					18	5.05E-05	
GOTERM_MF_FAT	GO:0003700:transcription factor activity	87	6.68E-19	117	0.000289	197	8.95E-15	
GOTERM_MF_FAT	GO:0030528:transcription regulator activity	90	6.36E-17	130	0.000376	213	1.42E-13	
GOTERM_MF_FAT	GO:0016563:transcription activator activity			19	0.000367	24	0.000119	
INTERPRO	IPR002283:Isopenicillin N synthase	6	0.000181					
GOTERM_MF_FAT	GO:0009055:electron carrier activity			58	4.96E-05	71	0.000137	META-BOLISM
GOTERM_MF_FAT	GO:0005506:iron ion binding			64	1.97E-05	81	1.38E-05	
KEGG_PATHWAY	ath00360:Phenylalanine metabolism					14	0.000354	
KEGG_PATHWAY	ath00380:Tryptophan metabolism					9	0.000608	
GOTERM_MF_FAT	GO:0020037:heme binding			36	0.000331			
GOTERM_BP_FAT	GO:0042545:cell wall modification	11	0.000382					EXTRACELLULAR CELL WALL
GOTERM_CC_FAT	GO:0048046:apoplast			33	0.000659	40	0.001071	
GOTERM_CC_FAT	GO:0005576:extracellular region			91	3.05E-06	118	4.31E-07	
GOTERM_CC_FAT	GO:0044421:extracellular region part			9	0.000443	10	0.00066	
GOTERM_CC_FAT	GO:0016021:integral to membrane					169	0.000641	
GOTERM_CC_FAT	GO:0031224:intrinsic to membrane					201	0.000451	
GOTERM_CC_FAT	GO:0009505:plant-type cell wall					33	0.000232	
GOTERM_CC_FAT	GO:0044459:plasma membrane part					31	0.001184	
GOTERM_CC_FAT	GO:0012505:endomembrane system			229	0.000161	303	3.45E-05	SECRETION

Figure 4. Term enrichment for genes nearest to dynamic DHSs. A, GO term, KEGG pathway, and INTERPRO domain enrichments of genes nearest to dynamic DHSs. Only the enrichments with a Benjamini-corrected p-value less than 0.05 are shown. Enrichments were manually summarized into very general categories denoted in colored boxes at the right.

Motif families in activated and deactivated DHSs are compartmentalized

To better understand the molecular basis for deactivated and activated DHSs we looked at their transcription factor motif densities. We analyzed motifs in two ways: first we compared motif densities in dynamic DHSs relative to all DHSs, the vast majority of which are static (**Figure 5A**); second, we compared motif densities in dynamic DHSs to the genome (**Figure 5B**). Motif enrichments were performed using hypergeometric tests, and only the TFs that are significantly enriched after multiple testing correction are shown in Figure 5. We observed compartmentalization of TF families between activated and deactivated DHSs when compared to all DHSs and the genome. For example, MYBs and NACs are enriched in activated DHSs whereas TCPs are enriched in deactivated DHSs (**Figure 5A, B**). All of the MYBs with enriched motifs are of the same subfamily, the R2R3 MYBs. There are also motif families whose motifs are enriched in dynamic DHSs when compared to the genome, but not when compared to all DHSs. These motifs belong to bZIP and bHLH families.

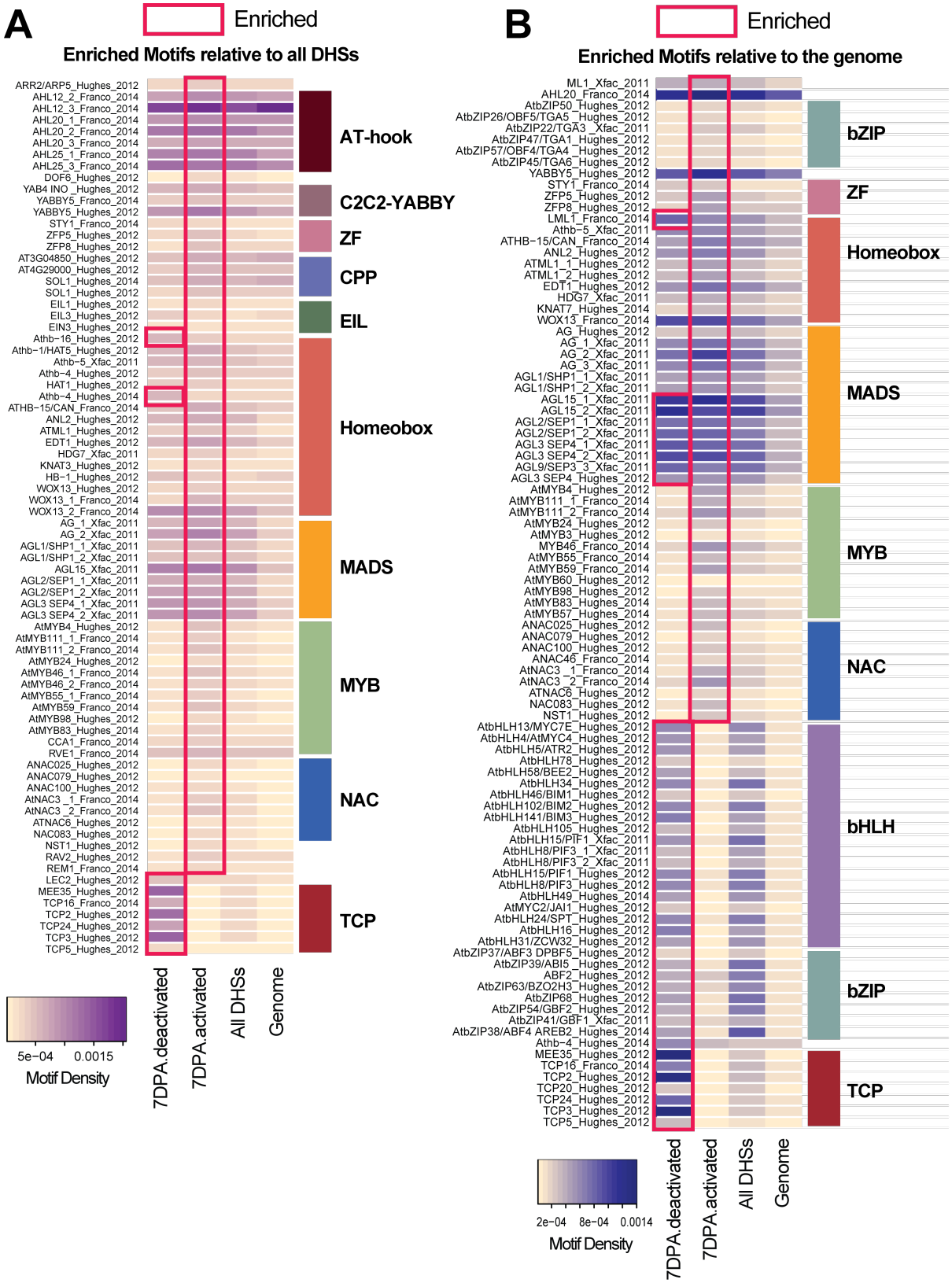

A
Enriched
Enriched Motifs relative to all DHSs
ARR2/ARP5_Hughes_2012
AHL12_2_Franco_2014
AHL12_3_Franco_2014
AHL20_1_Franco_2014
AHL20_2_Franco_2014
AHL20_3_Franco_2014
AHL25_1_Franco_2014
AHL25_3_Franco_2014
DOF6_Hughes_2012
YAB4 INO _Hughes_2012
YABBY5_Franco_2014
YABBY5_Hughes_2012
STY1_Franco_2014
ZFP5_Hughes_2012
ZFP8_Hughes_2012
AT3G04850_Hughes_2012
AT4G29000_Hughes_2012
SOL1_Hughes_2014
SOL1_Hughes_2012
EIL1_Hughes_2012
EIL3_Hughes_2012
EIN3_Hughes_2012
Athb−16_Hughes_2012
Athb−1/HAT5_Hughes_2012
Athb−5_Xfac_2011
Athb−4_Hughes_2012
HAT1_Hughes_2012
Athb−4_Hughes_2014
ATHB−15/CAN_Franco_2014
ANL2_Hughes_2012
ATML1_Hughes_2012
EDT1_Hughes_2012
HDG7_Xfac_2011
KNAT3_Hughes_2012
HB−1_Hughes_2012
WOX13_Hughes_2012
WOX13_1_Franco_2014
WOX13_2_Franco_2014
AG_1_Xfac_2011
AG_2_Xfac_2011
AGL1/SHP1_1_Xfac_2011
AGL1/SHP1_2_Xfac_2011
AGL15_Xfac_2011
AGL2/SEP1_1_Xfac_2011
AGL2/SEP1_2_Xfac_2011
AGL3 SEP4_1_Xfac_2011
AGL3 SEP4_2_Xfac_2011
AtMYB4_Hughes_2012
AtMYB111_1_Franco_2014
AtMYB111_2_Franco_2014
AtMYB24_Hughes_2012
AtMYB46_1_Franco_2014
AtMYB46_2_Franco_2014
AtMYB55_1_Franco_2014
AtMYB59_Franco_2014
AtMYB98_Hughes_2012
AtMYB83_Hughes_2014
CCA1_Franco_2014
RVE1_Franco_2014
ANAC025_Hughes_2012
ANAC079_Hughes_2012
ANAC100_Hughes_2012
AtNAC3 _1_Franco_2014
AtNAC3 _2_Franco_2014
ATNAC6_Hughes_2012
NAC083_Hughes_2012
NST1_Hughes_2012
RAV2_Hughes_2012
REM1_Franco_2014
LEC2_Hughes_2012
MEE35_Hughes_2012
TCP16_Franco_2014
TCP2_Hughes_2012
TCP24_Hughes_2012
TCP3_Hughes_2012
TCP5_Hughes_2012
AT-hook
C2C2-YABBY
ZF
CPP
EIL
Homeobox
MADS
MYB
NAC
TCP
7DPA.deactivated
7DPA.activated
All DHSs
Genome
5e−04
0.0015
Motif Density
B
Enriched
Enriched Motifs relative to the genome
ML1_Xfac_2011
AHL20_Franco_2014
AtbZIP50_Hughes_2012
AtbZIP26/OBF5/TGA5_Hughes_2012
AtbZIP22/TGA3 _Xfac_2011
AtbZIP47/TGA1_Hughes_2012
AtbZIP57/OBF4/TGA4 _Hughes_2012
AtbZIP45/TGA6_Hughes_2012
YABBY5_Hughes_2012
STY1_Franco_2014
ZFP5_Hughes_2012
ZFP8_Hughes_2012
LML1_Franco_2014
Athb−5_Xfac_2011
ATHB−15/CAN_Franco_2014
ANL2_Hughes_2012
ATML1_1_Hughes_2012
ATML1_2_Hughes_2012
EDT1_Hughes_2012
HDG7_Xfac_2011
KNAT7_Hughes_2014
WOX13_Franco_2014
AG_Hughes_2012
AG_1_Xfac_2011
AG_2_Xfac_2011
AG_3_Xfac_2011
AGL1/SHP1_1_Xfac_2011
AGL1/SHP1_2_Xfac_2011
AGL15_1_Xfac_2011
AGL15_2_Xfac_2011
AGL2/SEP1_1_Xfac_2011
AGL2/SEP1_2_Xfac_2011
AGL3 SEP4_1_Xfac_2011
AGL3 SEP4_2_Xfac_2011
AGL9/SEP3_3_Xfac_2011
AGL3 SEP4_Hughes_2012
AtMYB4_Hughes_2012
AtMYB111_1_Franco_2014
AtMYB111_2_Franco_2014
AtMYB24_Hughes_2012
AtMYB3_Hughes_2012
MYB46_Franco_2014
AtMYB55_Franco_2014
AtMYB59_Franco_2014
AtMYB60_Hughes_2012
AtMYB98_Hughes_2012
AtMYB83_Hughes_2014
AtMYB57_Hughes_2014
ANAC025_Hughes_2012
ANAC079_Hughes_2012
ANAC100_Hughes_2012
ANAC46_Franco_2014
AtNAC3 _1_Franco_2014
AtNAC3 _2_Franco_2014
ATNAC6_Hughes_2012
NAC083_Hughes_2012
NST1_Hughes_2012
AtbHLH13/MYC7E_Hughes_2012
AtbHLH4/AtMYC4_Hughes_2012
AtbHLH5/ATR2_Hughes_2012
AtbHLH78_Hughes_2012
AtbHLH58/BEE2_Hughes_2012
AtbHLH34_Hughes_2012
AtbHLH46/BIM1_Hughes_2012
AtbHLH102/BIM2_Hughes_2012
AtbHLH141/BIM3_Hughes_2012
AtbHLH105_Hughes_2012
AtbHLH15/PIF1_Xfac_2011
AtbHLH8/PIF3_1_Xfac_2011
AtbHLH8/PIF3_2_Xfac_2011
AtbHLH15/PIF1_Hughes_2012
AtbHLH8/PIF3_Hughes_2012
AtbHLH49_Hughes_2014
AtMYC2/JAI1_Hughes_2012
AtbHLH24/SPT_Hughes_2012
AtbHLH16_Hughes_2012
AtbHLH31/ZCW32_Hughes_2012
AtbZIP37/ABF3 DPBF5_Hughes_2012
AtbZIP39/ABI5_Hughes_2012
ABF2_Hughes_2012
AtbZIP63/BZO2H3_Hughes_2012
AtbZIP68_Hughes_2012
AtbZIP54/GBF2_Hughes_2012
AtbZIP41/GBF1_Xfac_2011
AtbZIP38/ABF4 AREB2_Hughes_2014
Athb−4_Hughes_2014
MEE35_Hughes_2012
TCP16_Franco_2014
TCP2_Hughes_2012
TCP20_Hughes_2012
TCP24_Hughes_2012
TCP3_Hughes_2012
TCP5_Hughes_2012
bZIP
ZF
Homeobox
MADS
MYB
NAC
bHLH
bZIP
TCP
7DPA.deactivated
7DPA.activated
All DHSs
Genome
2e−04
8e−04
0.0014
Motif Density

Figure 5. Motif enrichments within dynamic DHSS. A, Motif densities (motifs/bp) of motifs within each type of DHS are shown. Only motifs that are enriched (Bonferroni-corrected p-value < 0.05) **relative to all DHSs** are shown in the heatmap and are highlighted in a red outline. Motif name and source are noted at the left. Color blocks on the right indicate transcription factor family. **B,** Motif densities (motifs/bp) of motifs within each type of DHS are shown. Only motifs that are enriched (Bonferroni-corrected p-value < 0.05) **relative to the genome** are shown in the heatmap and are highlighted in a red outline. Motif name and source are noted at the left. Color blocks on the right indicate transcription factor family.

Discussion

In this study we mapped regulatory elements and their dynamics using cell-type specific DNase I-seq. We applied this method to study the development of seed coat cells as they transition from a state of growth to a state of mucous production and secretion. We identified 2,111 reproducibly dynamic DHSs during this developmental window.

DHSs are a gene regulatory phenomenon and thus we expect that dynamic DHSs are indicators of gene expression changes of 'nearby' genes. Consistent with this hypothesis, we found genes that are differentially expressed during seed coat cell differentiation were enriched near dynamic DHSs. Differentially expressed genes were also often the nearest genes to dynamic DHSs. However, connecting dynamic DHSs to the genes they regulate is not simple for several reasons. First, a change in DNase I accessibility is not always a predictor of altered expression since regulatory DNA can be poised for transcription activation (Elgin, 1988) and DHSs can remain after transcription is shut off (Groudine and Weintraub, 1982). Our analysis is based on dynamic DHSs and thus static DHSs and the genes they reside near are not emphasized in our analysis. Second, the binding of activators (Morgan et al., 1987) or repressors (Banuagnad et al.,, 1990) has the potential to create DHSs. Therefore, an increase in accessibility does not necessarily indicate an increase in gene expression, though we do see a fair amount of agreement between the direction of DHS change and expression change in our analysis. Finally, we know that enhancers can work at long distances, agnostic to orientation (Banerji et al., 1981). Therefore, we can map DHSs to genes based on their proximity to the TSS; however, we may be missing long range interactions, or making the incorrect gene assignment.

Despite these limitations, dynamic DHSs are potentially useful in identifying new candidate genes that control seed coat development, and suggest potential TFs that may be driving or responding to DHS dynamics. Term enrichments performed on genes nearest to dynamic DHSs revealed TF families (e.g. MYBs, homeoboxes) that may be involved in seed coat cell maturation. 7DPA-deactivated DHSs were associated with hormone response, reflecting an early role for hormones in growth and specification of seed coat fate. The genes near 7DPA-activated DHSs are primarily associated with the secretion system, cell wall modification, and the extracellular compartment, which is consistent with these cells switching to a state of mucous production and secretion of compounds outside the cell membrane and building the secondary cell wall structure of the columella.

Motif enrichments within 7DPA-activated and 7DPA-deactivated DHSs revealed distinct TF families and individual TFs that may be regulating seed coat cell maturation. The activated and deactivated DHSs were enriched in different TF families, showing strong compartmentalization of TF and function.

Motifs for several TCPs and LEC2 were enriched in 7DPA-deactivated DHSs, suggesting that these TFs act early in growth and fate specification and must be deactivated for cells to acquire a terminal fate. Consistent with this hypothesis, TCPs and LEC2 play roles in embryogenesis, cell growth, and identity. For example, modulation expression of the maize TCP TF *tb1* accounts for morphological traits that differentiate wild and domestic maize species (Clark et al., 2006). In cotton, TCPs function in auxin-mediated differentiation of seed coat epidermal cells into cotton fibers; silencing of *GbTCP* leads to a shortened cotton fiber (Hao et al., 2012; Wang et al., 2013). Mutant *lec2 Arabidopsis* plants

are severely affected in embryo development; seed coats are affected and have a possible pigmentation defect (Meinke et al.,, 1994). Ectopic expression of *lec2* also confers embryonic features to somatic tissues (Guo et al., 2013).

Many TFs were enriched in 7DPA-activated DHSs. NACs, MYBs, and Homeobox TFs are known to be involved in seed coat cell maturation (Table 1). All the MYBs with enriched motifs are from the R2R3 MYB family. We also found zinc finger TFs, MADS-box, and AT-hook TFs were enriched in activated DHSs; these TFs have not been implicated previously in seed coat cell maturation. However, MADS-box TFs are required for proper reproductive organ development (Honma and Goto, 2001).

Finally, we have identified a number of seed coat mucilage candidate genes based on expression and/or chromatin accessibility profiles (Table 4). T-DNA insertion mutants for these lines have been obtained and are currently being analyzed for defects in seed coat mucilage production or extrusion (Table 3).

Gene	Description	Stock
AT4G30280	XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 18 (ATXTH18)	SALK_025862C
AT5G10430	ARABINOGALACTAN PROTEIN 4	SALK_024865C
AT5G56870	BETA-GALACTOSIDASE 4	SALK_022796C
AT5G59120	SUBTILASE 4.13	SALK_082062C
AT2G23560	METHYL ESTERASE 7	SALK_054303C
AT2G31750	UDP-GLUCOSYL TRANSFERASE 74D1	SALK_011286C
AT2G31960	GLUCAN SYNTHASE-LIKE 3	SALK_011560C
AT3G13750	BETA-GALACTOSIDASE 1	SALK_025142C
AT1G61820	BETA GLUCOSIDASE 46	CS369123
AT1G66280	BGLU22	CS332191
AT2G03210	FUCOSYLTRANSFERASE 2	SALK_027979
AT2G25630	BETA GLUCOSIDASE 14	CS65839
AT3G10740	ALPHA-L-ARABINOFURANOSIDASE 1	CS351979
AT3G13784	CELL WALL INVERTASE 5	CS341273

Table 3. Mucilage synthesis and secretion candidates. Candidates were identified based on their proximity to 7DPA-activated DHSs. ATG, description, and ABRC stock center ID are listed.

Methods

Sample preparation

Siliques (seed pods) of appropriate ages were collected by first marking young flowers using a fine paint brush and water based paint as described in Western et al., 2001. In brief, recently opened flowers are chosen at the stage the anthers are almost at the same level as the pistil and fertilization is able to occur, usually 2 per plant per day at this stage. The flower is marked with paint and silique collected 4 or 7 days later. Samples were prepared using INTACT nuclei isolation (Deal and Henikoff, 2011) followed by DNase I-seq (Sullivan et al., 2014). A detailed protocol is located at plantregulome.org.

Microscopy

A. thaliana (accession Col-0) seeds were imbibed with water for 1 hour with shaking at 200 rpm. Seeds were then stained with 0.01% Ruthenium red (SIGMA) for 1 hour. Seeds were washed and mounted on a slide and viewed with a light microscope.

Data processing

Each sample was subsampled to 6,650,000 reads. Mitochondrial, chloroplast, and centromeric reads regions from (Clark et al., 2007) converted to TAIR9 coordinates (ch1:13698788-15897560; chr2: 2450003-5500000; chr3:11298763-14289014; chr4:1800002-5150000, chr5:10999996-13332770) were excluded during subsampling.

DHSs and density tracks were generated for the subsampled data using previously described methods (Sullivan et al., 2014). Low-depth 7DPA and 4DPA samples were pooled before subsampling.

Dynamic DHSs

Dynamic DHSs were identified by their relative difference ((7DPA-4DPA)/mean (7DPA, 4DPA)) in DNase I accessibility. In order to calculate relative differences we took the union of DHSs found in any of the 3 samples and tallied per base DNase I cleavages within each union DHS for each sample. We then normalized cleavage tallies for sample quality by dividing by the proportion of DNase I cleavages within 1% FDR threshold hotspots. We defined the relative difference threshold for dynamic DHSs as "-1" for 7DPA deactivated DHSs and "1" for 7DPA activated DHSs. The set of DHSs that reproducibly meet these thresholds make up the "reproducible" dynamic DHS set.

Genomic distribution of DHSs

DHS midpoints were used to determine overlaps with genomic elements. Genomic elements (5'UTR, coding regions, 3'UTR, intergenic, TE) were extracted from the TAIR10 gff file on arabidopsis.org. Some TSSs are not precisely known, so we defined TSSs as 400 bp upstream of the transcription start site. Centromeric regions were excluded from the analysis. To simplify the analysis, only the primary transcript of each gene (AT*.1) was considered. When a single DHS midpoint landed in two different elements, both element overlaps were tallied, thus our overlapping DHS counts sum to greater than the initial number of DHSs. We tallied the total number of base pairs within each element type in the genome, double-counting base pairs that are assigned to overlapping elements.

Integration with expression data sets

Genes from Dean et al., 2011 and Belmonte et al., 2013 were considered to be differentially expressed if there was a 2-fold change in expression between time points. Dean et al., 2011 identify the genes that change 2-fold between 3DPA and 7DPA; the genes that they identify in their supplemental tables were used for integration with dynamic DHS data. The genes that change expression by 2 or more fold in Belmonte et al., 2013 were extracted from the published normalized expression data (Belmonte et al., Dataset S2). Differentially expressed genes were mapped to dynamic DHSs if their TSS was closest to the DHS midpoint (BEDOPS (Neph et al., 2012a); closest-features --closest). Distances to the DHS represented in Figure 3 are the distance between the DHS midpoint and the TSS.

Term enrichment

Term enrichments were performed using DAVID (Huang da et al., 2009). Only the enrichments with a Benjamini-corrected p-value less than 0.05 are shown in Figure 4.

Motif enrichment

Instances of high quality motifs from TRANSFAC (Matys et al., 2006), Franco-Zorrilla et al., 2014, Weirauch et al., 2014, and Sullivan et al., 2014 were determined using FIMO (p-value $1e^{-4}$). Motifs overlapping 7DPA-deactivated, 7DPA-activated DHSs, and all DHSs by 1bp were counted and used in hypergeometric tests in R (phyper) to calculate significant enrichment over all DHSs or the genome. Only the enrichments with a Bonferroni corrected p-value less than 0.05 are shown in Figure 5.

Chapter 4. Mapping regulatory DNA in diverse *A. thaliana* ecotypes[4]

Acknowledgements

James Urton created transgenic lines for each *A. thaliana* ecotype, and DNase I-treated ecotype samples. Kerry Bubb created the ecotype reference genomes, subsampled the data, called hotspots and DHSs, predicted deletions, and identified dynamic DHSs. I assisted Kerry in designing and directing the analysis. I created the figures, interpreted the data in light of the literature, performed GO enrichments and gene family enrichments, and am the author of the text in this chapter. Members of the Stamatoyannopoulos Lab performed DNA sequencing and alignments.

Introduction

Cis-regulatory elements direct the patterns of gene expression that are essential for development and physiology. Modifications of gene expression patterns (i.e. regulatory changes) are hypothesized to be the driving force for evolution of morphological diversity across living organisms (Carroll, 1995; King and Wilson, 1975). Extensive genetic and phenotypic variation exists among *A. thaliana* populations (Koornneef et al., 2004; Mitchell-Olds and Schmitt, 2006) and cis-variations associated with expression changes are common (de Meaux et al., 2005; DeCook et al., 2006; Gan et al., 2011). One study of 18 diverse A. thaliana ecotypes (Gan et al., 2011) showed that extensive variation is concentrated within 100bp of transcription start sites and is associated with expression

[4] This chapter is a part of a manuscript in preparation

changes in adjacent genes. Another study revealed that in addition to *cis*-regulatory variation, deletions, often involving the transcription unit, play a large role in explaining expression level polymorphisms with simple inheritance (Plantegenet et al., 2009). In this chapter we will explore the accessible chromatin landscape in 5 geographically diverse *A. thaliana* ecotypes: Bay-0, Bur-0, Est-1, Tsu-1 and Col-0. Our primary goals are 1) identify regions of the genome which exhibit differences in DNase I accessibility among the accessions, 2) characterize the genetic variation within variable DHSs, 3) describe the genes adjacent to variable DHSs which represent candidates for further study. This project is ongoing; our progress to date is presented here.

Results

DNase I mapping of seven-day-old ecotype seedlings

For each ecotype we created a transgenic line in which nuclei from all cell types are biotin-tagged and can be captured with streptavidin beads (Deal and Henikoff, 2010; Sullivan et al., 2014). Nuclei from seven-day-old light-grown seedlings were treated with DNase I, giving rise to five DNase I-seq libraries that were sequenced and aligned using reference-guided assembly (methods).

Ecotype	Geographic origin	DNase I-seq library	Chr 1-5 reads	SPOT	DHSs (1%FDR)
Bay-0	Bayreuth, Germany	DS22973	10M	0.7522	35,909
Bur-0	Burren, Ireland	DS23077	10M	0.5969	32,417
Est-1	Estland, Estonia	DS22974	10M	0.7714	36,249
Tsu-1	Tsushima, Japan	DS22968	10M	0.7849	35,045
Col-0	Columbia, USA	DS21094	10M	0.5559	32,986

Table 1. *A. thaliana* DNase I-seq samples. Ecotype abbreviations, geographical origins, DNase I-seq library ID, number of reads mapping to Chr. 1-5, SPOT quality score (percentage of reads in hotspots), and the number of DHSs found for each ecotype are listed.

Nine percent of all DHSs identified in the ecotypes are variable

A total of 37,617 DHSs were identified among the five ecotypes. Of these DHSs, 3,311 exhibited variable accessibility in at least one ecotype. DHS variability was determined based on the coefficient of variation (CV) of DHS accessibility across the five ecotypes. A CV threshold for variability was set such that all putative DHS deletions (DHSs with zero DNase I accessibility) would be called as "variable" among the accessions. This threshold captures DHSs in the top 9% CV range (**Figure 1A**).

We found examples of DHSs within predicted genomic deletions (methods) (**Figure 1B**). One such DHS is located in the promoter region of *AGL77,* a MADS-box transcription factor with an unknown function (**Figure 1B**). We also found examples of variable DHSs that were not due to deletions, for example, the intergenic DHS near *microRNA161*; *miRNA161* is predicted to target pentatricopeptide repeat family members for degradation (**Figure 1C**). We also found examples of differential DHSs with complex patterns, such as those near the transcription factor *ABERRANT TESTA SHAPE*, which is important in ovule integument formation in the seed coat (Leon-Kloosterziel et al., 1994)(**Figure 1D**).

Over half of the variable DHSs are most accessible in Col-0

Of the 3,311 variable DHSs, 60% (n=2000) were most accessible in Col-0 (**Figure 1E,** *pattern). An approximately equal number of the remaining DHSs were most accessible

in each of the other ecotypes. The most likely explanation for this skew towards Col-0 DHSs is ascertainment bias introduced by reference-guided assembly (methods); we are missing DHSs in regions of each ecotype genome that are not present in the Col-0 reference. Gan and co-authors have noted that about 10% (14.9Mb) of the Col-0 reference is missing from one or more *de novo* assembled *A. thaliana* ecotype genomes (Gan et al., 2011). Another group has estimated as much as 27% of the Col-0 genome is missing or diverged in at least one ecotype (Zeller et al., 2008). This suggests that it is possible that each of the ecotypes in our study contains large portions of the genome that are not in Col-0. We do not think that this skew towards Col-0 is due to more straight forward alignment issues (e.g. mismatches), because the DHSs that are preferentially accessible in Col-0 have a lower single nucleotide polymorphism rate (1.13% polymorphic sites) than DHSs in which Col-0 is not the most accessible ecotype (1.63% polymorphic sites), excluding DHSs overlapping a deletion in each case.

Variable DHSs are more likely to be in TSSs, TEs, or genes

Next, we asked if the genomic distribution of variable DHSs relative to genomic elements such as TSSs and coding regions, was different from the distribution of all DHSs. We observed that generally variable DHSs are distributed similarly to all DHSs, except they are found more often in TSSs, TEs and within genes, and less often in intragenic regions (**Figure 1F**).

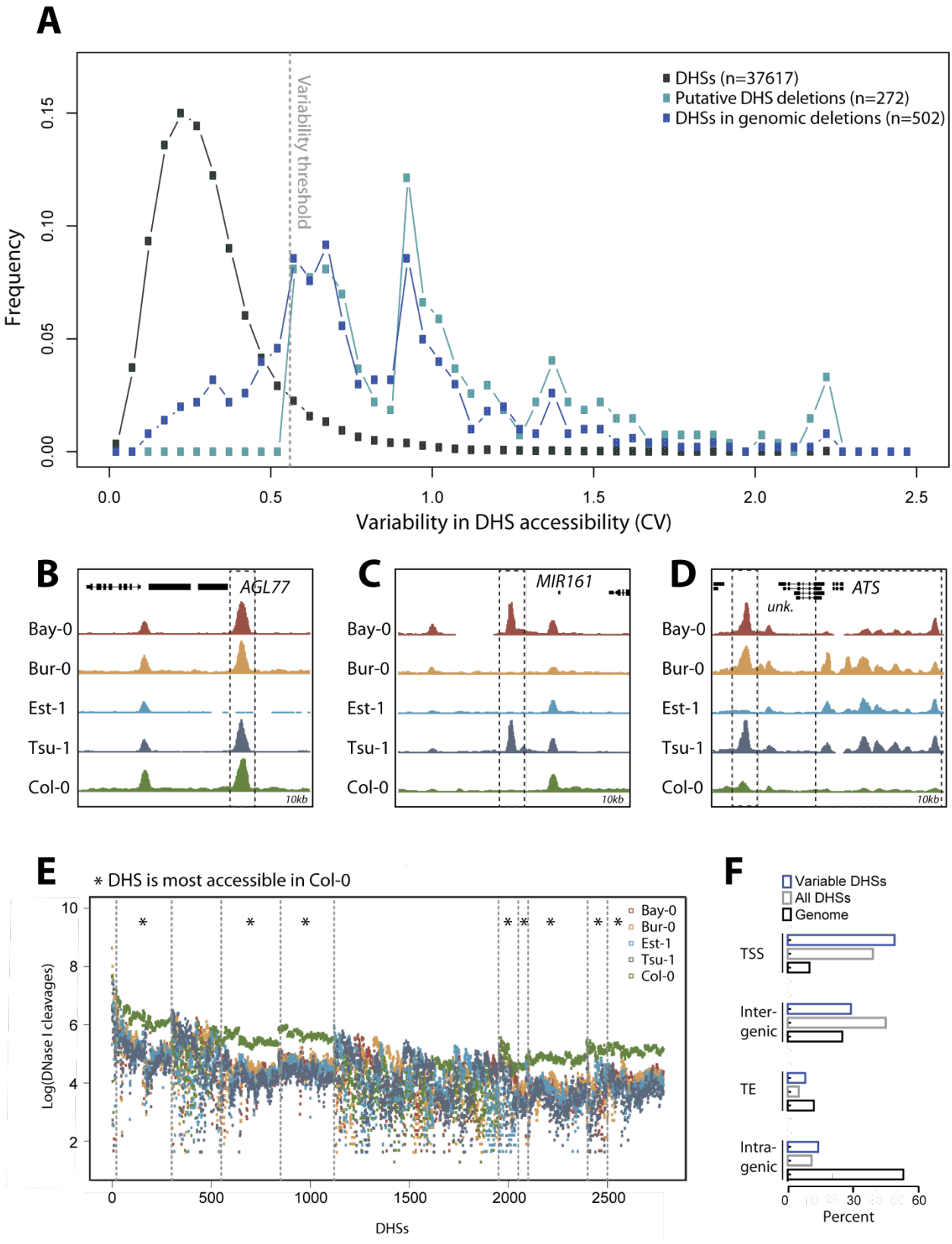

A
DHSs (n=37617)
Putative DHS deletions (n=272)
DHSs in genomic deletions (n=502)
Variability threshold
Frequency
0.00
0.05
0.10
0.15
0.0
0.5
1.0
1.5
2.0
2.5
Variability in DHS accessibility (CV)
B
AGL77
C
MIR161
D
ATS
unk.
Bay-0
Bur-0
Est-1
Tsu-1
Col-0
10kb
E
* DHS is most accessible in Col-0
Log(DNase I cleavages)
2
4
6
8
10
0
500
1000
1500
2000
2500
DHSs
F
Variable DHSs
All DHSs
Genome
TSS
Inter-genic
TE
Intra-genic
0
10
20
30
40
50
60
Percent

Figure 1. Mapping variable DHSs in *A. thaliana* ecotypes. A, Variable DHSs were selected based on the coefficient of variation (CV) in DNase I accessibility across the ecotypes. For a DHS to be considered "variable" it must have a CV greater than 0.5587933 (dashed grey line). Putative DHS deletions (light blue line) and DHSs that sit in predicted deleted genomic regions (dark blue line) in at least one ecotype have CVs primarily above the variability threshold. CV distribution for all DHSs is shown in black. **B,** Example of a variable DHS that is in a genomic region that is deleted in Est-1. **C,** Example of an intergenic, variable DHS that is not predicted to sit in a genomic deletion that is near a miRNA gene. **D,** Example of a region with complex changes in chromatin accessibility near the transcription factor *ATS*. **E,** Clustering analysis revealed patterns of DHS accessibility among the 3,311 variable DHSs. The pattern in which the DHS is most accessible in Col-0 is most common (denoted with a *). **F,** Distribution of variable DHSs (purple), all DHSs (grey), and genomic base pairs (black) within TSSs (400bp upstream of TSS), intergenic regions, transposable elements, and intragenic regions (coding, intron, 5'UTR, 3'UTR).

Variable DHSs are more polymorphic than static DHSs

Next we asked what kind of genetic variation coincided with variable DHSs by examining their overlap with predicted deletions (methods) and single nucleotide polymorphisms ((Gan et al., 2011); unpublished 1001 Genome Data http://1001genomes.org/). Fifteen percent of the variable DHSs (494 of 3,311) overlapped predicted deletions by at least 1bp; 13% of variable DHSs (441/3311) overlapped a large deletion by at least 75bp, half the width of most DHSs. Overall, DHSs have a lower polymorphism rate (0.91% polymorphic sites) than the entire genome (1.01% polymorphic sites). Variable DHSs have an elevated polymorphism rate of 1.35%. Variable DHSs overlapping a deletion in at least one ecotype show an even higher polymorphism rate of 1.41%.

Specific gene functions and families are enriched near variable DHSs

Next we asked if certain kinds of genes are more likely to be located near variable DHSs. To explore this question we mapped each variable DHS to its nearest TSS then used the resulting genes to look for term enrichment and gene family enrichment. We hypothesized that variable DHSs coincided with genes involved in the response to biotic environment, including pathogen defense, since these genes are known to be differentially expressed among *A. thaliana* ecotypes (Clark et al., 2007; Gan et al., 2011). The 3,311 variable DHSs map to 2,960 unique genes. Using term enrichment (Huang da et al., 2009) we found that genes involved in regulated cell death, defense response, and genes associated with chloroplasts or plastids were enriched near variable DHS (**Figure 2**).

Next we looked at gene family data from TAIR and ask if members of certain gene families were more likely to be near variable DHSs. This analysis revealed several families, including Zinc-finger-homeobox genes and Receptor-like kinase (RLK) protein families were enriched near variable DHSs (**Figure 3A**). Because deletions play a large role in disrupting DHSs, we also looked at gene family enrichments for genes (n = 1,626) within large genomic deletions to see if any families overlapped with our variable DHS analysis. We found RLKs and chloroplast and mitochondria gene families were enriched in both deletions and near variable DHSs (**Figure 3B**). It should be noted that none of the gene family enrichments or depletions in deletions or near variable DHSs were significant after Bonferroni multiple testing correction due to the relatively small numbers of genes in each family and the large number of families tested.

A

GENES NEAREST VARIABLE DHSs

Category	Term	Count	p-value
GOTERM_BP_FAT	GO:0006915:apoptosis	42	4.41E-07
GOTERM_BP_FAT	GO:0016265:death	52	4.86E-06
GOTERM_BP_FAT	GO:0008219:cell death	52	4.86E-06
GOTERM_BP_FAT	GO:0012501:programmed cell death	46	1.51E-05
INTERPRO	IPR000157:Toll-Interleukin receptor	31	1.03E-05
INTERPRO	IPR002182:NB-ARC	36	5.50E-06
GOTERM_BP_FAT	GO:0006955:immune response	54	6.70E-05
GOTERM_CC_FAT	GO:0009535:chloroplast thylakoid membrane	54	5.43E-05
GOTERM_CC_FAT	GO:0055035:plastid thylakoid membrane	54	5.43E-05
GOTERM_CC_FAT	GO:0044436:thylakoid part	59	5.26E-04
GOTERM_CC_FAT	GO:0044435:plastid part	138	4.74E-04
GOTERM_CC_FAT	GO:0031090:organelle membrane	122	4.11E-04
GOTERM_CC_FAT	GO:0034357:photosynthetic membrane	57	2.66E-04
GOTERM_CC_FAT	GO:0009570:chloroplast stroma	67	7.22E-04
GOTERM_CC_FAT	GO:0044434:chloroplast part	135	3.78E-04
GOTERM_CC_FAT	GO:0009534:chloroplast thylakoid	57	9.39E-04
GOTERM_CC_FAT	GO:0031976:plastid thylakoid	57	9.39E-04
GOTERM_CC_FAT	GO:0031984:organelle subcompartment	57	0.00108
GOTERM_CC_FAT	GO:0009532:plastid stroma	68	0.001716
GOTERM_CC_FAT	GO:0042651:thylakoid membrane	56	5.79E-05

CELL DEATH

DISEASE RESISTANCE

CHLOROPLAST / PLASTID

Figure 2. A, Top GO term enrichments (Huang da et al., 2009) for genes located near variable DHSs.

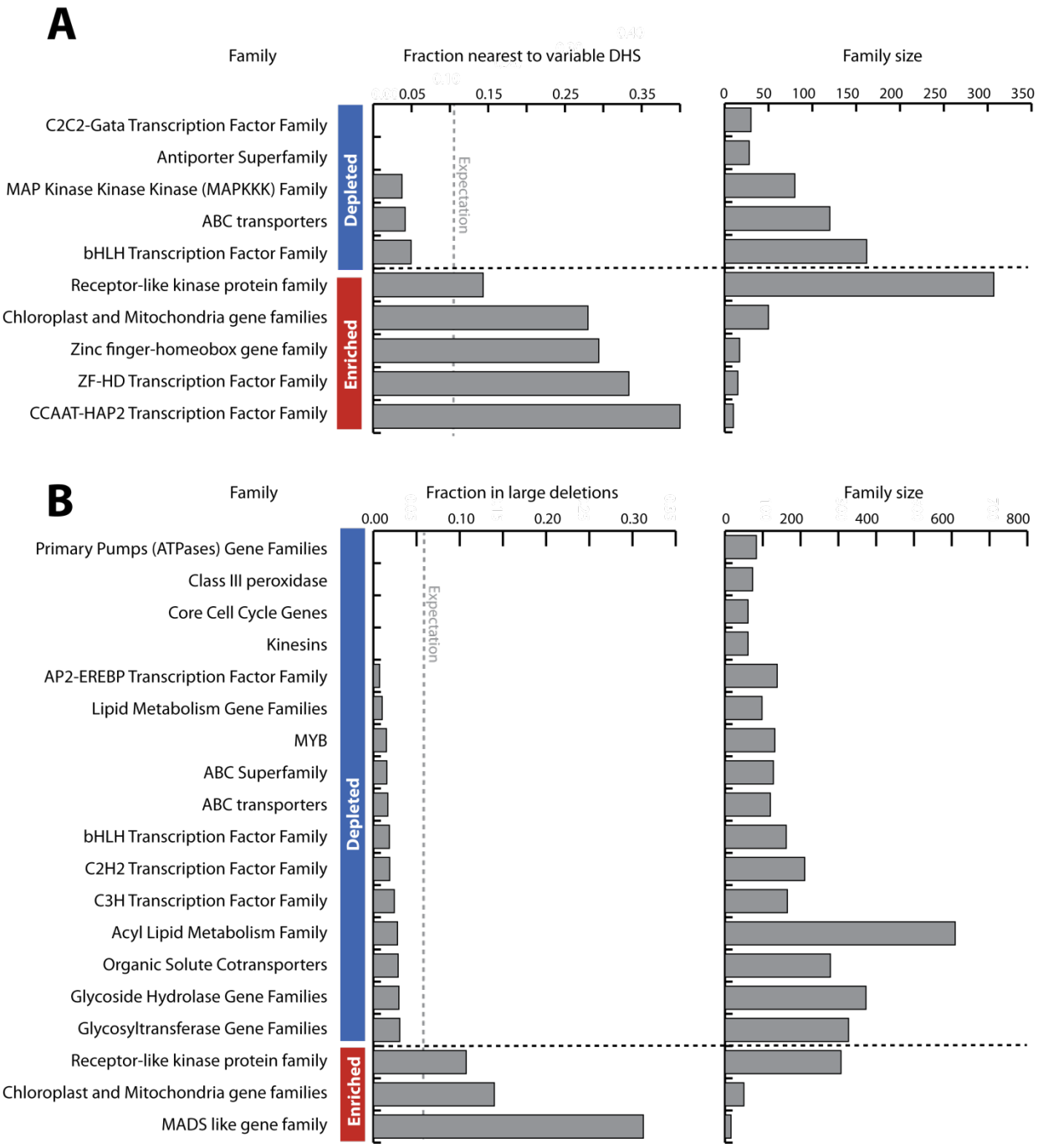


Figure 3. Gene family enrichment near variable DHSs. A, The fraction of gene family members that are also the nearest TSS to a variable DHS (left) are shown together with their corresponding family sizes (right). **B,** The fraction of gene family members that are found within large genomic deletions (left) are shown together with their corresponding family sizes (right).

Discussion

Our findings show that genetic variation is enriched within variable DHSs, and that a large portion of variable DHSs can be explained by larger (>300bp) deletions, which had been suggested previously (Plantegenet et al., 2009). We also observe what appears to be ascertainment bias towards Col-0 DHSs, possibly due to our reference guided assembly strategy for comparing the ecotypes with Col-0. This is concerning because we may be missing diverged and important variable DHSs in the ecotypes. We confirmed that extensive variation coincides with defense related genes. We also showed that cell death related genes and chloroplast related genes are enriched near variable DHSs; many cell death related genes are related to disease resistance and defense.

We found certain gene families to be enriched or depleted near variable DHSs. One enriched family was the receptor-like kinase family. RLKs are membrane-bound, signal transduction proteins that phosphorylate serine/threonine residues. They are involved in signaling in development, environmental response, and immunity and share a common origin with animal serine/threonine and tyrosine kinases (Shiu and Bleecker, 2001). In A. thaliana, 30% of RLK family members exist in tandem repeats of 2 to 19 genes, suggesting these genes may have arisen from tandem and large-scale duplication events (Shiu and Bleecker, 2001). This tandem arrangement of RLKs may in part explain their enrichment near variable DHSs and within deletions if the tandem configuration leads to increased genetic instability.

A major goal of the A. thaliana community has been to identify the molecular polymorphisms that control fitness relevant traits in *A. thaliana* ecotypes; many of these

molecular polymorphisms will be regulatory in nature. Traditional mapping approaches can identify *cis*-regulatory variation responsible for the phenotypic differences (Clark et al., 2006), however, these studies are difficult and limited in terms of resolution (Weigel and Mott, 2009). GWAS studies work well in identifying regulatory SNPs associated with phenotypic (Atwell et al., 2010b; Gan et al., 2011; Meijon et al., 2014) or expression level variation (Gan et al., 2011) with good resolution, but often identifying the "candidate" SNP is difficult since linkage disequilibrium between loci <50kb from each other is substantial *A. thaliana* (Mitchell-Olds and Schmitt, 2006). Furthermore, GWAS works best with common variants. Although well-characterized examples of regulatory variation associated with phenotypic changes are rare, several studies have demonstrated that natural variation in promoter elements accounts for differing expression levels of potent flowering-time genes *FLC* (Li et al., 2014) and *FT* (Liu et al., 2014).

By mapping the active and poised regulatory regions of *A. thaliana* ecotype seedlings we have created a resource that can help guide mapping and GWAS studies to regions of the genome that are potentially functional and should be prioritized in candidate searches. From our own previous work we know highly trait-associated SNPs are enriched in DHSs (Sullivan et al., 2014). Delineation of these regulatory elements will also provide an orthogonal validation for studies of individual promoter variation in the accessions. In the near future we will continue this work, first, to look whether highly trait-associated variants identified in GWAS are enriched in variable DHSs, and second, to see if expression quantitative trait loci (eQTL) are enriched in variable DHSs relative to conserved or static DHSs.

Methods

Plant material

Transgenic INTACT lines harboring the UBQ10:NTF construct (Sullivan et al., 2014) and ACT2:BirA construct (Deal and Henikoff, 2010) were created for each ecotype using double agrobacterium transformation. Transformants were selected on plates containing BASTA (15uM) and KAN (50ug/ml for all ecotypes except Ler-1 which received 15ug/ml). Transgenic lines are available from ABRC under the following accession numbers: CS68650 (Bay-0), CS68651 (Bur-0), CS68652 (Cvi-0), CS68653 (Est-1), CS68654 (Ler-1), CS68655 (Shah), CS68656 (Tsu-1), CS68649 (Col-0).

Sample preparation

*S*eeds (0.1g) were surface sterilized by treating with 70% EtOH with 0.5% triton for 10 minutes followed by 5 minutes in 95% EtOH. Seeds were dried completely on sterile filter paper and plated on 150mm petri plates containing 50ml 1XMS with 0.8% agar covered by a sterile #1 filter circle cut to size (Whatman, GE Healthcare UK ltd). Plates were sealed with micropore tape, double wrapped with aluminum foil and stratified for 3 days at 4C. Stratified plates were unwrapped and moved to LD conditions (16hr light, 22°C; 8hr dark, 20°C) in a growth chamber (Conviron CMP5090, Controlled environment ltd. Winnipeg, Manitoba, Canada) and grown for 7 days. Samples were harvested at the same time of day and nuclei were collected and DNase I treated as in Sullivan et al., 2014.

Preparation of ecotype data

Short read data (single and paired-end reads) for Bur-0, Bay-0, Est-1, Tsu-1 and Col-0 were downloaded from the 1001 genomes project for *A. thaliana* ((Gan et al., 2011) http://1001genomes.org/index.html). Short read data were used to generate ecotype reference genomes in which small polymorphic sites in the reference genome (Col-0) were replaced with the ecotype variant using reference-guided assembly. DNase I-seq reads were aligned to the appropriate reference genomes. ChrC and ChrM reads, and centromeric regions from (Clark et al., 2007) (ch1:13698788-15897560; chr2: 2450003-5500000; chr3:11298763-14289014; chr4:1800002-5150000, chr5:10999996-13332770) were filtered out, and the remaining reads from each sample were subsampled to 10 million reads. Per-base DNase I cleavages, hotspots, and DHSs (peaks) were called on the subsampled data sets as in Sullivan et al., 2014.

Identification of differential DHSs

Bay-0, Bur-0, Est-1, Tsu-1 and Col-0 DHSs were merged to create a "union" set of 47,022 DHSs. Per-base DNase I cleavages within each merged DHS were calculated for each ecotype. DNase I cleavages within each DHS were then summed across all five ecotypes; DHSs with sums in the bottom 20% were excluded from further analysis, leaving 37,617 remaining DHSs from which variable DHSs were determined. Variable DHSs were then identified based on their coefficient of variation (CV), which is the standard deviation of DHS accessibility across the five ecotypes, divided by the mean in DHS accessibility across the five ecotypes. The CV threshold (CV = 0.5587933) was chosen because all DHSs that are predicted to be affected by a deletion (i.e. DHSs with zero per base DNase I cleavages in at

least one ecotype) had a CV greater than or equal to the threshold. This CV threshold corresponds to the top 9th percentile of DHS variability.

Identification of predicted deletions >300bp in length in the ecotypes

Deletions were identified by calculating the mean x-coverage in all 150bp sliding windows (overlap = 20bp) for each ecotype reference using the 1001 genomes (http://1001genomes.org/) short read data. The x-coverage in each window was normalized by dividing by the x-coverage observed for Col-0 short reads mapped back to the reference, which controls for regions of the genome that are more readily sequenced with short reads. We identified putative deletions by taking the 150bp windows in the bottom 1% of normalized coverage and merged overlapping windows. We then merged putative deletions from different ecotypes that were within 1kb of each other, to generate the predicted deletions set used in this analysis.

Variable DHSs near gene families & GO term enrichments

Gene family information was downloaded from TAIR (gene_families_sep_29_09_update.txt). Family members that did not have a corresponding ATG number were filtered out. Overlaps between genes nearest to variable DHSs and family members were tabulated. Hypergeometric tests (phyper in R; q = number of genes nearest to a variable DHS in each family, m = number of genes nearest to variable DHSs in the background, n = number of gene that are not nearest to a variable DHS in the background, k = number of family members in the family) were used to determine the significance of the overlaps between genes nearest to variable DHSs or genes in deletions and family members. GO term, KEGG pathway, and INTERPRO enrichments were performed using DAVID (Huang da et al.,

2009). Only the enrichments that had a Benjamini corrected p-value smaller than 0.05 are presented in Figure 2.

Chapter 5. Conclusions and future directions

The large collections of DNase I mapping data (atlas of *cis*-elements, dynamic elements, TF footprints, and TF networks) generated and analyzed as part of my dissertation work have generated a resource for the plant community. Aside from this large resource, the data support the notion that plants have complex, dynamic, developmentally regulated patterns of *cis*-regulatory elements more similar to that of an animal, for example *Drosophila*, than that of single-celled yeast in which few distal elements exist and promoters tend to follow a simple recipe of upstream activating or silencing elements (Levine and Tjian, 2003). This similarity is perhaps not surprising given that plants are multicellular and muticellularity requires regulatory complexity. Despite this fact, animal-centric regulatory literature (Levine, 2010a) often leaves the impression that transcriptional regulation in plants is somehow less complex than their multicellular metazoan counterparts. On account of its apparent complexity, *A. thaliana* (and other plant) regulatory DNA warrants further study. Below I propose follow up studies that are based on the conclusions drawn in previous chapters; these studies are aimed at outstanding questions about plant transcriptional control.

Assigning function to individual DHSs

Rationale: While chromatin accessibility mapping in *A. thaliana* improved our understanding of regulatory elements, the factors that bind them, and their dynamics, we are still missing half the picture: we do not have a good way of connecting regulatory elements, especially the distal ones, to the genes they regulate (**Figure 1A**). Our current analytical methods assume regulatory elements are near their target genes. This assumption is passable because 1) more than a third of DHSs fall within 400bp of the

transcription start site suggesting they are part of proximal promoter machinery (Sullivan et al., 2014; Zhang et al., 2012b)(**Chapters 2, 3, & 4**), 2) genes nearest to dynamic DHSs implicate known biological processes or are differentially expressed genes (Sullivan et al., 2014)(**Chapter 2 & Chapter 3**), and 3) the organization of *cis*-regulatory elements in the similarly compact genome of *Drosophila* indicates that the greatest proportions of developmental enhancers are gene-proximal or intragenic (Kvon et al., 2014). However, we know proximity assignments do not always work. For example, in vertebrate development, the enhancer controlling *Shh* expression in the developing limb bud is located in the intron of a second gene *Lmbr1* (Amano et al., 2009) which is more than a megabase from *Shh* (Calhoun and Levine, 2003). From our own data, we know that a quarter of differentially expressed genes have a dynamic DHS within 5 kb (**Chapter 3**). The average distance between genes in *A. thaliana* is just 1kb, therefore, even at a distance of 1 or 2kb from the TSS, proximity assignments can be problematic for "choosing the correct gene".

Connecting regulatory elements to their target genes is a difficult problem that generally requires the use of transgenic reporter constructs or *in situ* hybridizations (Kvon et al., 2014; Nord et al., 2013). Because these methods are labor intensive, ideally they would not be used without some prior knowledge of the location or function of regulatory elements. Alternatively, chromatin conformation capture (3-C) based assays such as Hi-C and ChIA-PET (Dekker et al., 2013; van Steensel and Dekker, 2010) can capture physical interactions between distal elements and promoters. Three-C has been used to measure the long-range (40-60kb) interactions between enhancer elements of the locus control region and the promoters of active globin genes (Tolhuis et al., 2002). However, genome-wide 3-C based technologies are technologically challenging, do not distinguish between functional

and non-functional interactions, do not capture dynamics, are muddled by the great degree of heterogeneity in chromatin architecture among cells, and for the most part cannot identify the proteins responsible for the observed association (Dekker et al., 2013).

Experiment: Map the function of developmental enhancers through targeted DHS deletion. Classical genetics is based on the experimental paradigm of knocking out gene function and assaying molecular and morphological phenotypes to determine gene function. Using the same logic, deleting DHSs (and the footprints inside them), should reveal their regulatory function. A major advantage of this strategy is that the regulatory element is studied in its native chromatin context rather than on a plasmid or in newly integrated T-DNA.

Targeted deletions can be achieved using CRISPR/Cas system which has been successfully used in *A. thaliana,* as well as several crops (Feng et al., 2014b; Jiang et al., 2013; Miao et al., 2013; Shan et al., 2013). The most common mutation introduced by CRISPR/Cas are 1bp insertions or deletions; however the second most common mutation is a small (<20bp) deletion, which is ideal for deleting footprints. Guide RNAs can be multiplexed to target more than one site simultaneously (Cong et al., 2013), which raises the possibility of tiling guide RNAs across a single DHS to generate a library of mutations, or even targeting multiple DHSs simultaneously to create a more complex library of DHS mutations. If the technology continues to develop, it may eventually be possible to use CRISPR/Cas to create "DHS traps" by inserting GFP/LUC reporter constructs with minimal promoters in a targeted fashion to locations adjacent to DHSs, though this is to my knowledge not yet possible.

Evaluation of the effects of targeted DHS deletion can be done in several ways. The most straightforward method is to measure the transcriptome and ask which genes have altered expression patterns (**Figure 1B)**. A second strategy would be to target DHSs in plants that are carrying an enhancer or gene trap construct that exhibits tissue specific expression (**Figure 1C, D)**. If the enhancer or gene trap locations were known (most are not, though a large collection of these lines exist in *A. thaliana*), DHSs near the trap (minimal promoter fused with GFP or LUC) could be specifically targeted and then changes in reporter construct expression patterns would be assayed in the whole organism.

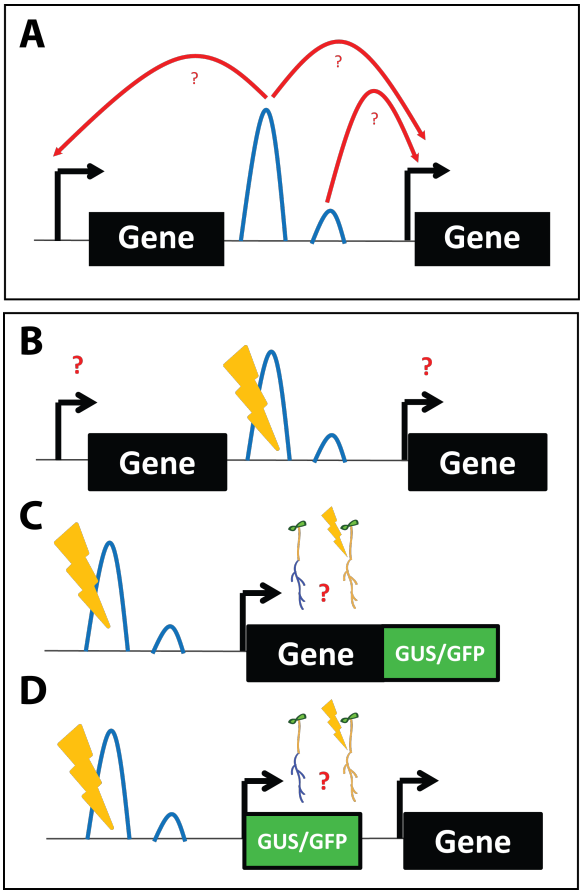


Figure 1. Assigning function to individual DHSs. A, Schematic illustrating the ambiguity of mapping DHSs to their target genes using proximity alone. This is especially true when genes are close to each other as is the case in *A. thaliana*. Light blue lines are DHSs, red arrows are regulatory interactions, yellow thunderbolts are targeted deletions that may result in loss of reporter construct activity (plantlets with blue roots vs. normal roots).**B,** Targeted deletion of a DHS and subsequent analysis of gene expression of nearby genes. **C,** Targeted deletion of a DHS near a trapped gene with a root specific expression pattern. **D,** Targeted deletion of a DHS near a minimal promoter enhancer trap which displays a root specific expression pattern.

Identification of plant protein domains with insulator properties

Rationale: Insulators are DNA elements that block enhancer promiscuity and prevent chromatin domains from spreading (West et al., 2002). In general insulator sequences bind proteins or protein complexes that block enhancers from interacting with target promoters or form a barrier to advancing chromatin domains (West et al., 2002). Proteins with blocking/barrier function vary among organisms; most of these proteins have been identified in *Drosophila*. CTCF is the best-known insulator binding protein that blocks enhancers (Bell et al., 1999). CTCF is also found at the boundaries of chromatin domains that are topologically associated in *Drosophila* and human (Dixon et al., 2012; Ho et al., 2014). To date, proteins with insulator capabilities have not been identified in plants, and plants do not have CTCF. Plant matrix attachment regions (MARs), similar to animal MARs, have enhancer-blocking function in plants but the proteins involved are unknown (Hilly, Singer Liu, Plant Cell Reports, 2009). Interestingly, *A. thaliana* does not appear to have topologically associated chromatin domains similar to those found in animals (Feng et al., 2014a), though they are predicted to exist based on patterns of co-expressed genes (Zhan et al., 2006). These observations lead to the following question: how are plant enhancers and chromatin domains insulated?

Experiment: Screen plant protein domains for insulator properties in a yeast system. To identify *A. thaliana* protein domains with chromatin blocking capabilities, I would take advantage of a previously developed assay in yeast (Fourel et al., 2001). Briefly, in this system two different selectable markers are separated by four Gal4-binding sites and are inserted in proximity to the telomere VIIL. When telomeric silencing is unhindered, both

selectable markers are inactive. When a chimeric protein containing the Gal4 DNA binding domain fused to an insulator protein (or protein domain) binds to the Gal4-binding sites and blocks telomeric silencing of the selectable marker furthest from the telomeric sequences, only one selectable marker is active. This method was successfully used to identify two insulator domains of the chicken CTCF (Defossez and Gilson, 2002). Using this assay a library of barcoded protein domains from *A. thaliana* would be fused to the Gal4 DNA binding domain. Yeast colonies that are positive for containing a functional insulator domain would be harvested and sequenced to reveal the identity of functional plant insulator domains. While not outlined here, a similar type of reporter construct could be designed to assay *A. thaliana* protein domains with enhancer blocking function.

Final thoughts

There are questions that remain about plant chromatin dynamics and gene regulation. The first question that interests me is how cell lineages, which were derived independently in animals and plants, have left their mark on chromatin landscapes. In humans, cell fate, lineage, and maturity are recorded in DHS patterns (Stergachis et al., 2013b). In plants it is still too early to tell whether this is the case, we simply have not looked in enough cell types to be sure. However, we have reasons to believe plants will exhibit differences. For example, plant cells cannot be made into immortalized, differentiated cell lines the way human cells can; in fact when put into culture with just two hormones, nearly all plant cells dedifferentiate into totipotent callus cells. Plants have fewer cell lineages established during embryogenesis; one could argue that plants really have only two: root and shoot (Laux et al., 2004). Most plant cell types arise from pools of

stem cells called meristems after embryogenesis has finished. Using meristems, plants develop in a continuous fashion in step with environmental cues.

The next question that interests me is whether shadow enhancers exist in plants. Shadow enhancers are functionally redundant enhancers, which allow for developmental robustness in animals (Hong et al., 2008; Perry et al., 2010). Their existence has not been shown in plants, though it seems likely they should exist. Targeted deletion of DHSs may uncover such elements.

The next question that interests me is the effect of genome size on regulatory element organization. I often like to say that nothing is "gene-distal" in Arabidopsis. I say this because the genome is small (152Mb) and is packed with ~32,000 genes. The only other plant in which DHSs have been mapped is rice, which still has a relatively small genome of 389Mb (International_Rice_Genome_Sequencing_Project, 2005). Without yet knowing the function of DHSs in rice and *A. thaliana* it is tempting to think that maybe distal regulatory elements really do not exist. However, we know long-range, distal elements exist in *Drosophila,* which has a similarly compact genome to *A. thaliana* (Levine, 2010a). Maize, on the other hand, has a 2.3 Gb genome (Schnable et al., 2009), more similar to the size of the human genome. Interestingly, the only long-range regulatory element that has been identified in plants is the *tb1* enhancer in maize, which acts at distances greater than ~60kb (Clark et al., 2006). My question is whether the distal nature of an element is just an artifact of genome size, or if it actually serves a purpose.

Finally, I would like to know if regulatory elements, such as enhancers, in plants can be predicted from chromatin signatures as has been found in animals (Heintzman et al.,

2009; Heintzman et al., 2007). The limitation here is the availability of high quality ChIP-seq experiments on chromatin marks, and also the availability of data from the exact tissues and conditions.

Appendix A. A fast-evolving plant polymerase

Project summary

Pol IV (NRPD1and Pol V (NRPE1)) are two unusual DNA-dependent RNA-polymerases involved in RNA-directed DNA methylation (RdDM) in plants. RdDM leads to *de novo* cytosine methylation in all sequence contexts (CG, CNG, and CNN where N = A,T,C) and epigenetic silencing at target loci (Matzke et al., 2007). The two polymerases arose from duplications in Pol II and continue to share subunits with each other and with Pol II (Huang et al., 2009; Ream et al., 2009). Their functions in RdDM, however, appear largely non-redundant.

While screening candidate genes for novel phenotypic capacitors, Janne Lempe, a post-doc in the Queitsch lab, analyzed the developmental stability of various isogenic mutants involved in small RNA pathways. Her analysis revealed that mutations in the largest (NRPE1) subunit of Pol V significantly increased developmental instability (**Figure 1**). Furthermore, like Hsp90, the Pol V mutation revealed the effects of cryptic genetic or epigenetic variation on several morphological traits in *A. thaliana* accessions Ler and Col-0. Pol V also contained a highly variable C-terminal glutamine-serine repeat (QS) (**Figure 2-4**) that we hypothesized would be important for capacitor function.

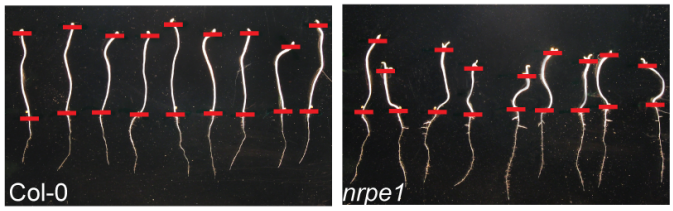


Figure 1. Images of dark grown wild-type phenotypically stable Col-0 seedlings and variable *pol V* (*nrpe1*) mutant seedlings. Area marked by red lines indicates the embryonic stem, or hypocotyl, which is used as a measure of developmental instability.

Sadly, I found the phenotype of *nrpe1* (*drd3-1*)was actually due to a background mutation in the strain in which the mutagenesis screen was carried out (Kanno et al., 2005). The background mutation was closely linked to the *nrpe1* allele, and thus in crossing experiments it appeared that nrpe1 was responsible for the phenotype. A complementation experiment between the *nrpe1* allele and the background (DRD) revealed 1) that the background did not complement the *nrpe1* mutation, and 2) ruled out the possibility that the DRD background, which looked to have a similar phenotype to *nrpe1*, was not due to "something bad happening to the tube of seeds". Cristina Alexandre, another post doc in the Queitsch lab, has since shown that the variable phenotype of *nrpe1* is actually due to the disruption of expression of *LTP2*, a lipid transfer protein important for cell wall expansion.

Pol V may not play a role as a capacitor in phenotypic robustness, but it does play an important role in silencing transposable elements in the genome (Zhong et al., 2012). Thus, the QS-repeat and its fast evolving nature is perhaps still interesting for future studies (**Figure 2-4**).

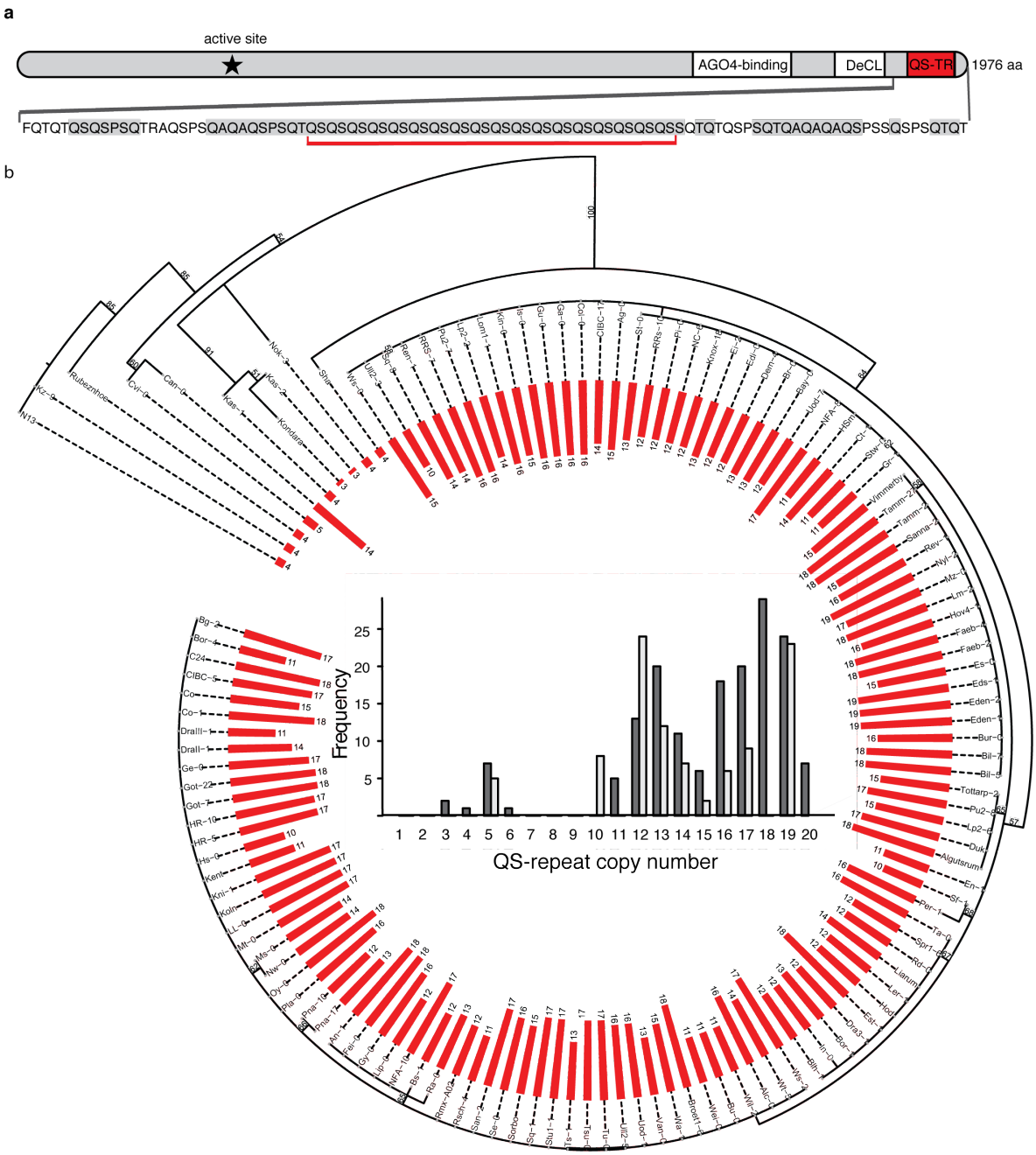

a
active site
AGO4-binding
DeCL
QS-TR
1976 aa
FQTQTQSQSPSQTRAQSPSQAQAQSPSQTQSQSQSQSQSQSQSQSQSQSQSQSQSQSQSQSQSSQTQTQSPSQTQAQAQAQSPSSQSPSQTQT
b
Frequency
QS-repeat copy number

Figure 2: *NRPE1* is hypervariable in global and local *A. thaliana* accessions. a, The highly variable QS-TR (shaded in red) is located close to the NRPE1 C-terminus. Hotspots for additional polymorphisms are in grey (Supplementary Figs. 3 and 4 for accession-specific polymorphisms). NRPE1 QS-TR copy number ranged from 3-20 in global accessions sample (inset, Supplementary Fig. 2). Most accessions carried long TRs (≥10 QS), and a few carried short TR (≤5). Certain QS-TR alleles (7-9 QS) were not observed. This pattern of *NRPE1* QS-TR variation was mirrored in a local accession sample (Supplementary Fig. 4). In both samples, we found additional polymorphisms adjacent to the QS-TR, which tended to be shared either among accessions with short or those with long QS-TR (Supplementary Figs 3 and 4). Common ancestry may explain this pattern, particularly in Central Asian and Southern Russian accessions (Cao et al., 2011b); however, short QS-TRs were also present in accessions from other regions and in individuals of the local sample (Supplementary Tables 6 and 7). **b,** As short TRs are unlikely to re-expand, intragenic suppressors may compensate for their potentially deleterious effects. WMax Press and I asked whether short QS-TRs correlated with increased variation elsewhere in *NRPE1* by building a maximum likelihood tree with 131 *NRPE1* sequences (Cao et al., 2011b), excluding the polymorphic C-terminus. Accession-specific QS-TR copy number is represented with proportional red bars. Most branches had low bootstrap support, those ≥ 50 are displayed. Accessions with short QS-TRs were well separated from those with long QS-TRs with the exception of the island accession Can-0, which shares a branch with Cvi-0, another island accession. This separation defined the *NRPE1* divergence class. Tree inset**:** distribution of NRPE1 QS-TR alleles in a sample of 164 global accession and in a local sample of 12 accessions with 8 individuals each highlights the absence of some QS-TR alleles.

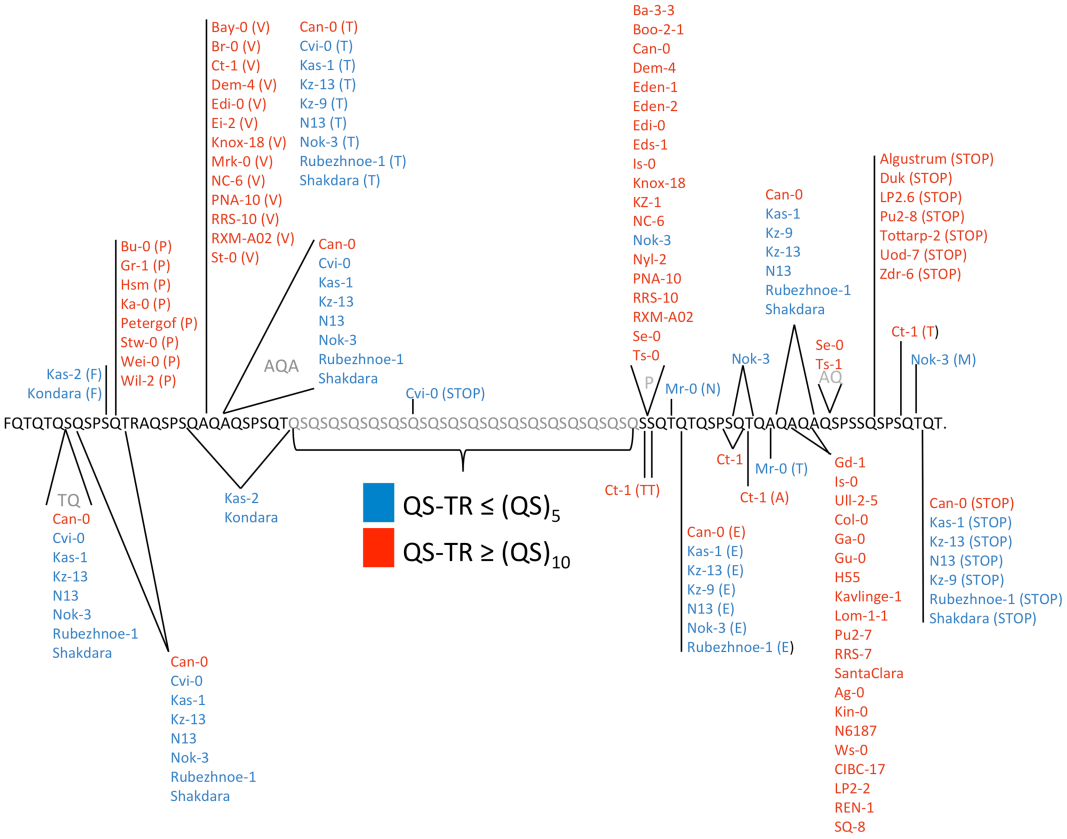


Figure 3: C-terminal polymorphisms in 164 global *A. thaliana* acessions. Acessions with short QS-TR alleles are in blue, accessions with long QS-TR are in red. Amino acid substitutions are represented as single lines, the substituted amino acid is indicated in parenthesis after the accession name. Insertions are represented as open-faced ∧∨ symbols with inserted amino acids represented in grey. Deletions are represented as close-faced ∧∨ surrounding deleted amino acids.

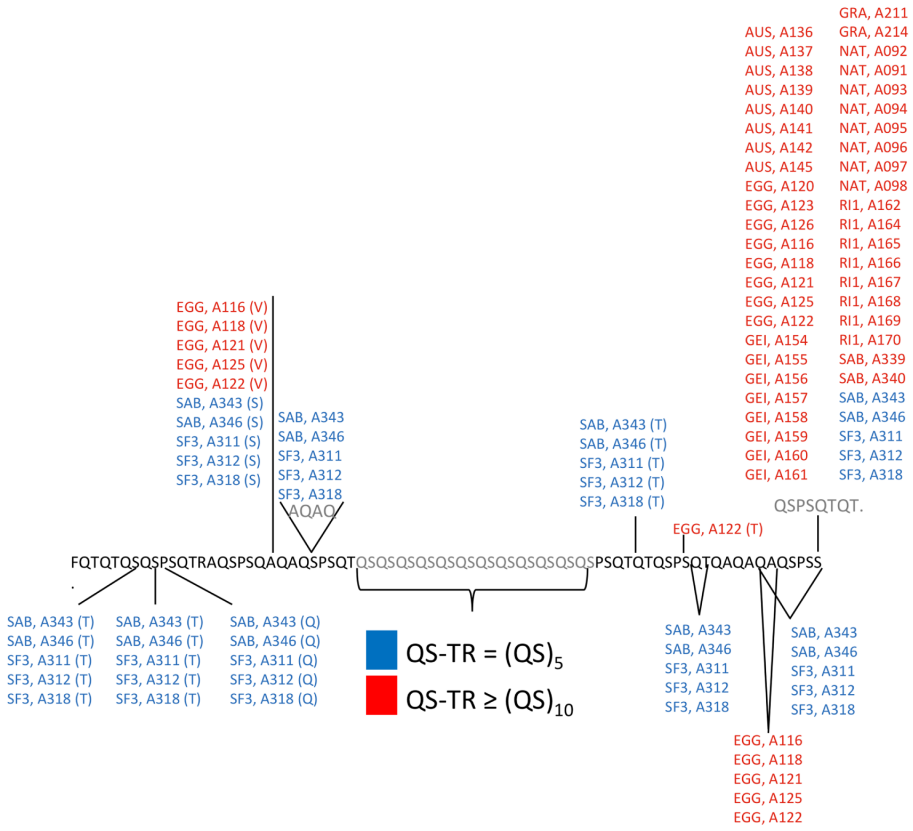


Figure 4: C-terminal polymorphisms in 12 local *A. thaliana* accessions containing a total number 96 individuals plants from the Swiss Alps. Each individual is labeled according to their accession (location abreviation) followed by a unique identifier. Individuals with short QS-TR alleles are in blue, individuals with long QS-TR are in red. Amino acid substitutions are represented as single lines, the substituted amino acid is in parenthesis after the unique identifier. Additional amino acid sequence after the alternate translation stop site is shown in grey. Insertions are represented as open-faced ∧∨ symbols with inserted amino acids represented in grey. Deletions are represented as close-faced ∧∨ surrounding the deleted amino acids.

Methods

I amplified and sanger-sequenced the variable *NRPE1* tandem repeat using ExTaq polymerase (Takara Bio INC) and primers 14 (Supplementary Table 15) in 164 global accessions (Nordborg et al., 2005; Shindo et al., 2005) and 96 individuals representing 12 local accessions from the Swiss Alps. The following PCR cycle was used for amplification: Step 1 95°C 10:00; Step 2 95°C 00:30; Step 3 56°C 00:30; Step 4 72°C 01:00; Step 5 GO TO Step 2, 29 more times; Step 6 72°C 5:00. PCR products were sequenced from both ends, and aligned and trimmed using DNASTAR® (Lasergene 8.0 software suite). QS-TR units were counted. To build the *NRPE1* gene tree for the sample of global accessions, Max Press and I used 1001genomes data (Cao et al., 2011b) (www.1001genomes.org) for all accessions with QS-TR sequence data. Non-coding regions and the polymorphic C-terminus (after amino acid position Q1876) were excluded for this analysis. The remaining sequences were aligned using Clustal Omega (Sievers et al., 2011) and used as input for PhyML to build a maximum likelihood tree (Guindon et al., 2005). Support for all branches was assessed with 100 nonparametric bootstraps. Tree visualization and labeling was performed with iTOL (Letunic and Bork, 2007).

Appendix B. Supplementary Materials for Chapter 2

Contents

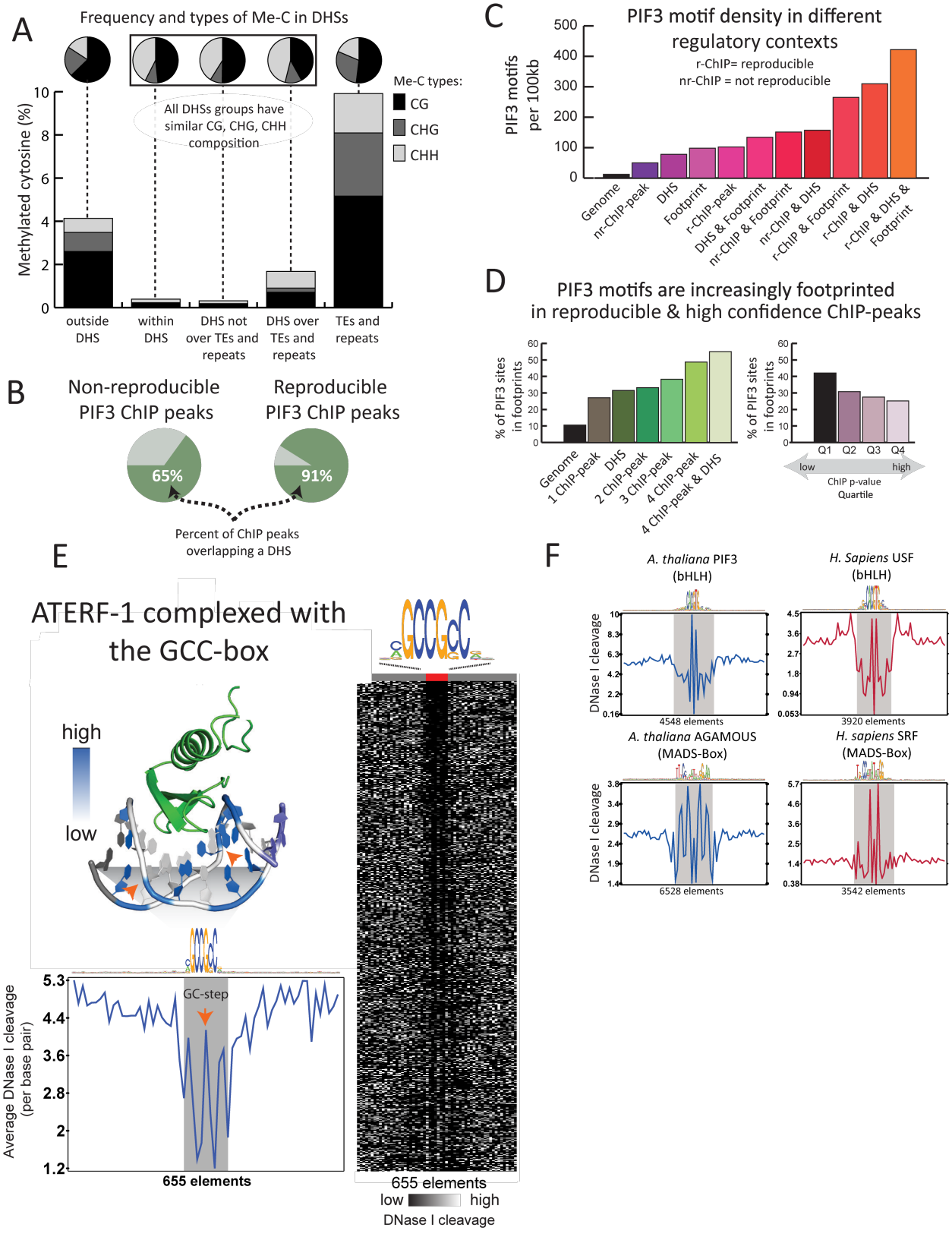

A
Frequency and types of Me-C in DHSs
Methylated cytosine (%)
All DHSs groups have similar CG, CHG, CHH composition
Me-C types:
CG
CHG
CHH
outside DHS
within DHS
DHS not over TEs and repeats
DHS over TEs and repeats
TEs and repeats
B
Non-reproducible PIF3 ChIP peaks
Reproducible PIF3 ChIP peaks
65%
91%
Percent of ChIP peaks overlapping a DHS
C
PIF3 motif density in different regulatory contexts
r-ChIP= reproducible
nr-ChIP = not reproducible
PIF3 motifs per 100kb
Genome
nr-ChIP-peak
DHS
Footprint
r-ChIP-peak
DHS & Footprint
nr-ChIP & Footprint
nr-ChIP & DHS
r-ChIP & Footprint
r-ChIP & DHS
r-ChIP & DHS & Footprint
D
PIF3 motifs are increasingly footprinted in reproducible & high confidence ChIP-peaks
% of PIF3 sites in footprints
Genome
1 ChIP-peak
DHS
2 ChIP-peak
3 ChIP-peak
4 ChIP-peak
4 ChIP-peak & DHS
Q1
Q2
Q3
Q4
low
high
ChIP p-value Quartile
E
ATERF-1 complexed with the GCC-box
high
low
Average DNase I cleavage (per base pair)
GC-step
655 elements
655 elements
low
high
DNase I cleavage
F
A. thaliana PIF3 (bHLH)
H. Sapiens USF (bHLH)
DNase I cleavage
4548 elements
3920 elements
A. thaliana AGAMOUS (MADS-Box)
H. sapiens SRF (MADS-Box)
6528 elements
3542 elements

Figure S1. The *cis*-regulatory landscape of *A. thaliana*. DHSs were significantly depleted for cytosine methylation (*P < $2.2x10^{-16}$) and contained significantly different ratios of cytosine methylation contexts (CG:CHG:CHH; *P < $1.0x10^{-50}$). DHSs had similar ratios of cytosine methylation contexts irrespective of whether the DHSs coincided with transposons and repetitive DNA. Here, 'n' is the sum of per-cytosine methylation frequencies found in immature flowers for all cytosines of a given context (CHH, CHG, CG) (Lister et al., 2008). **B,** PIF3 ChIP-seq peaks (Zhang et al., 2013) overlapped with DHSs. Reproducible ChIP-seq peaks were more likely to overlap with DHSs than non-reproducible ChIP-seq peaks. **C,** PIF3 motif density, measured as number of motifs per 100kb, steadily increased when regulatory DNA was stratified by reproducibility of ChIP-seq signal (Zhang et al., 2013) and DNase I accessibility. **D,** The percentage of occupied PIF3 motifs increased with ChIP-seq peak reproducibility and confidence (measured by ChIP-seq peak p-value) (Zhang et al., 2013). **E,** At left, the DNase I aggregate cleavage profile of ATERF-1 reflected the known DNA interface topology of ATERF-1 in complex with the GCC-box. The solution structure indicates a kink at the central GC step of the motif, which is consistent with the increased DNase I accessibility within the center of the motif (bottom left, orange arrows) (Allen et al., 1998) (PDB ID 1GCC). At right, DNase I cleavage at 655 footprinted ATERF-1 motifs. **F,** DNase I aggregate cleavage profiles for *A. thaliana* and *H. sapiens* bHLH TFs and *A. thaliana* and *H. sapiens* MADS-box TFs revealed similar DNase I profiles, reflecting the similarities in DNA binding domains shared among orthologous TFs. *A. thaliana* aggregate cleavage profiles were from whole-seedling data, human profiles are from a normal human fibroblast cell line (NHDF_Ad) (Neph et al., 2012c).

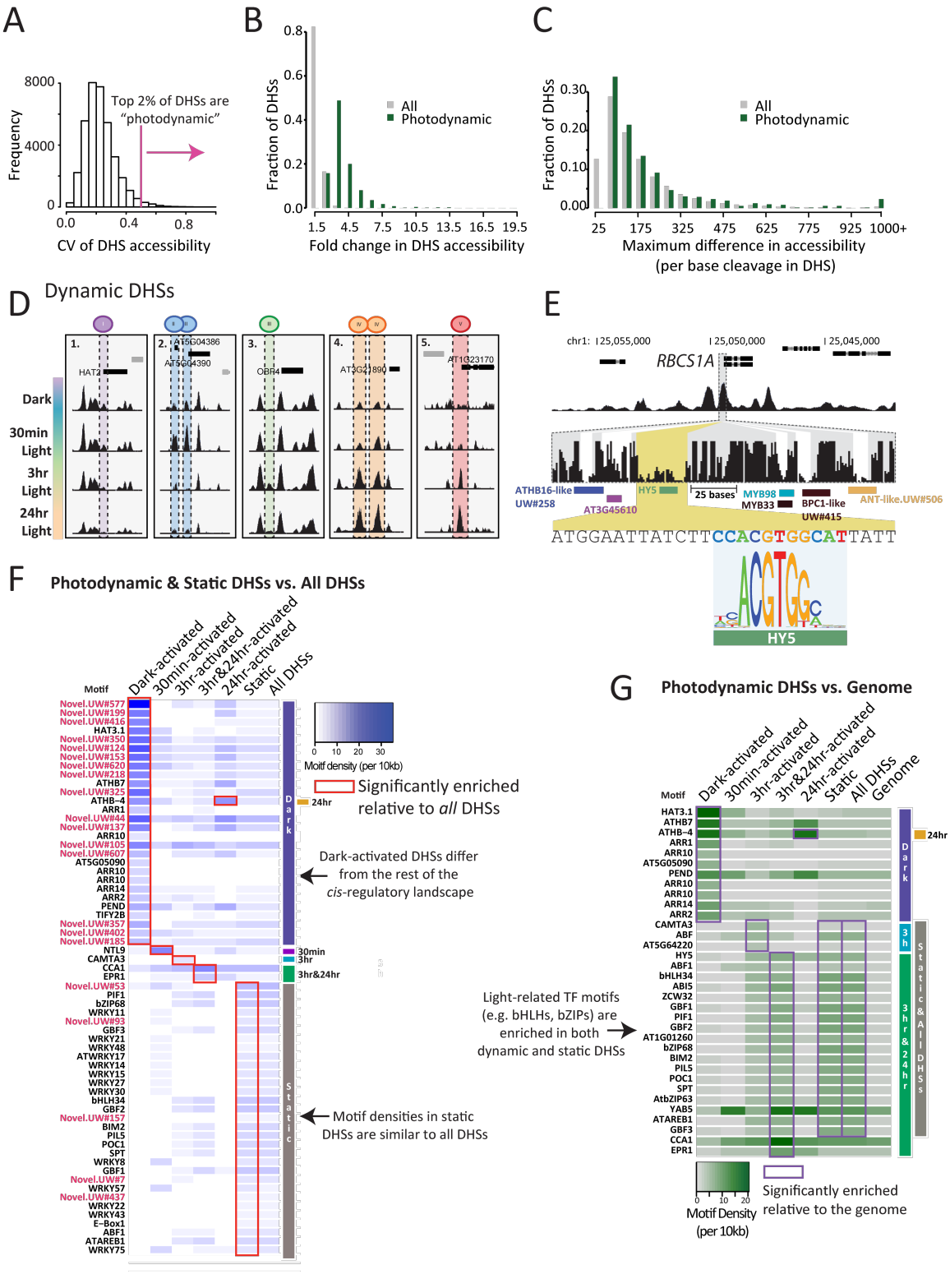

A
Frequency
Top 2% of DHSs are "photodynamic"
CV of DHS accessibility
B
Fraction of DHSs
All
Photodynamic
Fold change in DHS accessibility
C
Fraction of DHSs
All
Photodynamic
Maximum difference in accessibility (per base cleavage in DHS)
D
Dynamic DHSs
Dark
30min Light
3hr Light
24hr Light
HAT2
AT5G04386
AT5G04390
OBP4
AT3G21890
AT1G23170
E
chr1: 25,055,000
25,050,000
25,045,000
RBCS1A
ATHB16-like UW#258
AT3G45610
HY5
25 bases
MYB98
MYB33
BPC1-like UW#415
ANT-like.UW#506
ATGGAATTATCTTCCACGTGGCATTATT
HY5
F
Photodynamic & Static DHSs vs. All DHSs
Motif
Dark-activated
30min-activated
3hr-activated
3hr&24hr-activated
24hr-activated
Static
All DHSs
Motif density (per 10kb)
Significantly enriched relative to all DHSs
Dark-activated DHSs differ from the rest of the cis-regulatory landscape
Motif densities in static DHSs are similar to all DHSs
G
Photodynamic DHSs vs. Genome
Dark-activated
30min-activated
3hr-activated
3hr&24hr-activated
24hr-activated
Static
All DHSs
Genome
Light-related TF motifs (e.g. bHLHs, bZIPs) are enriched in both dynamic and static DHSs
Motif Density (per 10kb)
Significantly enriched relative to the genome

Figure S2. The dynamic chromatin landscape upon photomorphogensis. The DHSs with the top 2% CV (standard deviation / mean) in DNase I accessibility (cleavages) are considered "photodynamic". **B,** Fold change of DNase I accessibility within all (grey) and photodynamic (green) DHSs. **C,** Maximum cleavage differences within all (grey) and photodynamic (green) DHSs. **D,** Representative examples of DHSs (1-5) from each condition-specific DHS cluster were located near novel photomorphogenesis-related genes. Each window is 5kb, vertical ranges vary but are consistent for each DHS example to emphasize fold change. Examples are color-coded to indicate respective condition-specific DHS clusters. **E**, A HY5 binding site (green) upstream of the RBCS1A gene, originally discovered with traditional DNase I footprinting (Chattopadhyay et al., 1998), coincided with a 1% FDR footprint in 7d-old light-grown seedlings. **F**, Motif density (shades of blue) within dark-activated DHSs differed dramatically from light-activated and static DHSs. Motif density within static DHSs is similar to the motif density within all DHSs; though quite a few differences are significant, they are subtle. Motif enrichments (hypergeometric tests w/ Bonferroni correction) are relative to all seedling DHSs and are outlined in red. **G**, Relative to their frequency in the genome, light-related TF recognition sequences (e.g. HY5 and PIF motifs) are enriched in photo-activated DHSs, static DHSs, and all seedling DHSs. Motif enrichments (hypergeometric tests w/ Bonferroni correction) are relative to the genome and are outlined in purple.

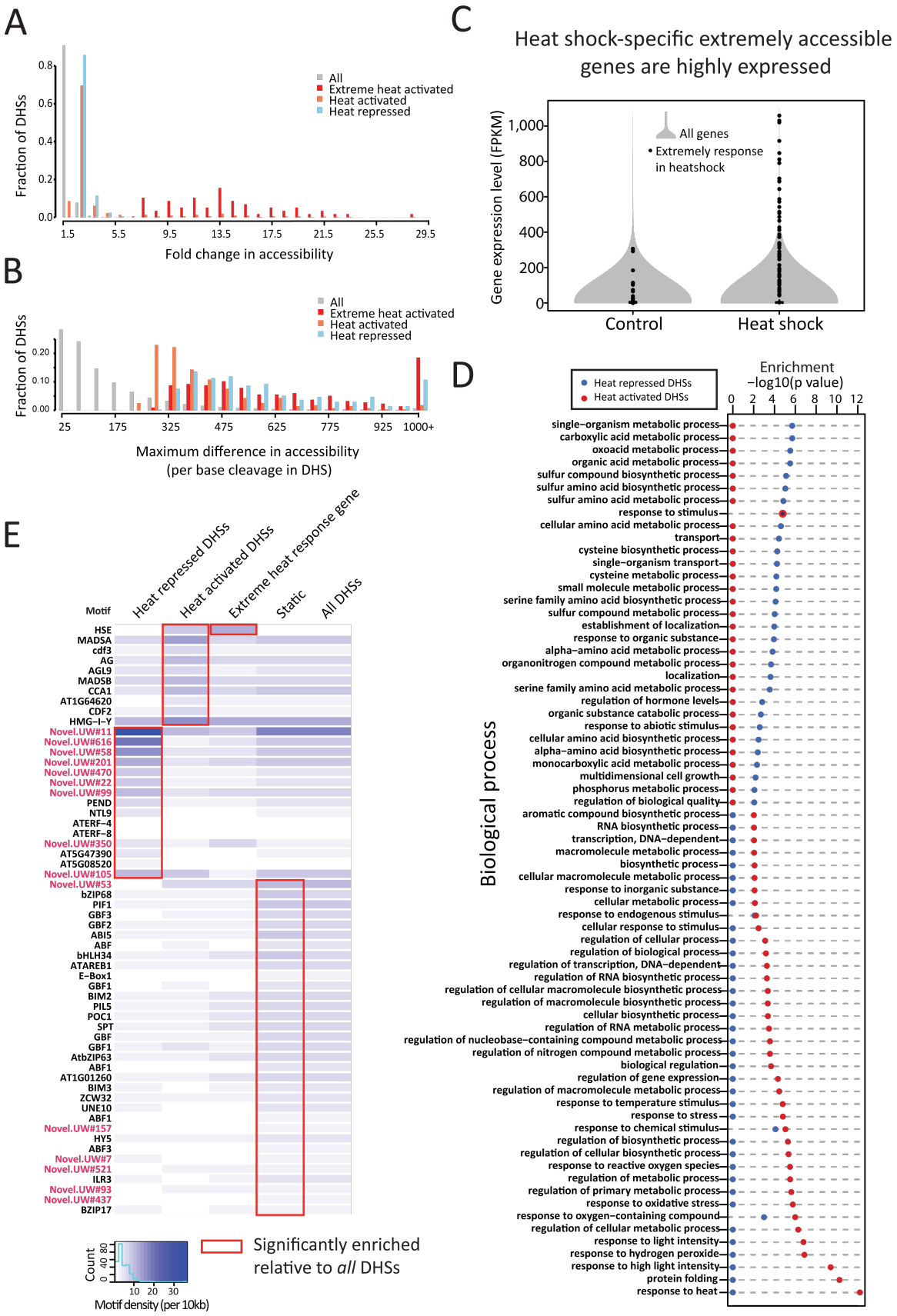

A
Fraction of DHSs
All
Extreme heat activated
Heat activated
Heat repressed
Fold change in accessibility
B
Maximum difference in accessibility
(per base cleavage in DHS)
C
Heat shock-specific extremely accessible genes are highly expressed
Gene expression level (FPKM)
All genes
Extremely response in heatshock
Control
Heat shock
D
Heat repressed DHSs
Heat activated DHSs
Enrichment −log10(p value)
Biological process
single−organism metabolic process
carboxylic acid metabolic process
oxoacid metabolic process
organic acid metabolic process
sulfur compound biosynthetic process
sulfur amino acid biosynthetic process
sulfur amino acid metabolic process
response to stimulus
cellular amino acid metabolic process
transport
cysteine biosynthetic process
single−organism transport
cysteine metabolic process
small molecule metabolic process
serine family amino acid biosynthetic process
sulfur compound metabolic process
establishment of localization
response to organic substance
alpha−amino acid metabolic process
organonitrogen compound metabolic process
localization
serine family amino acid metabolic process
regulation of hormone levels
organic substance catabolic process
response to abiotic stimulus
cellular amino acid biosynthetic process
alpha−amino acid biosynthetic process
monocarboxylic acid metabolic process
multidimensional cell growth
phosphorus metabolic process
regulation of biological quality
aromatic compound biosynthetic process
RNA biosynthetic process
transcription, DNA−dependent
macromolecule metabolic process
biosynthetic process
cellular macromolecule metabolic process
response to inorganic substance
cellular metabolic process
response to endogenous stimulus
cellular response to stimulus
regulation of cellular process
regulation of biological process
regulation of transcription, DNA−dependent
regulation of RNA biosynthetic process
regulation of cellular macromolecule biosynthetic process
regulation of macromolecule biosynthetic process
cellular biosynthetic process
regulation of RNA metabolic process
regulation of nucleobase−containing compound metabolic process
regulation of nitrogen compound metabolic process
biological regulation
regulation of gene expression
regulation of macromolecule metabolic process
response to temperature stimulus
response to stress
response to chemical stimulus
regulation of biosynthetic process
regulation of cellular biosynthetic process
response to reactive oxygen species
regulation of metabolic process
regulation of primary metabolic process
response to oxidative stress
response to oxygen−containing compound
regulation of cellular metabolic process
response to light intensity
response to hydrogen peroxide
response to high light intensity
protein folding
response to heat
E
Motif
Heat repressed DHSs
Heat activated DHSs
Extreme heat response gene
Static
All DHSs
HSE
MADSA
cdf3
AG
AGL9
MADSB
CCA1
AT1G64620
CDF2
HMG−I−Y
Novel.UW#11
Novel.UW#616
Novel.UW#58
Novel.UW#201
Novel.UW#470
Novel.UW#22
Novel.UW#99
PEND
NTL9
ATERF−4
ATERF−8
Novel.UW#350
AT5G47390
AT5G08520
Novel.UW#105
Novel.UW#53
bZIP68
PIF1
GBF3
GBF2
ABI5
ABF
bHLH34
ATAREB1
E−Box1
GBF1
BIM2
PIL5
POC1
SPT
GBF
GBF1
AtbZIP63
ABF1
AT1G01260
BIM3
ZCW32
UNE10
ABF1
Novel.UW#157
HY5
ABF3
Novel.UW#7
Novel.UW#521
ILR3
Novel.UW#93
Novel.UW#437
BZIP17
Count
Motif density (per 10kb)
Significantly enriched relative to all DHSs

Figure S3. The dynamic chromatin landscape upon heat shock. Fold change of DNase I accessibility within all (grey), extreme heat-activated (red), heat-activated (orange) and heat-repressed (blue) DHSs. **B,** Maximum cleavage differences within all (grey), extreme heat-activated (red), heat-activated (orange) and heat-repressed (blue) DHSs. **C,** The extreme-accessibility genes were also highly expressed in response to heat shock. In gray, expression level distribution for all genes; black dots, expression levels of heat shock-signature genes. **D**, GO enrichment analysis (AmiGO; http://amigo.geneontology.org) for genes near heat-activated or heat-repressed DHSs revealed that heat- and oxidative stress-responsive genes were enriched near heat-activated DHSs. In contrast, genes involved in metabolism, biosynthesis, and transport were enriched near heat-repressed DHSs. **C,** Motif density (shades of blue) within heat-repressed DHSs, heat-activated DHSs, extreme response genes (including 500 bp of promoter region), and static DHSs, compared to all DHSs showed characteristic enrichments (hypergeometric tests w/ Bonferroni correction) for known motifs, such as the HSE, as well as several novel *de novo* motifs. Motif enrichments (hypergeometric tests w/ Bonferroni correction) are relative to all seedling DHSs and are outlined in red.

Table S1. Sample overview

Table includes samples used in the analysis grouped as replicates. DHS and footprint information is located at **plantregulome.org** and GEO GSE53322. SPOT (Signal Portion of Tags), is a quality measure for short-read sequence experiments. SPOT is the proportion of tags in a 5 million tag sub-sample that fall in non-1%FDR thresholded hotspots. ChrC and ChrM tags are removed before SPOT calculation. Pearson correlations of cleavage densities between replicates are indicated; in cases with more than two replicates, Pearson correlation values were averaged.

Sample #	Tissue	Conditions	SPOT score	Mapped reads	Mapped reads ChrM	Mapped reads ChrC	DHSs (1%FDR)	Foot-prints	Pears. Corr.
20969	whole seedling	7 day old dark+24hr LD conditions	0.46	30,764,878	292,452	6,006,228			
20968	whole seedling	7 day old dark+24hr LD conditions	0.49	105,908,132	977,439	21,575,778	26,712	113,003	0.998
20960	whole seedling	7 day old dark+30min light	0.37	8,471,763	77,904	1,225,965			
22140	whole seedling	7 day old dark+30min light	0.55	37,293,231	357,525	4,680,869	28,842	128,888	0.926
20963	whole seedling	7 day old dark+3hr light	0.35	8,917,581	70,816	1,021,967			
22465	whole seedling	7 day old dark+3hr light	0.51	9,287,330	65,841	1,504,522			
20412	whole seedling	7 day old dark+3hr light	0.62	98,684,296	690,590	21,280,274	31,336	151,381	0.947
21310	whole seedling	7 day old, Dark	0.35	18,526,333	183,500	2,223,109			
22138	whole seedling	7 day old, Dark	0.45	31,976,750	384,042	7,129,190	28,917	91,891	0.962
21094	whole seedling	7 day old, LD conditions	0.52	119,711,257	706,708	48,025,284	34,637	422,008	
20423	whole seedling	7 day old, LD conditions, 30min 45°C HS	0.50	185,381,712	6,700,558	105,648,521	32,755	433,522	0.998
20424	whole seedling	7 day old, LD conditions, 30min 45°C HS	0.56	9,492,630	324,896	5,407,729			
19992	whole seedling	7 day old, LD, spray control	0.44	195,436,271	1,743,480	65,454,373	34,288	697,899	0.967
19994	whole seedling	7 day old, LD, spray control	0.51	10,321,280	55,050	2,189,559			
19996	whole seedling	7 day old, LD, spray control	0.62	7,343,446	30,718	2,979,610			
20418	whole seedling	7 day old, LD, spray control	0.70	64,475,005	266,165	25,677,199			
21513	root hair cells	7 day old, LD	0.63	10,150,323	33,383	307,810	25,452	46,511	0.920
21512	root hair cells	7 day old, LD	0.64	6,082,493	18,493	151,191			
18564	root hair cells	7 day old, LD	0.62	10,628,240	62,960	676,882			
17498	root-non hair cells	7 day old, LD	0.57	45,891,881	388,666	1,677,353			
21511	root maturation zone	7 day old, LD	0.60	30,976,402	76,307	457,192	29,471	204,220	

Table S2. DHS and footprint distributions over genomic elements in 7-day-old seedling

We determined overlaps of DHSs and footprints with genomic elements to establish if certain elements were enriched or depleted for DHSs or footprints. Column headings include: genomic element type, number of base pairs within genomic element type, percent of the genome contained in genomic element type, number of DHS midpoints contained in genomic element type, number of footprint midpoints contained in genomic element type, the percentage of DHSs in genomic element type, the percentage of footprints in genomic element type, DHS and footprint p-values determined in R using the binom.test() function. TSS was defined as 400 bp upstream of the coding region for TAIR10 annotated genes.

Element	**Genomic Portion (bp)**	**% gen-ome**	**DHSs in element**	**Foot-prints in element**	**% of DHSs**	**% of foot-prints**	**DHS enrichment/ depletion (p-value)**	**Footprint enrichment/ depletion (p-value)**
TSS	11,174,920	9.87	12,647	228,568	37.42	32.75	$<2.2x10^{-16}$	$<2.2x10^{-16}$
5'UTR	2,664,398	2.35	1,703	52,962	5.04	7.59	$<2.2x10^{-16}$	$<2.2x10^{-16}$
Coding	32,250,392	28.50	3,590	95,441	10.62	13.68	$<2.2x10^{-16}$	$<2.2x10^{-16}$
Intron	17,295,174	15.28	1,141	28,598	3.38	4.10	$<2.2x10^{-16}$	$<2.2x10^{-16}$
3'UTR	4,561,920	4.03	1,061	33,642	3.14	4.82	$<2.2x10^{-16}$	$<2.2x10^{-16}$
TE	20,921,849	18.49	1,021	25,081	3.02	3.59	$<2.2x10^{-16}$	$<2.2x10^{-16}$
Intergenic	24,302,578	21.47	12,630	233,607	37.37	33.47	$<2.2x10^{-16}$	$<2.2x10^{-16}$

Table S3. TF Introns containing a DHS

Excel file. TF introns are enriched for DHSs. Column headings include: TF intron chromosome location, intron start and stop, TF gene name (all genes are ".1" gene models), number of DHS midpoints located within the intron, and TF name. TFs were taken from AGRIS TFDB (Davuluri et al., 2003)(http://arabidopsis.med.ohio-state.edu/AtTFDB/).

Table S4. Methylation within DHSs

To determine if DHSs were enriched or depleted for cytosine methylation, we examined methylation frequencies and methylation contexts within DHSs. Column headings include: number and type of methylated cytosines: outside of DHSs, within DHSs, within DHSs that *do not* overlap transposable elements or repeats, within DHSs that *do* overlap transposable elements or repeats, and within all transposable elements and repeats. Methylation data is from a previously published study (Lister et al., 2008), transposons are from TAIR10, repeats are from the RepeatMasker library for Arabidopsis.

Methylation context	**Me-C outside DHSs**	**Me-C within DHSs**	**DHSs not over TEs and repeats**	**DHSs over TEs and repeats**	**TEs and repeats**
Me-CG	1,062,362	3,723	2,913	810	515,294
Me-CHG	361,530	717	490	227	291,981
Me-CHH	264,550	3,251	2,367	884	182,041
Total me-C	1,688,442	7,690	5,770	1,920	989,317
Total C	40,884,885	1,970,359	1,855,813	114,546	9,976,395
me-C/Total C	0.0413	0.0039	0.0031	0.0167	0.0991

Table S5. D*e novo* motif sequences

Excel file. *De novo* motif discovery (Neph et al., 2012c) within footprints yielded 636 motifs. One hundred and twelve motifs were novel. Column headings include: motif consensus, UW motif ID, motif Z-score, and novelty.

Table S6. Cloned TF position weight matrices (PWMs)

Excel file. Sheet 1. Cloned TFs used in PBM assay (n=46). Forty six cloned TF DNA binding domains used in protein binding microarray assays. Column headings include: Backbone, Plasmid ID, Gene_Name, Gene_ID, Pfam ID, Insert DNA, Insert AA, Insert Mw, DNA length, AA length, Tag location, Tag Mw, Main Cloning sites. **Sheet 2. Cloned TF PWMs.** Position weight matrices for 46 cloned TFs. PWMs are also available through the following database: **http://cisbp.ccbr.utoronto.ca/.**

Table S7. SNV enrichment in DHSs for 107 GWAS phenotypes

Excel file. Enrichment of significant GWAS SNVs in DHSs was determined for 107 *A. thaliana* GWAS phenotypes. Column headings include: phenotype name from Atwell *et al.,* 2010, SNV enrichment p-value, phenotype category from Atwell *et al.,* 2010.

Table S8. Photomorphogenesis DHSs and motif enrichments

Excel file. Sheet 1. Dynamic DHSs during photomorphogenesis (n=734). DHSs that changed during photomorphogenesis were identified and mapped to their nearest TSS. Column headings include: DHS type, dynamic DHS genomic coordinates, SPOT-normalized DNase I cut counts, relative difference between conditions, nearest TSS, gene model type, and gene function. **Sheet 2. Superset DHSs used in photomorphogenesis analysis.** Superset DHSs are merged DHS regions created so that DNase I cleavage in hypersensitive sites can be easily compared between samples. Column headings include: chromosome, start, stop, cut count dark, cut count, dark+30min light, dark+3hr light, dark+24hr light. DNase I cut counts in this sheet are non-SPOT-normalized. **Sheet 3**. **TF motif densities and enrichments in photodynamic and static DHSs.** Motif densities were calculated from predicted binding sites (p-value $1e^{-4}$ FIMO) within dynamic photomorphogenesis DHSs and static DHSs. P-values and Bonferroni corrected p-values are from hypergeometric tests comparing motifs in DHS subsets to all DHSs. Column headings include: Motif, Dark[counts], 30min[counts], 3hr[counts], 3hr_24hr[counts], 24hr[counts], static[counts], All DHS[counts], genome[counts], Dark[density per 10kb], 30min[density per 10kb], 3hr[density per 10kb], 3hr_24hr[density per 10kb], 24hr[density per 10kb], static[density per 10kb], All DHS[density per 10kb], genome[density per 10kb], Dark[pval], 30min[pval], 3hr[pval], 3hr_24hr[pval], 24hr[pval], Dark[bonferroni], 30min[bonferroni], 3hr[bonferroni], 3hr_24hr[bonferroni], 24hr[bonferroni], static[bonferroni].

Table S9. GO term enrichments of genes near photodynamic DHSs

GO term enrichments were performed on the genes with TSSs nearest to DHSs that changed during photomorphogenesis. Column headings include: GO term by DHS cluster, GO description, enrichment p-value, sample frequency, TAIR background frequency.

GO term	Description	P-value	Sample freq	Background freq
Cluster V – 24hr				
none				
Cluster IV -3hr & 24hr				
GO:0009628	response to abiotic stimulus	1.65E-05	43/183 (23.5%)	2833/30320 (9.3%)
GO:0010224	response to UV-B	2.32E-05	9/183 (4.9%)	105/30320 (0.3%)
GO:1901362	organic cyclic compound biosynthetic process	5.45E-05	46/183 (25.1%)	3289/30320 (10.8%)
GO:0009411	response to UV	6.21E-05	12/183 (6.6%)	249/30320 (0.8%)
GO:1901700	response to oxygen-containing compound	9.18E-05	38/183 (20.8%)	2460/30320 (8.1%)
GO:0019438	aromatic compound biosynthetic process	0.000394	42/183 (23.0%)	3052/30320 (10.1%)
GO:0042440	pigment metabolic process	0.000395	13/183 (7.1%)	354/30320 (1.2%)
GO:0009416	response to light stimulus	0.000463	24/183 (13.1%)	1208/30320 (4.0%)
GO:0010033	response to organic substance	0.000906	39/183 (21.3%)	2805/30320 (9.3%)
GO:0050896	response to stimulus	0.00106	69/183 (37.7%)	6616/30320 (21.8%)
GO:0044710	single-organism metabolic process	0.00123	60/183 (32.8%)	5428/30320 (17.9%)
GO:0009314	response to radiation	0.00139	24/183 (13.1%)	1285/30320 (4.2%)
GO:0046148	pigment biosynthetic process	0.00141	11/183 (6.0%)	274/30320 (0.9%)
GO:0009813	flavonoid biosynthetic process	0.00146	10/183 (5.5%)	220/30320 (0.7%)
GO:0044763	single-organism cellular process	0.00273	77/183 (42.1%)	7917/30320 (26.1%)
GO:1901576	organic substance biosynthetic process	0.00299	68/183 (37.2%)	6661/30320 (22.0%)
GO:1901360	organic cyclic compound metabolic process	0.00323	56/183 (30.6%)	5057/30320 (16.7%)
GO:0044711	single-organism biosynthetic process	0.00331	41/183 (22.4%)	3190/30320 (10.5%)
GO:0009718	Anthocyanin compound biosynthetic process	0.00399	6/183 (3.3%)	65/30320 (0.2%)
GO:0008152	metabolic process	0.00415	94/183 (51.4%)	10544/30320 (34.8%)
GO:0044237	cellular metabolic process	0.00448	87/183 (47.5%)	9492/30320 (31.3%)
GO:0009812	flavonoid metabolic process	0.00459	10/183 (5.5%)	250/30320 (0.8%)
GO:0009058	biosynthetic process	0.00539	69/183 (37.7%)	6909/30320 (22.8%)
GO:0009987	cellular process	0.00787	102/183 (55.7%)	11946/30320 (39.4%)
GO:0006725	cellular aromatic compound metabolic process	0.00925	53/183 (29.0%)	4835/30320 (15.9%)
GO:0009744	response to sucrose	0.00928	9/183 (4.9%)	213/30320 (0.7%)
Cluster III – 3hr				
none				
Cluster II – 30min				
none				
Cluster I - dark				
GO:0009733	response to auxin	2.26E-12	23/177 (13.0%)	442/30320 (1.5%)
GO:0009725	response to hormone	7.19E-07	33/177 (18.6%)	1662/30320 (5.5%)
GO:0009719	response to endogenous stimulus	4.08E-06	34/177 (19.2%)	1879/30320 (6.2%)
GO:0042221	response to chemical	0.000019	50/177 (28.2%)	3787/30320 (12.5%)
GO:0010033	response to organic substance	3.52E-05	41/177 (23.2%)	2805/30320 (9.3%)
GO:0050896	response to stimulus	0.000479	68/177 (38.4%)	6616/30320 (21.8%)
GO:0009641	shade avoidance	0.00179	4/177 (2.3%)	15/30320 (0.0%)

Table S10. TF network degree changes (input + output edges) during photomorphogensis

Excel file. Input and output edges were summed for each network TF in each photomorphogenesis treatment condition. Column headings include: transcription factor, network degree in the dark, network degree upon 30min light, network degree upon 3hr light, network degree upon 24hr light.

Table S11. Heat shock-activated and heat shock-repressed DHSs and motif enrichments

Excel file. Sheet 1. Heat shock-activated and heat shock-repressed DHSs (n=1980). Heat shock activated and heat shock-repressed DHSs were identified by their relative difference in accessibility between heat shock and control conditions. Column headings include: DHS category, genomic coordinates for dynamic DHS, DNase I cut counts in control, DNase I cut counts in heat shock, relative difference between heat shock and control, nearest TSS, TSS strand, gene model type, gene function. **Sheet 2. Superset DHSs used in heat shock analysis.** Superset DHSs are merged DHS regions created so that DNase I cleavage in hypersensitive sites can be easily compared between samples. Column headings include: chromosome, start, stop, cut count control, cut count heat shock**. Sheet 3. TF motif densities and enrichments in heat-dynamic and static DHSs.** Motif densities were calculated from predicted binding sites (p-value $1e^{-4}$ FIMO) within heat-dynamic DHSs and static DHSs. P-values and Bonferroni corrected p-values are from hypergeometric tests comparing motifs in DHS subsets to all DHSs. Column headings include: Motif, Heat repressed DHS [counts], Heat activated DHS [counts], extreme Response genes [counts], static DHS [counts], All DHS [counts], Heat repressed DHS [density per 10kb], Heat activated DHS [density per 10kb], Extreme Response genes [density per 10kb], static DHS [density per 10kb], All DHS [density per 10kb], Heat repressed DHS [pval], Heat activated DHS [pval], Extreme Response genes [pval], static DHS [pval], Heat repressed DHS [Bonferroni], Heat activated DHS [Bonferroni], Extreme Response genes [Bonferroni], static DHS [Bonferroni]. **Sheet 4. Extreme response genes in heat shock.** Heat shock signature genes were identified by their extreme hypersensitivity to DNase I upon heat shock. Many novel heat shock genes, including many with unknown function, were implicated by this analysis.

Supplemental Methods

Constructs and transgenic plants. Original INTACT lines ADF8:NTF::ACT2:BirA (root hair cell line) and GL2:NTF::ACT2:BirA (root non-hair cell line) were used for cell-type-specific DNase I mapping (Deal and Henikoff, 2010, 2011). To create a transgenic line in which nearly all nuclei could be captured with INTACT, Andrej Arsovski made a new construct, in which the *UBQ10* promoter was fused with the NTF segment used in the original INTACT lines, and placed this construct in the pGR I 0029 vector, replacing GFP with RFP. The UBQ10:NTF construct contains 2.1kb of the *UBQ10* promoter, including the 5'UTR. The *UBQ10* promoter was amplified using the following primers: UBQ10R520-GCTCAACAACAAACTTTCCATTCACCC and UBQ10R2546-CTGTTAATCAGAAAAACTCAGAT. Andrej Arsovski simultaneously transformed the UBQ10:NTF construct and the ACT2:BirA construct by floral dip to obtain a stable UBQ10:NTF::ACT2:BirA line which was then used for all whole seedling samples and root samples. The UBQ10:NTF plasmid is available upon request.

Growth & treatment conditions. *Root cell types.* 0.5ml of seeds were surface sterilized with 50% (v/v) bleach with 0.05% (v/v) Triton X-100 to generate 3-6 g of root tissue; this amount is typically sufficient for a single DNase I prep. Seeds were sown on 1/2XMS,1% sucrose, 0.8% agar plates and grown vertically in long day (LD) conditions (16H light 22°C, 8H dark 20°C) for 7 to 10 days. Root maturation zone or whole roots were collected, rinsed with cold water, blotted dry and weighed before homogenization.

Whole seedlings. For light-grown seedlings, 0.1g of seeds were surface sterilized by treating with 70% EtOH with 0.5% triton for 10 minutes followed by 5 minutes in 95% EtOH. For dark-grown and varying light exposure samples, 0.4g of seeds were used. Seeds were dried

completely on sterile filter paper and plated on 150mm petri plates containing 50ml 1XMS with 0.8% agar covered by a sterile #1 filter circle cut to size (Whatman, GE Healthcare UK ltd). Plates were sealed with micropore tape, double wrapped with aluminum foil and stratified for 3 days at 4C.

Light treatments. For whole seedling controls, stratified plates were unwrapped and moved to LD conditions (16hr light, 22°C; 8hr dark, 20°C) in a growth chamber (Conviron CMP5090, Controlled environment ltd. Winnipeg, Manitoba, Canada) and grown for 7 days. Dark-grown plates were moved from stratification conditions into the same chamber as the control plates and were unwrapped and exposed to light for 1hr to promote germination. Plates were then re-wrapped and grown for 7 days. To test response to increasing light treatments, dark-grown seedlings were unwrapped after 7 days of dark growth and grown in LD conditions for 30mins, 3hrs, and 24hrs respectively. To control for regulatory changes due to circadian regulation, experiments were timed such that all seedlings were harvested at the same time. To prevent light exposure, dark-grown seedlings were homogenized under green light.

Heat shock. Seedlings were grown for 7 days under LD conditions. Plates were moved from control conditions to 45°C for 30mins, immediately prior to harvesting. Control seedlings remained in LD conditions.

Purification of biotin labeled nuclei. *Root cell types.* 2-3 grams of root tissue were homogenized in 1-2 ml of 1XCB (15mM PIPES pH 6.5, 20mM NaCl, 80mM KCl, 1mM EDTA pH 8.0, 0.3M Sucrose, 2.5mM Spermidine, 0.0005% (v/v) 2-mercaptoethanol) by chopping with a razor blade 2 times for 2 minutes. After the first chop, resulting liquid was drawn off

by sweeping debris aside using a 250-350 micron nylon screen and passed through a syringe filter loaded with coarse nylon mesh (70-250 micron). After the last chop, remaining tissue was swirled with enough 1XCB such that tissue moved freely and the liquid drawn off, course filtered and pooled with the first chop. While a single 70 micron filtering step is generally sufficient to remove most of the debris, subsequent filtering steps using a 36 micron and 5 micron filter were used to further reduce debris if necessary. Debris removal was optimized according to the tissue type and weight. Homogenate was distributed to 1.5ml low-retention microfuge tubes and pelleted at 1000 x g for 5 minutes at 4°C. The supernatant was removed and the pellet re-suspended in 1ml 1XCB. Bead binding, incubation, and wash steps were performed according to Deal and Henikoff, 2010 & 2011 (Deal and Henikoff, 2010, 2011), however, 1XCB or 1XCB supplemented with 0.1% Triton X-100 were used in lieu of NPB solutions. Isolated nuclei were gently re-suspended in 200ul 1XCB for digestion with DNase I.

Whole seedlings. Seedling tissue was homogenized in 6-8ml of 1XCB (15mM PIPES pH 6.5, 20mM NaCl, 80mM KCl, 1mM EDTA pH 8.0, 0.3M Sucrose, 2.5mM Spermidine, 0.0005% (v/v) 2-mercaptoethanol) by chopping with a razor blade 3 times for 2 minutes. After each chop, the debris was separated using a 250 micron nylon mesh; the remaining liquid collected with a 10ml syringe, was passed again through a 250 micron nylon mesh. The combined filtrate from 3 chops was then sequentially filtered through 36 micron and 5 micron nylon meshes, respectively. The final triple-filtered homogenate was distributed to 1.5ml low-retention microfuge tubes and pelleted at 1000 x g for 5 minutes at 4°C. The supernatant was discarded and the pellets re-suspended and combined in 2ml 1XCB. Fifty microliters of Dynabeads M280 Streptavidin beads (Invitrogen) were added per 1ml of re-

suspended pellet and incubated on a rotator for 30 min at 4°C. Wash steps were performed according to Deal and Henikoff, 2010 & 2011 however, 1XCB or 1XCB supplemented with 0.1% Triton X-100 were used in lieu of NPB solutions. Captured nuclei were resuspended in 200ul 1XCB for digestion with DNase I.

DNase I digestion and sequencing. Samples were digested as described previously by Sabo et al., 2006 (Sabo et al., 2006) with slight modifications: *A. thaliana* samples were treated with 45u DNase I for 3 minutes at 25°C. Size fractionation and sequencing of double-cut DNA fragments was done as described previously (Hesselberth et al., 2009; Neph et al., 2012c). Short-read archives can be found in GEO accession GSE53322. See **http://www.plantregulome.org/protocols** for additional details.

RNA extraction. *A. thaliana* UBQ10::NTF seedling were grown as for nuclei extraction, including all light treatments. 100-200mg of tissue was ground in liquid nitrogen and total RNA extracted using the Spectrum Plant Total RNA Kit (Sigma), total RNA was treated with DNase I on columns (Qiagen). Ribosomal subtraction was done using the Ribo-Zero Plant Leaf kit (Epicentre). Library generation was done using the TruSeq Sample preparation kit v2 (Illumina). Sequencing was completed on the Illumina HiSeq platform.

Mapping DNase I hypersensitivity. Uniquely-mapping high-throughput sequencing reads (36-bp) were mapped to the TAIR9 genome. The 5' ends of sequence reads were used to calculate per-base-pair DNase I cleavage. DNase I sensitive regions, or hotspots, were identified as in John et al. 2011 (John et al., 2011). In brief, the Hotspot algorithm looks for an enrichment of tags along the genome by evaluating the concentration of tags within a small window (200-300 bp) relative to the local background (observed tags in 50 kb

surrounding window) based on the binomial distribution model. Each mapped tag was given a z-score relative to the surrounding small and background windows centered on the tag. A 'hotspot' was defined as a succession of neighboring tags within a 250 bp window, each of whose z-score was greater than two. Once the hotspot was identified, the hotspot was given a z-score relative to the small and background windows centered on the average position of the tags within the hotspot. Z-score calculations were performed on hotspots as in John et al., 2011. SPOT (Signal Portion of Tags) is a quality measure for short-read sequence experiments. SPOT is the proportion of tags in a 5-million-tag sub-sample that fall in non-1%FDR thresholded hotspots. ChrC and ChrM tags are removed before SPOT calculation. A region on chromosome 2 known for potential sequencing errors (Stupar et al., 2001) exhibited unusually high DNase I signal in the heat shock sample and was subsequently masked from the heat shock analyses. Mitochondrial and chloroplast DHSs were not considered. Read depth was normalized for all sample comparisons by subsampling each sample to the same read depth.

Identification of DNase I hypersensitive sites (DHSs). DHSs (150 bp peaks) were determined with modification from John et al. 2011. For Arabidopsis, the local maxima threshold used for Phase-I peaks was dropped to the ninety-fifth percentile from the ninety-ninth percentile. One percent FDR peaks were constructed from FDR-thresholded hotspots.

Footprints. Footprints were computed as previously described (Neph et al., 2012c).

DHS & footprint distribution across genomic features. DHSs or footprints overlapped a genomic feature if their middle base pair fell within that genomic feature. The transcription

start site (TSS) category of genomic features was defined as 400 bp upstream of the most 5' annotated TAIR10 transcription start site. Other feature categories were taken from TAIR10 annotations. Binomial distribution tests (binom.test() in R) were used on each genomic feature category to determine the probability of the observed DHS overlap, given the frequency of base pairs found within that genomic feature across the mappable genome.

DHSs within introns. A binomial distribution test (binom.test() in R) was used to determine whether introns within certain classes of genes were more likely to contain a DHS (7 day old seedling). The probability of observing 101 or more DHSs within transcription factor introns was less than 2.2×10^{-16}, given the probability of that a given intron base pair lies in a TF is 0.018. To simplify complex gene models, intron information from all TAIR10 protein-coding gene models ending in ".1" were used in the analysis. DHS midpoints were used to calculate the DHS-intron overlaps.

DNA methylation within DHSs. A binomial distribution test was used to test whether the probability of observing 7,690 or fewer methylated cytosines (Lister et al., 2008) out of 1,970,359 total cytosines in DHSs was less than 1.0×10^{-50}, given that the probability of cytosine methylation in the rest of the genome is 0.0413. We used a Chi-square test to determine whether ratios of cytosine methylation contexts (CG:CHG:CHH) within DHSs deviated significantly from ratios in regions outside of DHSs ($P < 2.2 \times 10^{-16}$). Methylation data are from a previously published study (Lister et al., 2008), transposons are from TAIR10, repeats are from the RepeatMasker library for *A. thaliana*.

DHS overlap with PIF3 ChIP-seq study. Dark-grown 7d-old seedling DHS and footprint data, randomly subsampled to 24 million sequence reads, was used in this analysis. To determine the ChIP-seq peak overlap with DHSs, PIF3 ChIP-seq peaks were separated into reproducible and non-reproducible categories using the criteria in Zhang and co-authors (Zhang et al., 2013). Reproducible ChIP-seq peaks were then merged in BEDOPS (Neph et al., 2012a). ChIP-seq peaks overlapping a DHS by 75 or more base pairs were counted as overlapping a DHS.

To determine PIF3 motif enrichments in ChIP-seq peaks overlapping DHSs and footprints, BEDOPS (Neph et al., 2012a) was used to count the number of FIMO (Bailey et al., 2009) p-value $1e^{-4}$ PIF3 motifs (protein binding microarray motif) overlapping different genomic contexts by 4 or more base pairs (the PIF3 motif is 8-bp long). The different genomic contexts examined were: genome-wide, within non-reproducible ChIP-seq peaks, reproducible ChIP-seq peaks, DHSs, footprints, the intersection of DHSs and non-reproducible or reproducible ChIP-seq peaks, the intersection of footprints and non-reproducible or reproducible ChIP-seq peaks, and the intersection of DHSs, footprints and reproducible ChIP-seq peaks.

To determine the PIF3 footprint occupancy in ChIP-seq peaks and DHSs, BEDOPS was used to count the number of FIMO p-value 1e-4 PIF3 motifs (protein binding microarray motif) within different genomic contexts: genome-wide, within DHSs, within 1, 2, 3, or 4 ChIP-seq peaks respectively, or within 4 ChIP-seq replicates and a DHS. BEDOPS was then used to count the number of PIF3 motifs in footprints. PIF3 motifs covered 50% or more by a 1% FDR footprint were considered to be in a footprint. PIF3 motifs within any ChIP-seq peak

were stratified by the minimum ChIP-seq p-value. The counts of total PIF3 motifs and footprinted PIF3 motifs were calculated for each p-value quartile.

***De novo* motif discovery and overlap with known motifs.** *De novo* motif discovery was performed using previously described methods (Neph et al., 2012c). In brief, 636 *de novo* motifs were discovered in *A. thaliana* by clustering sequences found within footprints from a deeply sequenced 7-day-old seedling sample. We validated the *de novo* motifs by comparing them to 382 protein binding microarray-derived motif models ((Weirauch, (in press, September 11, 2014)) (n=228), new protein binding microarray (PBM) motifs see **Table S6** (n=46), recently published PBMs (Franco-Zorrilla et al., 2014) (n=108)) using TOMTOM with default parameters and a min-overlap of 5. With these settings, 366 of the 382 models matched at least one *de novo* motif. A set of 112 *de novo* motifs (18%) did not match any protein binding microarray motif or motifs found in TRANSFAC (v2011) (Matys et al., 2006) and JASPAR (Bryne et al., 2007) and were declared "novel" motifs.

Protein Binding Microarray (PBM) analysis of TF binding specificities. We determined the DNA binding specificities for a diverse panel of 46 previously uncharacterized TFs using universal PBMs (Berger et al., 2006)(**Table S6**). We scanned TF protein sequences for putative DNA-binding domains (DBDs) using the 81 Pfam (Finn et al., 2010) models listed in (Weirauch and Hughes, 2011) and the HMMER tool (Eddy, 2009). We designed primers to clone the region encompassing all DBDs plus the 50 flanking endogenous AAs on either side (or until the termini of the protein) by conventional PCR methods into one of a panel of T7-GST vectors for expression in *E. coli* (referred to hereafter as "plasmid constructs"). PBM laboratory methods were identical to those

described in (Lam et al., 2011; Weirauch et al., 2013). Each plasmid was analyzed in duplicate on two different arrays with differing probe sequences (denoted 'ME' and 'HK'). Calculation of 8-mer Z- and E-scores was performed as previously described (Berger et al., 2006). To obtain a single representative motif for each protein, we used a procedure similar to a recent study in which motifs were generated for each array using four different algorithms, and the best-performing single motif was chosen on cross-replicate array evaluations (Weirauch et al., 2013). PWMs are available through the following database: **http://cisbp.ccbr.utoronto.ca/.**

Nucleotide diversity of known and novel motifs. We calculated nucleotide diversity as: $\pi = \frac{n}{n-1}\left(\sum_{i=1}^{S} 2p_i(1-p_i)\right)$, where n is the number of chromosomes and p_i is the frequency of the major allele for the ith segregating site, S. To obtain a per nucleotide estimate of π, we divided by the total number of bases considered for a particular analysis. Sequencing information from 80 *A. thaliana* accessions (Cao et al., 2011a) was used to calculate π, requiring at least 80% of the accessions to be called at the sites used in the analysis. Estimates of π for neutrally evolving DNA (all non-annotated bp in TAIR10 GFF file) and DNA under purifying selection (all protein coding sequences) were used as references. The same calculations were made for all motifs and motifs occurring within 1% FDR footprints. "Novel" motifs (112) are *de novo* motifs that did not match any existing motif model. "Known" motifs are *de novo* motifs that matched at least one protein binding microarray motif or a motif in TRANSFAC (v2011) (Matys et al., 2006) or JASPAR (Bryne et al., 2007).

Crystal structures. A PyMOL (MacPyMOL v1.3) figure was created for *A. thaliana* factor *ATERF-1_M0197_0.63*, using corresponding PDB (Bernstein et al., 1977) record *1GCC* (Allen

et al., 1998). Cut count data from 7-day-old seedlings, aggregated over the factor's 1% FDR hotspot-filtered FIMO p-value 1e-4 hits is indicated on the DNA strand, dark blue indicates high cleavage whereas grey or white indicates protected regions.

GWAS. For each GWAS phenotype (Atwell et al., 2010b), all SNVs or SNVs occurring only within DHSs were binned by p-value to create a distribution for each SNV grouping. A representative curve was then fit to each distribution for graphical clarity. For each GWAS phenotype, we used a non-parametric, one-sample Kolmogorov-Smirnov test to determine if low p-value SNVs were more likely to occur in DHSs. The Kolmogrorov-Smirnov test determined if the cumulative distribution function of SNVs within DHSs binned by p-value was significantly less (i.e. shifted to the left) than the reference. In this case, the reference was the distribution of all SNVs binned by p-value. For the flowering-time example shown in Figure 1E, the Kolmogorov-Smirnov test p-value was 0.00153. Of the 107 phenotypes analyzed, 72 displayed a significant enrichment of associated SNVs in DHSs **(Table S7)**. Some of the GWAS phenotypes analyzed included different numbers of accessions; this may explain some of the variance in p-value distributions and may explain why some SNVs for some phenotypes were not significantly enriched in DHSs (i.e., phenotypes with fewer accessions may have fewer low p-value GWAS SNVs).

TF trinucleotide preferences. TF trinucleotide preferences were determined by calculating trinucleotide frequencies within footprinted and non-footprinted genomic regions. Trinucleotide frequencies within footprinted and non-footprinted regions were tabulated for coding and non-coding portions of the human and *A. thaliana* genomes. Codons and non-coding trinucleotides overlapping footprints were determined using

BEDOPS --element-of operations (Neph et al., 2012a); partial overlaps were excluded. Codons were determined from coding sequences. Non-coding trinucleotide frequencies were obtained from non-coding genomic regions that are uniquely mappable by 36-mer sequencing tags. In human, RepeatMasker annotations were also excluded from non-coding trinucleotide frequency calculations.

Codon usage bias in human and *A. thaliana*. Codon usage in human was determined from consensus CDS gene annotations (Pruitt et al., 2009) downloaded from the UCSC genome browser, corresponding to human build GRCh37/hg19. Codon usage in *A. thaliana* was determined from coding sequences listed in the TAIR10_GFF3_genes.gff file from TAIR. Codons containing an 'N' in the reference were excluded. Individual codon locations were parsed into BED format, excluding start codons and any codons overlapping a splice site or that were ambiguous due to overlapping annotations in different reading frames. Coding annotations containing one or more internal stops in the reference sequence were also excluded.

Identifying dynamic DHSs. *Light series*. DHSs identified in dark, dark+30min light, dark+3hr light and dark+24hr light treatments of 7-day-old seedling samples (subsampled to 24 million reads) were merged to create a DHS superset **(Table S8)**. This DHS superset contains genomic regions that are hypersensitive in at least one of the light series conditions. DNase I cleavages were tallied within each region in the superset DHS for each condition. The cleavage tally was divided by the SPOT score to normalize for sample quality. The DHSs with the most diversity in normalized DNase I cleavages among the experimental conditions (those with a coefficient of variance (sd/mean) in the top two

percentile, a total of 734 regions) were clustered and plotted using the heatmap() function in R using Euclidean distances. Dynamic DHSs were assigned to the nearest TSS. A 400kb region in the centromere of chromosome 2 [3,237,500 to 3,630,000] was highly variable among samples, possibly due to issues with unique mapping. Hence, we excluded this genomic region from further analyses.

Heat shock. DHSs identified in Control and/or Heat shock 7-day-seedling samples (subsampled to 70 million reads) were merged to create a DHS superset **(Table S11)** similar to the light series samples. DNase I cut counts were not normalized for SPOT score because the samples had similar SPOT scores to begin with. Heat shock-responsive DHSs were identified by the relative difference in DNase I cleavages ((heatshock – control) / mean (heat shock, control)) between control and heat shock superset DHSs. The top and bottom 2.5% of DHSs demarcate heat-repressed and heat-activated DHSs. Dynamic DHSs were assigned to the nearest TSS. Extremely accessible, heat-induced genes were hand-curated.

Motif density word clouds. Motifs which were most variable (top 50% coefficient of variation in density across clusters) and enriched three-fold or more relative to the genome are represented as a word cloud. Motif font size reflects the fold increase in density over background genomic levels.

Motif Enrichment. Motif enrichments were performed by calculating the FIMO p-value $1x10^{-4}$ motif counts for the 636 *de novo* motifs, TRANSFAC motifs, JASPAR motifs, and the motifs from protein binding microarray experiments within a subset of DHSs (e.g. heat-

activated DHSs) and superset DHSs (e.g. all regulatory DNA) or the genome. Hypergeometric distribution tests were used to test if motif frequencies in the superset DHSs or the genome (i.e. the background) were different from motif frequencies in a subset of DHSs (e.g. heat-activated DHSs). The following data were used as arguments in R phyper (): q = number of motif X in DHS subset minus 1; m = number of motif X in the DHS superset (or genome); n = the number of all non-X motifs in the DHS superset; k = total number of motifs in DHS subset; lower.tail = FALSE. Bonferroni correction was used to determine significance. For motif counts, density, and enrichment p-values see **Tables S8 (photomorphogenesis) & S11 (heat shock)**.

GO term enrichment. GO enrichments (AmiGO; http://amigo.geneontology.org) were performed on the genes whose TSS were nearest to the dynamic DHSs from each category.

Networks. Edges in the network connect TF genes to the TFs that regulate them. Edge identification was accomplished by scanning TF regulatory regions for occupied TF motifs (footprints). Methods used were as in Neph and co-authors, 2012 (Neph et al., 2012b), except the TF gene region scanned for TF motifs within footprints included 500 bp upstream of the TSS and the entire gene model (longest gene models were used). Only motifs with experimental support were used to build networks. If multiple motifs existed for a given TF (e.g. JASPAR and TRANSFAC report two slightly different motif models for the same TF), the motif with the most substantial experimental support was selected (if such distinction could not be made, the motif was randomly chosen). Different TFs with highly similar motif models (e.g. GBF1 and GBF2) were included in the analysis, because

such TFs have the potential to bind the same motif. The total number of motifs used was 251. Potential TF binding sites were determined by scanning the TAIR9/10 genome for motifs using FIMO (Bailey et al., 2009), version 4.6.1, with a maximum p-value threshold of 10^{-4} and defaults for other parameters. Auto-regulatory loops were established when a target TF's motif resided in footprint(s) within its own regulatory region.

Network motif topology. Network topology was determined using previously described methods (Neph et al., 2012b). Auto-regulatory edges were removed from the *A. thaliana* 7-day-old seedling network before calculating motif frequencies using mfinder software (Milo et al., 2004). A z-score was calculated for each of 13 three-node network motifs, using randomized networks to generate a null distribution for each motif. Z-scores were used to generate a triad significance profile for *A. thaliana* **(Figure 4B),** which was comparable to triad significance profiles described previously (Milo et al., 2004; Neph et al., 2012b).

Light activated subnetworks. Upon light exposure**,** light-related transcription factors increased in connectivity. Light related factors were chosen from the literature or because they were associated with a light-related GO term. Regulatory connections observed in any of the light conditions but not in the dark are in red (**Figure 4E**); regulatory connections observed in the dark and not in any light condition are in blue (**Figure 4E**). We found 108 regulatory edges among 35 TFs with light-related GO annotations, which was significantly more edges than observed with randomly selected TFs from the same network ($p=0.006$). We simulated the random selection of 35 TFs' edges from the network without replacement 1000 times to test the significance of finding 108 regulatory edges among 35 TFs.

Network profile plots. For representative TFs, connectivity was dissected into in-degree (regulators, top), out-degree (targets, bottom), bi-directional loops (condition-specific, left; common across conditions, right), and auto-regulatory loops, revealing regulatory differences despite similarities in overall interaction degree (e.g. ABI5 and HY5) and demonstrating that differences in degree can reflect complex changes in regulation (**Figure 5A-B**).

First-degree TF neighborhoods. First-degree neighborhoods of the selected TFs visualize how a TF's connectivity percolated through the larger network. Using Cytoscape (Shannon et al., 2003) all first-degree nodes and adjacent edges (e.g. must share 1 edge with TF of interest) were selected and all other nodes were removed, leaving behind only first-degree neighbor nodes and edges, and edges among first-degree neighbors. In each circular network, TFs were sorted by number of regulatory connections in the network as a whole, such that connectivity increases clock-wise around the circle starting from the bottom.

DNase I-accessibility gene outliers. DNase I-accessibility gene outliers were identified in all tested conditions by calculating z-scores from DNase I cut counts overlapping TAIR10 genes, excluding genes in centromeric regions from (Clark et al., 2007) converted to TAIR9 coordinates (ch1:13698788-15897560; chr2: 2450003-5500000; chr3:11298763-14289014; chr4:1800002-5150000, chr5:10999996-13332770). Z-scores were calculated by subtracting the mean DNase I cleavages over all genes from the DNase I cleavages of the gene of interest, and dividing that number by the standard deviation. Nine genes with a Z-score above 12 (dotted line) were specific to the heat shock condition (red dots)(**Figure**

6D). Six genes were more accessible than any other similarly sized genes in any condition tested (**Figure 6D**).

Differential HSF regulation during heat shock. HSF-centric regulatory networks were constructed by scanning HSF gene regulatory regions for TF motifs within footprints (similar to other regulatory network construction). Because individual HSF motifs are not known, all the edges connecting the different HSFs (orange circles) to other TFs (grey circles) are input relationships without output relationships.

HSF feedback loops. The generic HSF motif (Megraw and Hatzigeorgiou, 2010) was used to represent any of the 21 possible HSFs when calculating HSF output edges. The non-HSF TFs in feedback loops are TFs that had an occupied HSF motif in their regulatory region, and also targeted an HSF gene.

Appendix C. Molecular mechanisms of robustness in plants [5]

Abstract

Robustness, the ability of organisms to buffer phenotypes against perturbations, has drawn renewed interest among developmental biologists and geneticists. A growing body of research supports an important role of robustness in the genotype to phenotype translation, with far-reaching implications for evolutionary processes and disease susceptibility. Like for animals and fungi, plant robustness is a function of genetic network architecture. Most perturbations are buffered; however, perturbation of network hubs destabilizes many traits. Here, we review recent advances in identifying molecular robustness mechanisms in plants that have been enabled by a combination of classical genetics and population genetics with genome-scale data.

Introduction

Phenotypic robustness is a measure of an organism's ability to buffer phenotype against genetic and environmental perturbations during development (Debat and David, 2001b; Waddington, 1942a; Whitacre, 2012) (Box 1). Robustness is commonly attributed to features of the underlying genetic networks, such as connectivity, redundancy, feedback, and oscillators, as well as to non-genetic mechanisms (Casanueva et al., 2012; Masel and Siegal, 2009; Whitacre, 2012). Targeted perturbation of these features decreases

[5] This chapter was published in Current Opinion in Plant Biology in February 2013
Molecular mechanisms of robustness in plants
Janne Lempe[1, 2], Jennifer Lachowiec[1, 3], Alessandra. M. Sullivan[1], Christine Queitsch[1]

phenotypic robustness and releases cryptic genetic or epigenetic variation. The release of accumulated variation has been invoked as an important factor in evolutionary processes (Masel, 2006) and in disease susceptibility in humans (Gibson, 2009).

Robustness is a quantitative trait. Traditionally, robustness of individuals has been measured as the degree of symmetry in morphological features (Clarke, 1998b). Another robustness measure is the degree of accuracy with which a genotype produces a phenotype across many isogenic siblings. Robustness thus measured is trait-specific and may not be predictive of robustness in other traits (Clarke, 1998b). Like any quantitative trait, robustness shows a distribution among genetically divergent individuals of a species and can be mapped to distinct genetic loci (Hall et al., 2007; Jimenez-Gomez et al., 2011; Sangster et al., 2008b). Non-genetic mechanisms also affect robustness, as mutation penetrance can vary among isogenic individuals (Casanueva et al., 2012; Masel and Siegal, 2009; Whitacre, 2012). Plants are excellent models to probe the molecular underpinnings of robustness. Due to their sessile life-style and continuous development, plants have likely optimized molecular mechanisms that buffer phenotype in the face of ever-changing environmental conditions. Here, we review some advances in identifying molecular mechanisms that contribute to robustness in plants and discuss future directions and challenges.

'Master regulators of robustness' affect connectivity of genetic networks.

One of the best characterized 'master regulators of robustness' is the molecular chaperone HSP90 (Burga et al., 2011; Casanueva et al., 2012; Gangaraju et al., 2011; Jarosz and Lindquist, 2010; Queitsch et al., 2002; Rutherford and Lindquist, 1998; Sangster et al.,

2008a; Sangster et al., 2008b; Sollars et al., 2003; Specchia et al., 2010; Yeyati et al., 2007) (Box 1). HSP90 assists the folding of key developmental proteins, a function that is of even greater importance under stresses that compromise protein folding (Taipale et al., 2010). HSP90 inhibition decreases robustness in plants, flies, yeast, and fish and releases previously cryptic genetic and epigenetic variation (Jarosz and Lindquist, 2010; Queitsch et al., 2002; Rutherford and Lindquist, 1998; Sangster et al., 2008a; Sangster et al., 2008b; Yeyati et al., 2007) (**Figure 1a, b, Figure 2a, b**). In worms, low HSP90 levels correlate with high mutation penetrance (Casanueva et al., 2012). HSP90's capacity to buffer many developmental phenotypes has been attributed to its high connectivity in genetic networks (Sangster et al., 2004). Perturbing HSP90 function impairs its numerous substrates, which is thought to reduce network connectivity and lead to decreased robustness and release of variation. In genetically divergent *A. thaliana* strains, every tested quantitative trait is affected by at least one HSP90-dependent polymorphism; most traits are affected by several (Sangster et al., 2008a; Sangster et al., 2008b).

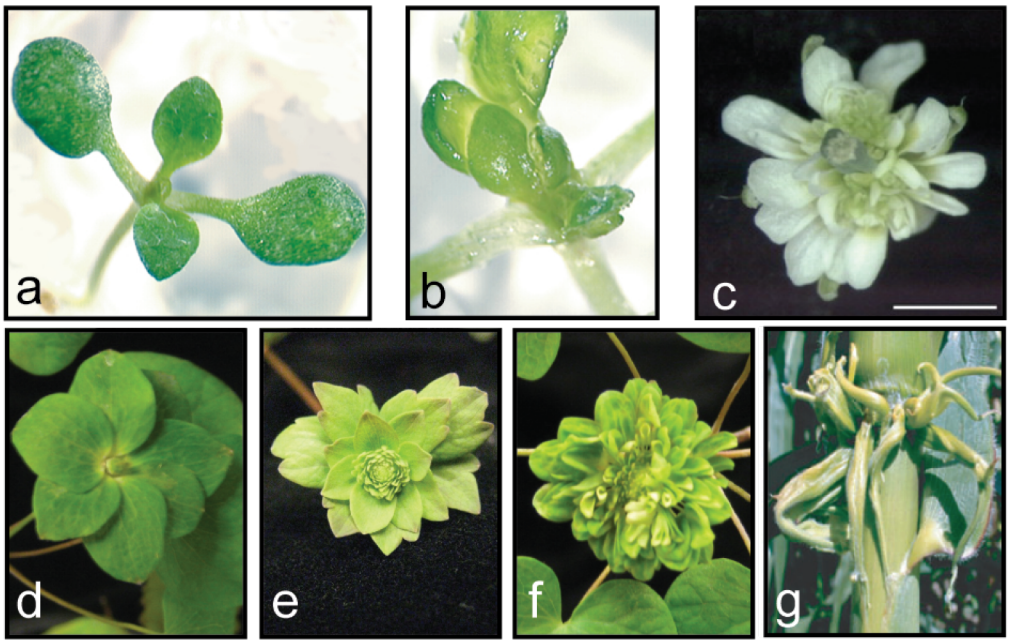


Figure 1. Loss of robustness in developmental traits. a, A wild-type *A. thaliana* seedling shows normal phyllotaxy and leaf shape. **b**, An HSP90-reduced *A. thaliana* seedling shows distorted phyllotaxy and meristem defects. **c**, *A . thaliana* transgenic plant with *miR172*-resistant *AP2* (*rAP2*) under the control of the AP3 promoter (B-gene, shows disrupted floral patterning). Transgenic plants mis-expressing wild-type *AP2* have mostly normal flowers, whereas those misexpressing *rAP2* often have petaloid stamens or as pictured here, complete conversion of stamens into petals in severe cases. Reproduced/adapted with permission (Wollmann et al., 2010). **d-f**, A homeotic flower *Thalictrum thalictroides* mutant with disrupted floral patterning (no carpels, little or no stamens,. Mutant expressivity differs for different floral meristems within the same plant. Three flowers of the same plant are shown (unpublished data. C. Snelson, V.S. diStilio). **g**, A *Zea mays* RNA polymerase IV mutant shows abnormal lateral outgrowths. Used with permission from (Hollick, 2010) and reproduced here from the article "Molecular mechanisms of robustness in plants" by Janne Lempe, Jennifer Lachowiec, Alessandra. M. Sullivan, Christine Queitsch, COPB, February 2013.

The circadian regulator *ELF4* is another gene that reduces robustness when perturbed (Doyle et al., 2002). Circadian clocks are endogenous oscillators with remarkably robust periods, which persist in the absence of light cues and under increased temperature (Mas and Yanovsky, 2009). The robustness of plant clocks is thought to arise from multiple interconnected feedback loops (Mas and Yanovsky, 2009). In reporter

assays, *elf4* mutants show highly variable periods before turning arrhythmic (without periods) (Doyle et al., 2002). It is unclear whether the initial, variable periods translate into increased variation of developmental traits or released cryptic variation; both seem likely given the importance of the circadian clock in orchestrating growth and development. In fact, HSP90's effect on robustness may arise in part from disrupted clock function: ZTL, a circadian regulator, is chaperoned by HSP90 (Kim et al., 2011).

Of course, HSP90 and ELF4 are not the only robustness master regulators in plants. However, unlike in yeast, in which a systematic mutant analysis identified 300 robustness master regulators, all highly connected 'network hubs' (Levy and Siegal, 2008b), in plants a similar analysis has not been conducted; the sheer number of genes and the lack of high-throughput robustness assays have so far made such analysis unfeasible. In our hands, most tested plant mutants, some of them affecting key developmental genes, do not affect robustness of quantitative seedling traits.

Fine-tuning of gene expression stabilizes developmental traits.

The origins and consequences of gene expression noise have been extensively studied in single celled organisms (Eldar and Elowitz, 2010; Munsky et al., 2012; Raser and O'Shea, 2005), but less so in multicellular organisms (Raj et al., 2010), including plants (Forde, 2009). In 2006, Hornstein and Shomron (Hornstein and Shomron, 2006) hypothesized that microRNAs (miRNAs) may reduce gene expression noise and sharpen developmental transitions. In particular, feed-forward loops, in which a transcription factor regulates both a target and its miRNA with opposing effects on target protein levels, were predicted to buffer stochastic expression fluctuations (Hornstein and Shomron, 2006). As

plant miRNAs tend to target key transcription factor and F-box genes, they modulate developmental transitions, variation in leaf morphogenesis, reproductive development, and root architecture (Rubio-Somoza and Weigel, 2011). miRNAs have recently been shown to facilitate robustness. For example, *miRNA164* miRNAs control plant development by dampening transcript accumulation of their targets *CUC1* and *CUC2*, wherever expression of miRNAs and targets overlap. *miRNA164* miRNAs define boundaries for target mRNA accumulation in addition to reducing target expression levels (Sieber et al., 2007).

In plants, small RNA-dependent regulation of gene expression is not limited to miRNAs – in fact, there are many plant-specific small interfering RNAs (siRNAs), some of which are mobile and facilitate robust pattern formation. Chitwood and co-authors (Chitwood et al., 2009) demonstrated that a subset of trans-acting siRNAs (tasiRNAs), the low-abundant and conserved tasiR-ARFs, move intercellularly from the upper leaf side (adaxial), where they originate, to the lower leaf side (abaxial), generating a small RNA gradient that defines the expression boundaries of the abaxial determinant ARF3. tasiR-ARF biogenesis requires both miRNA activity (*miR390*) and siRNA pathway components, including the specialized Argonaute AGO7. Although *miR390* accumulates in a seemingly non-specific pattern throughout the developing leaf, tasiR-ARF biogenesis is restricted to the most adaxial leaf cell layers by the localized expression of *AGO7* (Chitwood et al., 2009). Consistent with the notion that the tasiR-ARF gradient mediates robust adaxial-abaxial fate decision, *ago7* mutants show significantly increased variance in adaxial leaf width (**Figure 2d-g**).

Robust flower development through combinatorial gene interaction

In core eudicots, flower organs – sepals, petals, stamens and carpels – are organized in four concentric whorls, giving rise to a highly reproducible pattern that attracts pollinators and human admirers. First proposed for *A. thaliana* and *Antirrhinum majus*, the ABC model describes how three classes (A, B, and C) of homeotic transcription factors pattern flowers through antagonistic and combinatorial interactions (Bowman et al., 1991) (**Figure 1c-f**). The ABC model is conserved in flowering plants, with different flower phenotypes arising from fading expression boundaries and gene duplication or loss of the A, B, and C class transcription factor genes (Theissen and Melzer, 2007). In the original model, A and C activities are mutually exclusive, establishing the boundary between the sterile outer whorls (perianth, A function) and the reproductive inner whorls (C function) (Bowman et al., 1991). Curiously, the A gene *AP2* is uniformly expressed throughout young floral primordia. Wollmann and co-authors (Wollmann et al., 2010) reconciled this paradox with the discovery that *miR172* and *AP2* expression overlap transiently, which restricts *AP2* activity and reinforces the robust boundary between perianth and reproductive organs (**Figure 1c**). Robust boundaries may also be facilitated by the oligomerization dynamics of A, B, C, and E class proteins. The floral quartet hypothesis (Theissen, 2001) proposes the existence of tetrameric complexes of various ABCE proteins (floral quartets). The increased cooperativity and the higher local concentrations of specific A, B, C, and E class proteins involved in tetramer formation are predicted to sharpen organ identity boundaries.

Population genetics and large-scale phenotypic data demonstrate the role of genetic architecture in robustness

Hall and co-authors (Hall et al., 2007) mapped the first quantitative trait loci (QTL) for trait robustness rather than trait mean in two recombinant inbred populations (RILs), estimating within-genotype robustness with Levene's statistic. They identified 22 robustness QTL across five developmental traits in two conditions. Of these, only three QTL affected exclusively trait robustness, whereas all the others coincided with mean QTL. This strong correlation of robustness and mean QTL agrees with Waddington's view that decreased robustness is associated with decreased function (Hall et al., 2007). Nearly half of the robustness QTL were linked to *ERECTA*, for which a mutant allele segregated in both RILs. *ERECTA* controls aerial organogenesis, and *erecta* mutants show strong pleiotropic phenotypes (Shpak et al., 2004; Shpak et al., 2005). Using the same approach, Sangster and colleagues (Sangster et al., 2008b) mapped robustness QTL for two seedling traits in control and HSP90-reduced conditions, including a third RIL population without a segregating *erecta*-allele. Most robustness QTL did not coincide with mean QTL under control conditions. In contrast, under HSP90-reduced conditions, mean QTL strongly correlated with robustness QTL, consistent with Waddington's notion that newly released variants are poorly buffered (Waddington, 1942a). As in the previous study, heritability of robustness QTL was significantly lower than for mean QTL. Both studies provided empirical evidence for network elements that stabilize particular traits, which was subsequently also shown in maize (Ordas et al., 2008).

Moving from developmental traits to large-scale molecular traits, Jimenez-Gomez and colleagues (Jimenez-Gomez et al., 2011) mapped robustness QTL in Bay x Sha RILs for defense metabolite levels and genome-wide expression. For both datasets, the authors were limited to about four replicates per line for estimating within-genotype robustness

with the coefficient of trait variation (CV, standard deviation/mean). CV of small samples is an unreliable robustness estimate (**Figure 2c**). As CV is strongly mean-driven, the identified robustness QTL may be largely driven by mean differences. Countering this concern, the authors point out that not all mean QTL coincided with robustness QTL. However, all but one robustness QTL coincided with mean QTL. The authors addressed the sample size problem by mapping line-specific CV averages for all 22,746 transcripts, identifying loci that affect global gene expression CV. The major effect QTL contained *ELF3*, an important circadian and flowering time regulator. In a reference background, *ELF3*-Bay and -Sha alleles differentially affected robustness for some traits but not others, possibly due to buffering of the significant global CV expression differences. *ELF3*-Bay and -Sha alleles produce significant mean differences in circadian and developmental phenotypes in reference backgrounds (Coluccio et al., 2011; Davis, personal communication; Jimenez-Gomez et al., 2010). In contrast to *ELF4* (Doyle et al., 2002), *elf3* loss-of-function mutants do not show decreased robustness in circadian or developmental phenotypes (Coluccio et al., 2011; Davis, personal communication; Jimenez-Gomez et al., 2010), although the ELF4 and ELF3 proteins are known to interact (Nusinow et al., 2011).

Together, these studies prove that robustness is a quantitative trait with strong genetic contributions. In reference or hybrid backgrounds, natural alleles can cause different robustness levels in specific traits. In their natural backgrounds, however, these low robustness alleles may be buffered by compensatory mutations. Most *A. thaliana* robustness QTL were trait-specific, suggesting that natural robustness alleles reside in trait-specific sub-networks rather than in robustness master regulators that destabilize many traits and reduce fitness when perturbed. For example, an *HSP90* allele that subtly

decreases robustness in some traits has been found in wild flies, yet this slightly deleterious allele is exceedingly rare (Chen and Wagner, 2012; Sgrò et al., 2010).

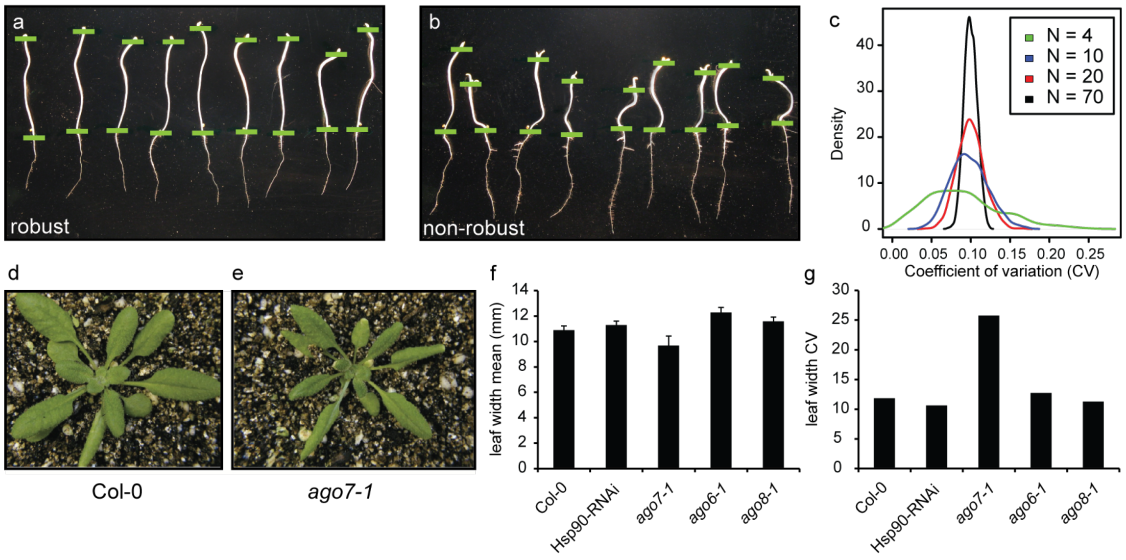


Figure 2. Using quantitative traits to measure robustness. a-b, The early seedling trait hypocotyl length is a sensitive robustness read-out. a) robust 7 day-old, dark-grown wild-type *A. thaliana* seedlings, **b**, non-robust 7 day-old, dark-grown mutant *A. thaliana* seedlings. Used with permission(Vazquez et al., 2004) and reproduced here from the article "Molecular mechanisms of robustness in plants" by Janne Lempe, Jennifer Lachowiec, Alessandra. M. Sullivan, Christine Queitsch, COPB, February 2013. **c**, Coefficient of variation (CV, standard deviation/mean) is an unreliable estimate of population robustness at small sample sizes. Here, we show the distribution of CV estimates for various sample sizes from simulated data. For each sample size N, 1000 samples of size N were drawn from a normal distribution with mean = 10 and SD = 1 (true CV = 0.1), and a CV estimate was calculated for each sample. **d-e,** *ago7* mutants show decreased robustness in leaf width. Representative plants of **d,** wild-type Col-0 and e) *ago7-1*. f) Mean width and **g**, CV of the longest leaf at flowering time was measured in Col-0, an HSP90-reduced line (HSP90-RNAi), and *ago6-1*, *ago7-1*, and *ago8-1*, n=15.

In 2009, Fu and colleagues (Fu et al., 2009) addressed a different robustness angle, genetic buffering, by identifying allelic variation that affects many different traits simultaneously. The authors mapped trait means for 139 developmental traits and 40,580 molecular traits in L*er* x Cvi RILs. Although the parental lines are highly divergent, the authors found only six QTL hot spots with major system-wide effects. These hotspots included well-studied candidate genes such as *ERECTA*, for which a mutant allele segregated, *CRY2*, *HUA2* or *FRL1*, all of which affect plant development pleiotropically. These results are consistent with pervasive genetic buffering that renders the majority of genetic variation phenotypically silent (Fu et al., 2009). These data are reminiscent of data in other organisms showing that most single loss-of-function mutations are phenotypically silent, even in pair-wise combinations (Boone et al., 2007; Davierwala et al., 2005; Giaever et al., 2002; Lehner et al., 2006; Levy and Siegal, 2008b), whereas a small group of highly connected network 'hubs' show epistasis with many different genes (Lehner et al., 2006; Levy and Siegal, 2008b). Although natural alleles are not equivalent to interaction studies with loss-of-function mutations, the fundamental message is the same: eukaryotic genetic networks are mostly robust to perturbation, unless one of a small number of 'fragile nodes' is perturbed (Fu et al., 2009; Lehner et al., 2006; Levy and Siegal, 2008b).

Most recently, Shen and colleagues (Shen et al., 2012) developed a statistical framework to associate trait variation across accessions with genetic variation (vGWAS), re-analyzing a published *A. thaliana* GWA dataset. Their analysis found loci in which allelic variation is associated with accessions that either vary little from each other in trait mean (e.g. flowering time for accessions with loss-of-function *flc* and *fri* mutants, high penetrance alleles) or vary greatly from each other (e.g. flowering time in accessions with functional

FLC and *FRI*, variable penetrance alleles) (**Figure 3**). The authors demonstrate that accounting for these variance or penetrance differences significantly improves heritability and hence mapping of trait means. By eliminating invariant accessions (e.g. accessions with loss-of-function *flc* and *fri* mutants) and focusing on those with variable penetrance (e.g. accessions with functional *FLC* and *FRI*) one could identify background-specific trait determinants. Curiously, the study did not attempt to identify loci that affect accession-variance for many traits.

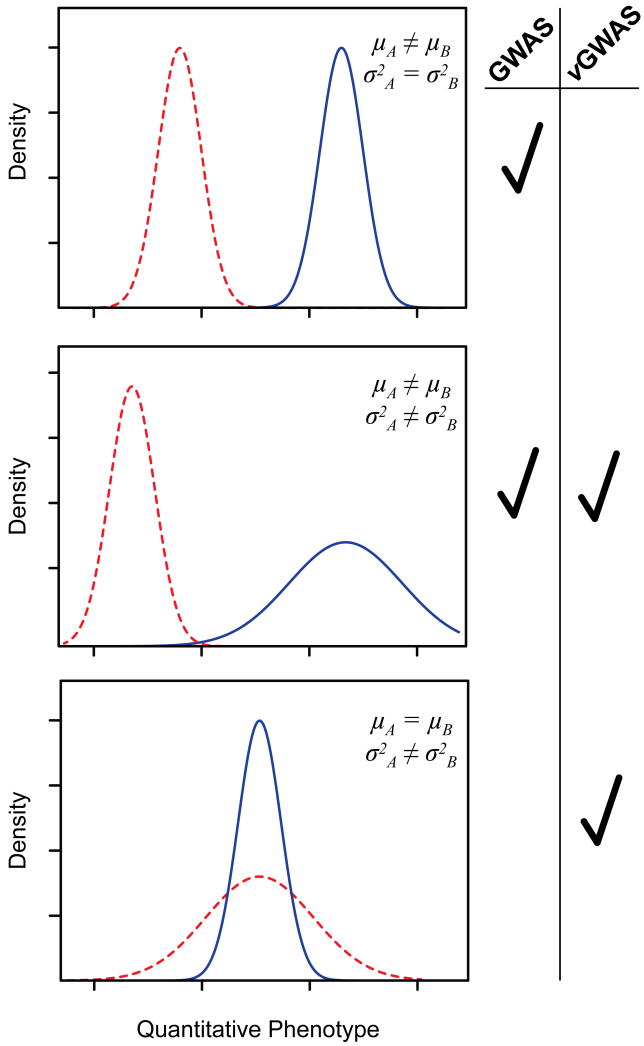


Figure 3. vGWAS loci that cannot be mapped with GWAS. Traditional GWAS associates significant mean differences in quantitative traits with genetic variants (top and middle panel). In contrast, vGWAS associates significant variance differences in trait means with common genetic variants in the same individuals. This approach does not consider within-genotype robustness, rather it is a measure of penetrance of a given polymorphism. Alleles associated with higher variance may or may not affect the trait mean (middle and bottom panels respectively). Thus, taking variance effects into account may improve heritability estimates relative to traditional GWAS.

Future directions and challenges

Similar to other organisms, plants achieve robustness by tightly controlling and buffering developmental decisions in a modular fashion. Whereas the vast majority of perturbations are either phenotypically silent or affect only small sub-networks (local), perturbation of a highly connected fragile node, network hub or robustness master regulator destabilizes many traits (global) (Boone et al., 2007; Davierwala et al., 2005; Fu et al., 2009; Giaever et al., 2002; Lehner et al., 2006; Levy and Siegal, 2008b). The importance and relative contribution of local and global robustness mechanisms to evolvability and adaptation to new environments remain unresolved.

In the absence of systematic deletion or double mutant analyses – thus far phenotype descriptions for 2400 single and 401 mutant combinations have been compiled (Lloyd and Meinke, 2012) – what is the best way to characterize robustness mechanisms in plants going forward? It seems reasonable to focus on robustness master regulators found in other organisms. For example, yeast and worm studies strongly implicate chromatin modifiers in robustness (Lehner et al., 2006; Levy and Siegal, 2008b). Consistent with a role of chromatin in plant robustness, maize mutants in RNA-directed DNA methylation (RdDM) show stochastic developmental defects (Hollick, 2010; Parkinson et al., 2007), suggesting yet another class of small RNAs as robustness agents (**Figure 1g**). Further, defects in ribosome function result in highly pleiotropic developmental defects in humans (Boulon et al., 2010; Freed et al., 2010) and maize (Degenhardt and Bonham-Smith, 2008), suggesting that genes involved in ribosome biogenesis, rRNA processing, and RNA splicing may also maintain robustness.

This approach, however, may not suffice to characterize plant-specific robustness mechanisms. Plants, in particular angiosperms, readily tolerate changes in chromosome number (aneuploidy, polyploidy) and interspecific hybridization (alloploidy) (Jackson and Chen, 2010; Leitch and Leitch, 2008), both of which are thought to be major drivers in angiosperm divergence. A recent analysis of *Arabidopsis* allotetraploids implicated RdDM-specific siRNAs in chromatin and genome stability, whereas miRNAS and tasiRNAs mediated gene expression diversity, possibly facilitating hybrid vigor and adaption (Ha et al., 2009). Curiously, the emergence of RNA polymerase V and IV, key enzymes in RdDM, coincides with the rise of angiosperms (Luo and Hall, 2007).

One of the major draw-backs for current robustness studies is the large sample sizes required for population-based robustness measures. Thus far, individual-based robustness measures remain less amenable to high-throughput analysis. We speculate that comparing organisms perturbed in functionally distinct robustness master regulators may reveal shared molecular features such as specific changes in gene expression, methylation or nucleolar function. These shared features could be leveraged as molecular robustness markers that are applicable to individuals and large populations. Molecular robustness markers would revolutionize the study of robustness in non-model organisms, including humans, and allow us to explore the role of robustness in evolutionary processes and disease susceptibility.

Box 1. The term phenotypic robustness is often conflated with other terms, some of which have slightly different meanings or denote entirely different phenomena. In the following, we attempt to clarify our view and usage of these terms:

Developmental stability – is equivalent to robustness as defined here, describes "the ability of organisms to withstand genetic and environmental perturbation during development, so as to produce a predetermined phenotype" (Auffray, 1999).

Canalization – describes the notion that genetic systems evolve to a robust optimum through stabilizing selection. This robust optimum is thought to arise through elimination of deleterious alleles and reduction of additive genetic effects. Canalization pertains to populations with most individuals clustering around an optimal phenotype (Gibson, 2009; Waddington, 1942a).

Cryptic genetic variation –is genetic variation that is phenotypically silent until revealed by environmental, genetic, or epigenetic perturbations (Gibson and Dworkin, 2004).

Developmental noise – was used originally by Waddington to refer to differences among homologous replicated parts within a single individual and to describe the absence of developmental stability (Debat and David, 2001b). The term is currently often used as "noise" to describe stochastic variation in traits such gene expression, caused by both intrinsic errors and extrinsic micro-environmental fluctuations (Munsky et al., 2012; Raser and O'Shea, 2005). Noise is thought to play an important role in fate determination and circadian clock function (Eldar and Elowitz, 2010; Forde, 2009; Mas and Yanovsky, 2009).

Fluctuating asymmetry – describes an organism's deviation from bilateral symmetry for the whole organism or particular morphological features such as fly wings or bristles. FA is an individual-based measure of robustness. Low FA is thought to correlate with high fitness (Clarke, 1998b).

Variable mutation penetrance – describes the phenomenon that certain mutations show different expressivity (i.e. severity of phenotypic effect) among isogenic individuals. We attribute these expressivity differences to differences in robustness among these individuals. Less robust individuals are expected to show higher mutation penetrance. Variable mutation penetrance among genetically divergent individuals arises from individual-specific genetic and non-genetic modifiers.

Phenotypic plasticity – is the ability of an organism to alter its physiology, morphology, and development in response to changes in its environment (Debat and David, 2001b). In our view, phenotypic plasticity describes changes in phenotype that are pre-determined in existing genetic networks, rather than consequences of stochastic errors in development (that may be ultimately due to extrinsic micro-environmental differences rather intrinsic errors).

Epistasis – is the nonreciprocal interaction of nonallelic genes, in which one gene masks the effects of another. More recently also used to describe interactions of variants with a gene or regulatory region.

Pleiotropy – describes the phenomenon in which a single gene is responsible for several distinct and seemingly unrelated phenotypic effects.

Robustness master regulator – is used here to denote genes that strongly affect robustness. We use this term interchangeably with the terms network hub and fragile node. In yeast, genes that strongly affect robustness are network hubs (Levy and Siegal, 2008b). Studies in plants (Fu et al., 2009) and worms (Lehner et al., 2006) have identified a small number of fragile nodes that affect the penetrance of mutants and natural variants in many other genes. Another frequently used term is 'capacitor', which refers to genes that keep genetic variation phenotypically silent when fully functional and release genetic variation when perturbed (Queitsch et al., 2002; Rutherford and Lindquist, 1998).

Acknowledgements

There have been many excellent studies with relevance to plant robustness in recent years. We apologize to all our colleagues whose work has not been discussed due to space limitations. We would like to thank Corey Snelson and Veronica Di Stilio for sharing unpublished data and providing images for Figure 1d-f, Seth Davis for sharing unpublished information on Bay-and Sha- *ELF3* circadian phenotypes. We thank Maximilian Press and Kerry Bubb for helpful discussions and Maximilian Press for contributing Figure 2c. We thank Stanley Fields and Maximilian Press for critical reading and comments. We acknowledge support by EMBO long-term and HFSP long-term fellowships to JL, by the National Human Genome Research Institute Interdisciplinary Training in Genomic Sciences (T32 HG00035) to JAL, and by a National Science Foundation Graduation Research Fellowship (DGE-0718124) to AMS and JAL. Our research on robustness is supported by the National Science Foundation (MCB-1242744) and the National Institute of Health (DP2OD008371).

Appendix D. Approaches to the study of developmental instability

Developmental instability is a measure of an organism's ability to buffer traits against minor genetic and environmental perturbations during development (Debat and David, 2001a; Waddington, 1942b). Developmental instability thus has a genetic and environmental component; the former component is determined by measuring trait stability of different genotypes in the same environment, and the latter by measuring trait stability of the same genotype in different environments (Debat et al., 2009; Sangster et al., 2008a; Sangster et al., 2008b; Waddington, 1942b). There are many causes of developmental instability. Genetic factors include inbreeding, hybridization and mutations, while environmental factors include deviant climatic conditions, food deficiency, pesticides, and parasitism (Mitton and Grant, 1984; Moller, 1997; Parsons, 1992). As researchers disagree on the nature of developmental instability, their approaches to studying it vary, often leading to conflicting results. Commonly used measures of developmental instability are left-right symmetry of bilateral traits in individuals and trait variability across isogenic populations. In general, early work focuses on characterizing patterns of developmental stability and their causes in various creatures and populations, while more recent studies using modern genetic techniques focus on mapping individual developmental stability genes.

Developmental stability studies have trouble resolving if developmental instability is a trait-specific, organism-specific, or even population-specific feature. One study of

symmetry in 11 invertebrate species shows that asymmetry of one trait within an individual does not predict asymmetry of other traits within the same individual (Clarke, 1998a). Therefore, they find that multiple trait symmetries when considered together have no predictive power to separate more asymmetrical individuals or populations from the rest of the sample. This result disagrees with a study of bilateral symmetry in island populations of lizards, which shows that populations of lizards with increased asymmetry in one trait show a tendency to be more asymmetrical for other traits (Soule, 1965; Soule, 1967). Another experiment analyzing isogenic drosophila wing mutants finds that the most variable genotypes for wing *size* are also the most asymmetric. Confusingly, the most variable genotypes for wing *shape* are *not* asymmetric (Debat et al., 2009). One plausible explanation for the inconsistency in results is that symmetry is a special case of developmental stability, limiting the general nature of the conclusions that can be drawn.

In general developmental stability is not controlled by a single gene or environmental input, but rather appears to be multigenic and responsive to many environmental inputs. Most studies support a hierarchical model in which a few general stabilizing mechanisms buffer global developmental noise and a number of trait-specific stabilizing mechanisms buffer local developmental noise (Takahashi et al., 2011b).

The capacitor and protein chaperone HSP90 is a good candidate for a general stabilizing mechanism because flies, plants, and fish with disrupted HSP90 all display increased developmental instability (Queitsch et al., 2002; Rutherford and Lindquist, 1998; Yeyati et al., 2007). In these organisms, HSP90 is known to buffer development against both genetic and environmental perturbation and was shown to do so using both

population-wide trait variance and individual-specific asymmetry measures (Queitsch et al., 2002; Rutherford and Lindquist, 1998; Yeyati et al., 2007). However, it is worth noting that while HSP90-mediated trait instability is often accompanied by a loss of symmetry in plants and fish, no such loss of symmetry has been observed in flies (Milton et al., 2003). In plants, HSP90-mediated developmental stability is a phenotypic trait that is even amenable to quantitative genetic mapping (Sangster et al., 2008b).

In an attempt to screen for genetic components of a global stabilizing mechanism, Levy and Siegal used data from a large-scale deficiency mapping experiment of 4,718 single-gene knockout haploid yeast strains to find genes that alter developmental stability (Levy and Siegal, 2008a). To do this they screened 220 measurable yeast phenotypes (e.g. budding angle, nuclear size and actin patches) seeking the mutants with increased variance of multiple traits simultaneously. They found that three hundred genes contribute to developmental instability when absent. These genes are characterized as 'network hubs' that encode proteins that are highly connected. Genes involved in critical cellular processes such as chromosome organization, DNA integrity, RNA elongation, protein modification, cell cycle, and response to stress, are overrepresented among the 300 developmental stability genes. Essential or redundant genes were not tested, which may explain why HSP90 was not identified in this study. Environmentally sensitive mechanisms for developmental stability were also not detected because the mutants were screened under a single environmental condition. It remains unclear how these findings apply to developmental instability of complex, multicellular traits.

In addition to deficiency mapping in yeast, there have been two deficiency mapping studies in flies that have shed some light on the genetic underpinnings of developmental instability. The first study sought to map regions responsible for increased developmental instability of trait variance by measuring population-wide variation in pre-adult developmental period (Takahashi et al., 2011a). In total, 11 regions are responsible for instability in developmental period; all but one region displays sex-specific effects. Regions that increase variability in developmental period do not significantly correlate with increases in wing shape variability or wing symmetry leading to the conclusion that they are at least partially independent. However, of the 11 deletion mutants that have increased developmental period instability, 4 have significantly longer mean developmental period, and 3 of these 4 mutants show an increase in wing asymmetry. This result suggests that severe developmental disruption may be more likely to cause disruptions in symmetry. A similar asymmetry study using the same collection of deletion mutants found 92 deficiencies increased trait asymmetry. This study finds significant correlation between asymmetry of wing traits; however only a few genomic regions had a significant effect on asymmetry of both wing *and* bristle traits (Takahashi et al., 2011b). The interpretation of these results is limited due to the fact that the average deficiency encompasses 63 genes. Fine scale mapping will be required to disentangle the genes ultimately responsible for developmental stability.

It is clear that even in the modern genomic era the mechanisms behind developmental stability remain a controversial topic. Much of the controversy is likely due to the limited number of traits assessed and the lack of continuity in traits and measurement techniques across experiments in different organisms. Developmental

stability, in particular trait symmetry, has long been hypothesized to be adaptive. While this hypothesis remains to be formally tested in the field, selection experiments in flies show that under any selection regime (fluctuating, disruptive and stabilizing) trait asymmetry increases (Pelabon et al., 2010). This result brings us to the apparent paradox that organisms must have stable phenotypes to be well adapted to their environment, yet stability must be disrupted to adapt to new environments. To date, the capacitor HSP90 remains the best and only known mechanism that allows for both stability under standard conditions and reversible instability in novel environments.

References

Adrian, J., Farrona, S., Reimer, J.J., Albani, M.C., Coupland, G., and Turck, F. (2010). cis-Regulatory Elements and Chromatin State Coordinately Control Temporal and Spatial Expression of FLOWERING LOCUS T in Arabidopsis. Plant Cell *22*, 1425-1440.

Allen, M.D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., and Suzuki, M. (1998). A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. The EMBO Journal *17*, 5484-5496.

Alvarez-Venegas, R. (2010). Regulation by polycomb and trithorax group proteins in Arabidopsis. Arabidopsis Book *8*, e0128.

Antoniou, M., Deboer, E., Habets, G., and Grosveld, F. (1988). The Human Beta-Globin Gene Contains Multiple Regulatory Regions - Identification of One Promoter and 2 Downstream Enhancers. Embo Journal *7*, 377-384.

Atwell, S., Huang, Y.S., Vilhjalmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T.*, et al.* (2010a). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature *465*, 627-631.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T.*, et al.* (2010b). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature *465*, 627-631.

Auffray, J.-C. (1999). Shape asymmetry and developmental stability. In On Growth and Form: Spatiotemporal Patterning in Biology, M.A.J. Chaplain, ed. (John Wiley & Sons), pp. 310-324.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic acids research *37*, W202-208.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729-740.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a Beta-Globin Gene Is Enhanced by Remote Sv40 DNA-Sequences. Cell *27*, 299-308.

Belmonte, M.F., Kirkbride, R.C., Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, M., Munoz, M.D.*, et al.* (2013). Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. P Natl Acad Sci USA *110*, E435-E444.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of molecular biology *112*, 535-542.

Bianchi, M.E., and Agresti, A. (2005). HMG proteins: dynamic players in gene regulation and differentiation. Curr Opin Genet Dev *15*, 496-506.

Boone, C., Bussey, H., and Andrews, B.J. (2007). Exploring genetic interactions and networks with yeast. Nature reviews Genetics *8*, 437-449.

Boulon, S., Westman, B.J., Hutten, S., Boisvert, F.M., and Lamond, A.I. (2010). The nucleolus under stress. Molecular cell *40*, 216-227.

Bowman, J.L., Smyth, D.R., and Meyerowitz, E.M. (1991). Genetic interactions among floral homeotic genes of Arabidopsis. Development *112*, 1-20.

Brown, B.A., and Jenkins, G.I. (2008). UV-B signaling pathways with different fluence-rate response profiles are distinguished in mature Arabidopsis leaf tissue by requirement for UVR8, HY5, and HYH. Plant Physiology *146*, 576-588.

Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2007). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research *36*, D102-D106.

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. Nat Methods *10*, 1213-1218.

Burga, A., Casanueva, M.O., and Lehner, B. (2011). Predicting mutation outcome from early stochastic variation in genetic interaction partners. Nature *480*, 250-253.

Calhoun, V.C., and Levine, M. (2003). Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. Proc Natl Acad Sci U S A *100*, 9878-9883.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C*., et al.* (2011a). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature genetics *43*, 956-963.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C*., et al.* (2011b). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet *43*, 956-963.

Cao, S., Kumimoto, R.W., Gnesutta, N., Calogero, A.M., Mantovani, R., and Holt, B.F., 3rd (2014). A distal CCAAT/NUCLEAR FACTOR Y complex promotes chromatin looping at the FLOWERING LOCUS T promoter and regulates the timing of flowering in Arabidopsis. Plant Cell *26*, 1009-1017.

Carroll, S.B. (1995). Evolution and Development of the Insect Body Plan. Molecular Biology of the Cell *6*, 5-5.

Casanueva, M.O., Burga, A., and Lehner, B. (2012). Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans. Science *335*, 82-85.

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

Chen, B., and Wagner, A. (2012). Hsp90 is important for fecundity, longevity, and buffering of cryptic deleterious variation in wild fly populations. BMC Evol Biol *12*, 25.

Chen, H., Zhang, J., Neff, M.M., Hong, S.W., Zhang, H., Deng, X.W., and Xiong, L. (2008a). Integration of light and abscisic acid signaling during seed germination and early seedling development. Proc Natl Acad Sci U S A *105*, 4495-4500.

Chen, H., Zhang, J., Neff, M.M., Hong, S.W., Zhang, H., Deng, X.W., and Xiong, L. (2008b). Integration of light and abscisic acid signaling during seed germination and early seedling development. P Natl Acad Sci USA *105*, 4495-4500.

Chitwood, D.H., Nogueira, F.T.S., Howell, M.D., Montgomery, T.A., Carrington, J.C., and Timmermans, M.C.P. (2009). Pattern formation via small RNA mobility. Gene Dev *23*, 549-554.

Chung, J.H., Bell, A.C., and Felsenfeld, G. (1997). Characterization of the chicken beta-globin insulator. Proc Natl Acad Sci U S A *94*, 575-580.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A*., et al.* (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science *317*, 338-342.

Clark, R.M., Wagler, T.N., Quijada, P., and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet *38*, 594-597.

Clarke, G.M. (1998a). The genetic basis of developmental stability. IV. Individual and population asymmetry parameters. Heredity (Edinb) *80*, 553-651.

Clarke, G.M. (1998b). The genetic basis of developmental stability. IV. Individual and population asymmetry parameters. Heredity *80*, 553-561.

Coluccio, M.P., Sanchez, S.E., Kasulin, L., Yanovsky, M.J., and Botto, J.F. (2011). Genetic mapping of natural variation in a shade avoidance response: ELF3 is the candidate gene for a QTL in hypocotyl growth regulation. Journal of experimental botany *62*, 167-176.

Cong, L., Ran, F.A., Cox, D., Lin, S.L., Barretto, R., Habib, N., Hsu, P.D., Wu, X.B., Jiang, W.Y., Marraffini, L.A*., et al.* (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819-823.
Crews, S.T., and Pearson, J.C. (2009). Transcriptional autoregulation in development. Curr Biol *19*, R241-246.

Davidson, E.H., and Britten, R.J. (1971). Control of Gene Expression during Development. Journal of Theoretical Biology *32*, 123-&.

Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y.*, et al.* (2005). The synthetic genetic interaction spectrum of essential genes. Nat Genet *37*, 1147-1152.

Davis, S.J. (personal communication).

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

de Meaux, J., Goebel, U., Pop, A., and Mitchell-Olds, T. (2005). Allele-specific assay reveals functional variation in the chalcone synthase promoter of Arabidopsis thaliana that is compatible with neutral evolution. Plant Cell *17*, 676-690.

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.

Dean, G., Cao, Y.G., Xiang, D.Q., Provart, N.J., Ramsay, L., Ahad, A., White, R., Selvaraj, G., Datla, R., and Haughn, G. (2011). Analysis of Gene Expression Patterns during Seed Coat Development in Arabidopsis. Mol Plant *4*, 1074-1091.

Debat, V., and David, P. (2001a). Mapping phenotypes: canalization, plasticity, and developmental stability. TRENDs in Ecology and Evolution *16*, 555-561.

Debat, V., and David, P. (2001b). Mapping phenotypes:canalization, plasticity and developmental stability. Trends in ecology & evolution *16*, 555-561.

Debat, V., Debelle, A., and Dworkin, I. (2009). Plasticity, Canalization, and developmental stability of the Drosophila wing: joint effects of mutations and developmental temperature. Evolution *63*, 2864-2876.
DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006). Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics *172*, 1155-1164.

Defossez, P.A., and Gilson, E. (2002). The vertebrate protein CTCF functions as an insulator in Saccharomyces cerevisiae. Nucleic Acids Res *30*, 5136-5141.

Degenhardt, R.F., and Bonham-Smith, P.C. (2008). Arabidopsis ribosomal proteins RPL23aA and RPL23aB are differentially targeted to the nucleolus and are disparately required for normal development. Plant Physiology *147*, 128-142.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature reviews Genetics *14*, 390-403.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Dombrecht, B., Xue, G.P., Sprague, S.J., Kirkegaard, J.A., Ross, J.J., Reid, J.B., Fitt, G.P., Sewelam, N., Schenk, P.M., Manners, J.M.*, et al.* (2007). MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. Plant Cell *19*, 2225-2245.

Doyle, M.R., Davis, S.J., Bastow, R.M., McWatters, H.G., Kozma-Bognar, L., Nagy, F., Millar, A.J., and Amasino, R.M. (2002). The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. Nature *419*, 74-77.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform *23*, 205-211.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167-173.
Elgin, S.C.R. (1988). The Formation and Function of Dnase-I Hypersensitive Sites in the Process of Gene Activation. Journal of Biological Chemistry *263*, 19259-19262.

ENCODE (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.
Erhard, K.F., Jr., Parkinson, S.E., Gross, S.M., Barbour, J.E., Lim, J.P., and Hollick, J.B. (2013). Maize RNA polymerase IV defines trans-generational epigenetic variation. Plant Cell *25*, 808-819.

Esfandiari, E., Jin, Z., Abdeen, A., Griffiths, J.S., Western, T.L., and Haughn, G.W. (2013). Identification and analysis of an outer-seed-coat-specific promoter from Arabidopsis thaliana. Plant Mol Biol *81*, 93-104.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014a). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Mol Cell *55*, 694-707.

Feng, Z.Y., Mao, Y.F., Xu, N.F., Zhang, B.T., Wei, P.L., Yang, D.L., Wang, Z., Zhang, Z.J., Zheng, R., Yang, L.*, et al.* (2014b). Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. P Natl Acad Sci USA *111*, 4632-4637.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K.*, et al.* (2010). The Pfam protein families database. Nucleic Acids Res *38*, D211-222.
Forde, B.G. (2009). Is it good noise? The role of developmental instability in the shaping of a root system. Journal of experimental botany *60*, 3989-4002.

Fourel, G., Boscheron, C., Revardel, E., Lebrun, E., Hu, Y.F., Simmen, K.C., Muller, K., Li, R., Mermod, N., and Gilson, E. (2001). An activation-independent role of transcription factors in insulator function. EMBO Rep *2*, 124-132.

Franco-Zorrilla, J.M., Lopez-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A *111*, 2367-2372.

Freed, E.F., Bleichert, F., Dutca, L.M., and Baserga, S.J. (2010). When ribosomes go bad: diseases of ribosome biogenesis. Mol Biosyst *6*, 481-493.

Fu, J., Keurentjes, J.J.B., Bouwmeester, H., America, T., Verstappen, F.W.A., Ward, J.L., Beale, M.H., de Vos, R.C.H., Dijkstra, M., Scheltema, R.A.*, et al.* (2009). System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genet *41*, 166-167.

Fujimoto, S.Y., Ohta, M., Usui, A., Shinshi, H., and Ohme-Takagi, M. (2000). Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. The Plant Cell *12*, 393-404.

Galas, D.J., and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res *5*, 3157-3170.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T.*, et al.* (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature *477*, 419-423.

Gangaraju, V.K., Yin, H., Weiner, M.M., Wang, J., Huang, X.A., and Lin, H. (2011). Drosophila Piwi functions in Hsp90-mediated suppression of phenotypic variation. Nat Genet *43*, 153-158.

Garcia-Fayos, P., Bochet, E., and Cerda, A. (2010). Seed removal susceptibility through soil erosion shapes vegetation composition. Plant Soil *334*, 289-297.

Garwood, N.C. (1985). The Role of Mucilage in the Germination of Cuipo, Cavanillesia-Platanifolia (H-and-B) Hbk (Bombacaceae), a Tropical Tree. Am J Bot *72*, 1095-1105.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K.*, et al.* (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775-1787.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B.*, et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387-391.

Gibson, G. (2009). Decanalization and the origin of complex disease. Nature reviews Genetics *10*, 134-140.

Gibson, G., and Dworkin, I. (2004). Uncovering cryptic genetic variation. Nat Rev Genet *5*, 681-690.

Groudine, M., and Weintraub, H. (1982). Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell *30*, 131-139.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res *33*, W557-559.

Guo, F., Liu, C., Xia, H., Bi, Y., Zhao, C., Zhao, S., Hou, L., Li, F., and Wang, X. (2013). Induced expression of AtLEC1 and AtLEC2 differentially promotes somatic embryogenesis in transgenic tobacco plants. PLoS One *8*, e71714.

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

Gutierrez, C. (2009). The Arabidopsis cell division cycle. The Arabidopsis book / American Society of Plant Biologists *7*, e0120.

Gutterman, Y., and Shem-Tov, S. (1997). The efficiency of the strategy of mucilaginous seeds of some common annuals of the Negev adhering to the soil crust to delay collection by ants. Israel J Plant Sci *45*, 317-327.

Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K.D., Chapman, E.J., Carrington, J.C., Chen, X., Wang, X.-J., and Chen, Z.J. (2009). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proceedings of the National Academy of Sciences *106*, 17835-17840.

Hall, M.C., Dworkin, I., Ungerer, M.C., and Purugganan, M. (2007). Genetics of microenvironmental canalization in Arabidopsis thaliana. Proc Natl Acad Sci U S A *104*, 13717-13722.

Hao, J., Tu, L., Hu, H., Tan, J., Deng, F., Tang, W., Nie, Y., and Zhang, X. (2012). GbTCP, a cotton TCP transcription factor, confers fibre elongation and root hair development by a complex regulating system. J Exp Bot *63*, 6267-6281.

Haughn, G.W., and Western, T.L. (2012). Arabidopsis Seed Coat Mucilage is a Specialized Cell Wall that Can be Used as a Model for Genetic Analysis of Plant Cell Wall Structure and Function. Front Plant Sci *3*, 64.

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.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y.T., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C.X., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics *39*, 311-318.

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.*, et al.* (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Meth *6*, 283-289.

Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A.*, et al.* (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449-452.

Hollick, J.B. (2010). Paramutation and development. Annu Rev Cell Dev Biol *26*, 557-579.

Hong, J.W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Honma, T., and Goto, K. (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature *409*, 525-529.

Hornstein, E., and Shomron, N. (2006). Canalization of development by microRNAs. Nat Genet *38 Suppl*, S20-24.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huang, L., Jones, A.M., Searle, I., Patel, K., Vogler, H., Hubner, N.C., and Baulcombe, D.C. (2009). An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. Nat Struct Mol Biol *16*, 91-93.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S.*, et al.* (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res *40*, D306-312.

International_Rice_Genome_Sequencing_Project (2005). The map-based sequence of the rice genome. Nature *436*, 793-800.

Jackson, S., and Chen, Z.J. (2010). Genomic and expression plasticity of polyploidy. Current opinion in plant biology *13*, 153-159.

Jarosz, D.F., and Lindquist, S. (2010). Hsp90 and environmental stress transform the adaptive value of natural genetic variation. Science *330*, 1820-1824.

Jiang, W., Zhou, H., Bi, H., Fromm, M., Yang, B., and Weeks, D.P. (2013). Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. Nucleic Acids Res *41*, e188.

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

Jimenez-Gomez, J.M., Corwin, J.A., Joseph, B., Maloof, J.N., and Kliebenstein, D.J. (2011). Genomic Analysis of QTLs and Genes Altering Natural Variation in Stochastic Noise. PLoS Genet *7*, e1002295.

Jimenez-Gomez, J.M., Wallace, A.D., and Maloof, J.N. (2010). Network analysis identifies ELF3 as a QTL for the shade avoidance response in Arabidopsis. PLoS genetics *6*.

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. Nature genetics *43*, 264-268.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nat Genet *37*, 761-765.

Keene, M.A., Corces, V., Lowenhaupt, K., and Elgin, S.C. (1981). DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. P Natl Acad Sci USA *78*, 143-146.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant journal : for cell and molecular biology *50*, 347-363.

Kim, T.S., Kim, W.Y., Fujiwara, S., Kim, J., Cha, J.Y., Park, J.H., Lee, S.Y., and Somers, D.E. (2011). HSP90 functions in the circadian clock through stabilization of the client F-box protein ZEITLUPE. Proc Natl Acad Sci U S A *108*, 16843-16848.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.

Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004). Naturally occurring genetic variation in Arabidopsis thaliana. Annu Rev Plant Biol *55*, 141-172.

Kotak, S., Vierling, E., Baumlein, H., and von Koskull-Doring, P. (2007). A novel transcriptional cascade regulating expression of heat stress proteins during seed development of Arabidopsis. The Plant Cell *19*, 182-195.

Krishna, P., and Gloor, G. (2001). The Hsp90 family of proteins in Arabidopsis thaliana. Cell stress & chaperones *6*, 238-246.

Kumar, S.V., and Wigge, P.A. (2010). H2A.Z-containing nucleosomes mediate the thermosensory response in Arabidopsis. Cell *140*, 136-147.

Kuno, N., Moller, S.G., Shinomura, T., Xu, X., Chua, N.H., and Furuya, M. (2003). The novel MYB protein EARLY-PHYTOCHROME-RESPONSIVE1 is a component of a slave circadian oscillator in Arabidopsis. The Plant Cell *15*, 2476-2488.

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.

Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A., and Hughes, T.R. (2011). Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucleic Acids Res *39*, 4680-4690.
Laux, T., Wurschum, T., and Breuninger, H. (2004). Genetic regulation of embryonic pattern formation. Plant Cell *16 Suppl*, S190-202.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics *11*, 204-220.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet *38*, 896-903.

Leitch, A.R., and Leitch, I.J. (2008). Genomic plasticity and the diversity of polyploid plants. Science *320*, 481-483.

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

Leon-Kloosterziel, K.M., Keijzer, C.J., and Koornneef, M. (1994). A Seed Shape Mutant of Arabidopsis That Is Affected in Integument Development. Plant Cell *6*, 385-392.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127-128.

Levine, M. (2010a). Transcriptional enhancers in animal development and evolution. Curr Biol *20*, R754-763.

Levine, M. (2010b). Transcriptional Enhancers in Animal Development and Evolution. Current Biology *20*, R754-R763.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

Levy, S., and Siegal, M. (2008a). Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol *6*, e264-

Levy, S.F., and Siegal, M.L. (2008b). Network hubs buffer environmental variation in Saccharomyces cerevisiae. PLoS Biol *6*, e264.

Li, G., Siddiqui, H., Teng, Y., Lin, R., Wan, X.Y., Li, J., Lau, O.S., Ouyang, X., Dai, M., Wan, J.*, et al.* (2011). Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis. Nature cell biology *13*, 616-622.

Li, P., Filiault, D., Box, M.S., Kerdaffrec, E., van Oosterhout, C., Wilczek, A.M., Schmitt, J., McMullan, M., Bergelson, J., Nordborg, M.*, et al.* (2014). Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in Arabidopsis thaliana. Genes Dev *28*, 1635-1640.

Lindquist, S. (1986). The heat-shock response. Annual review of biochemistry *55*, 1151-1191.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

Liu, L., Adrian, J., Pankin, A., Hu, J., Dong, X., von Korff, M., and Turck, F. (2014). Induced and natural variation of promoter length modulates the photoperiodic response of FLOWERING LOCUS T. Nat Commun *5*, 4558.

Lloyd, J., and Meinke, D. (2012). A Comprehensive Dataset of Genes with a Loss-of-Function Mutant Phenotype in Arabidopsis thaliana. Plant Physiology.

Luo, J., and Hall, B.D. (2007). A multistep process gave rise to RNA polymerase IV of land plants. J Mol Evol *64*, 101-112.

Luo, J., Yoshikawa, N., Hodson, M.C., and Hall, B.D. (2007). Duplication and paralog sorting of RPB2 and RPB1 genes in core eudicots. Mol Phylogenet Evol *44*, 850-862.

Ma, L., Li, J., Qu, L., Hager, J., Chen, Z., Zhao, H., and Deng, X.W. (2001). Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways. The Plant Cell *13*, 2589-2607.

Mas, P., and Yanovsky, M.J. (2009). Time for circadian rhythms: plants get synchronized. Curr Opin Plant Biol *12*, 574-579.

Masel, J. (2006). Cryptic genetic variation is enriched for potential adaptations. Genetics *172*, 1985-1991.

Masel, J., and Siegal, M.L. (2009). Robustness: mechanisms and consequences. Trends Genet *25*, 395-403.

Mason, M.G., Mathews, D.E., Argyros, D.A., Maxwell, B.B., Kieber, J.J., Alonso, J.M., Ecker, J.R., and Schaller, G.E. (2005). Multiple type-B response regulators mediate cytokinin signal transduction in Arabidopsis. Plant Cell *17*, 3007-3018.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research *34*, D108-110.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A.J. (2007). Targets of RNA-directed DNA methylation. Curr Opin Plant Biol *10*, 512-519.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012a). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, NY) *337*, 1190-1195.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012b). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

Megraw, M., and Hatzigeorgiou, A.G. (2010). MicroRNA Promoter Analysis. In Plant MicroRNAs, B.C. Meyers, and P.J. Green, eds. (Humana Press), pp. 149-161.

Meijon, M., Satbhai, S.B., Tsuchimatsu, T., and Busch, W. (2014). Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis. Nat Genet *46*, 77-81.

Mendes, M.A., Guerra, R.F., Berns, M.C., Manzo, C., Masiero, S., Finzi, L., Kater, M.M., and Colombo, L. (2013). MADS domain transcription factors mediate short-range DNA looping that is essential for target gene expression in Arabidopsis. Plant Cell *25*, 2560-2572.

Meyerowitz, E.M. (2002). Plants compared to animals: the broadest comparative study of development. Science *295*, 1482-1485.

Miao, J., Guo, D., Zhang, J., Huang, Q., Qin, G., Zhang, X., Wan, J., Gu, H., and Qu, L.J. (2013). Targeted mutagenesis in rice using CRISPR-Cas system. Cell Res *23*, 1233-1236.

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M.*, et al.* (2008). Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS genetics *4*, e14.

Michaels, S.D. (2009). Flowering time regulation produces much fruit. Current Opinion in Plant Biology *12*, 75-80.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. Science (New York, NY) *303*, 1538-1542.

Milton, C.C., Huynh, B., Batterham, P., Rutherford, S.L., and Hoffmann, A.A. (2003). Quantitative trait symmetry independent of Hsp90 buffering: distinct modes of genetic canalization and developmental stability. Proc Natl Acad Sci U S A *100*, 13396-13401.

Mitchell-Olds, T., and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature *441*, 947-952.

Mittler, R., Kim, Y., Song, L., Coutu, J., Coutu, A., Ciftci-Yilmaz, S., Lee, H., Stevenson, B., and Zhu, J.K. (2006). Gain- and loss-of-function mutations in Zat10 enhance the tolerance of plants to abiotic stress. FEBS letters *580*, 6537-6542.

Mitton, J.B., and Grant, M.C. (1984). Associations among protein heterozygositym, growth rate, and developmental homeostasis. Annual Review of Ecology and Systematics *15*, 479-499.

Mok, D.W.S., and Mok, M.C. (1994). CytokininsChemistry, Activity, and Function (CRC Press).
Moller, A.P. (1997). Developmental stability and fitness: a review. The American Naturalist *149*, 916-932.

Morgan, W.D., Williams, G.T., Morimoto, R.I., Greene, J., Kingston, R.E., and Tjian, R. (1987). 2 Transcriptional Activators, Ccaat-Box-Binding Transcription Factor and Heat-Shock Transcription Factor, Interact with a Human Hsp70-Gene Promoter. Molecular and Cellular Biology *7*, 1129-1138.

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.

Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. Science *336*, 183-187.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S.*, et al.* (2012a). BEDOPS: high-performance genomic feature operations. Bioinformatics (Oxford, England) *28*, 1919-1920.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012b). Circuitry and dynamics of human transcription factor regulatory networks. Cell *150*, 1274-1286.
Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K.*, et al.* (2012c). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Nord, A.S., Blow, M.J., Attanasio, C., Akiyama, J.A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V.*, et al.* (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell *155*, 1521-1531.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R.*, et al.* (2005). The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol *3*, e196.

North, H.M., Berger, A., Saez-Aguayo, S., and Ralet, M.C. (2014). Understanding polysaccharide production and properties using seed coat mutants: future perspectives for the exploitation of natural variants. Ann Bot.

Nover, L., Bharti, K., Doring, P., Mishra, S.K., Ganguli, A., and Scharf, K.D. (2001). Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? Cell stress & chaperones *6*, 177-189.

Nusinow, D.A., Helfer, A., Hamilton, E.E., King, J.J., Imaizumi, T., Schultz, T.F., Farre, E.M., and Kay, S.A. (2011). The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. Nature *475*, 398-402.

Ordas, B., Malvar, R.A., and Hill, W.G. (2008). Genetic variation and quantitative trait loci associated with developmental stability and the environmental correlation between traits in maize. Genet Res (Camb) *90*, 385-395.

Parkinson, S.E., Gross, S.M., and Hollick, J.B. (2007). Maize sex determination and abaxial leaf fates are canalized by a factor that maintains repressed epigenetic states. Developmental biology *308*, 462-473.
Parsons, P.A. (1992). Fluctuating asymmetry: a biological monitor of environmental and genomic stress. Heredity (Edinb) *68*, 361-364.

Pelabon, C., Hansen, T.F., Carter, A.J., and Houle, D. (2010). Evolution of variation and variability under fluctuating, stabilizing, and disruptive selection. Evolution *64*, 1912-1925.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow Enhancers Foster Robustness of Drosophila Gastrulation. Current Biology *20*, 1562-1567.

Plantegenet, S., Weber, J., Goldstein, D.R., Zeller, G., Nussbaumer, C., Thomas, J., Weigel, D., Harshman, K., and Hardtke, C.S. (2009). Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. Molecular systems biology *5*, 242.

Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J*., et al.* (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res *19*, 1316-1323.

Queitsch, C., Hong, S.W., Vierling, E., and Lindquist, S. (2000). Heat shock protein 101 plays a crucial role in thermotolerance in Arabidopsis. The Plant Cell *12*, 479-492.

Queitsch, C., Sangster, T.A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. Nature *417*, 618-624.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913-918.

Raser, J.M., and O'Shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.

Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Mol Cell *33*, 192-203.

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

Rossel, J.B., Wilson, P.B., Hussain, D., Woo, N.S., Gordon, M.J., Mewett, O.P., Howell, K.A., Whelan, J., Kazan, K., and Pogson, B.J. (2007). Systemic and intracellular responses to photooxidative stress in Arabidopsis. The Plant Cell *19*, 4091-4110.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F*., et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.

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

Rutherford, S.L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. Nature *396*, 336-342.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A*., et al.* (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Meth *3*, 511-518.

Sangster, T.A., Lindquist, S., and Queitsch, C. (2004). Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. Bioessays *26*, 348-362.

Sangster, T.A., Salathia, N., Lee, H.N., Watanabe, E., Schellenberg, K., Morneau, K., Wang, H., Undurraga, S., Queitsch, C., and Lindquist, S. (2008a). HSP90-buffered genetic variation is common in Arabidopsis thaliana. Proc Natl Acad Sci U S A *105*, 2969-2974.

Sangster, T.A., Salathia, N., Undurraga, S., Milo, R., Schellenberg, K., Lindquist, S., and Queitsch, C. (2008b). HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. Proc Natl Acad Sci U S A *105*, 2963-2968.

Scharf, K.D., Berberich, T., Ebersberger, I., and Nover, L. (2012). The plant heat stress transcription factor (Hsf) family: structure, function and evolution. Biochim Biophys Acta *1819*, 104-119.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A*., et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112-1115.

Schneitz, K., Hulskamp, M., and Pruitt, R.E. (1995). Wild-Type Ovule Development in Arabidopsis-Thaliana - a Light-Microscope Study of Cleared Whole-Mount Tissue. Plant J *7*, 731-749.

Schramm, F., Ganguli, A., Kiehlmann, E., Englich, G., Walch, D., and von Koskull-Doring, P. (2006). The heat stress transcription factor HsfA2 serves as a regulatory amplifier of a subset of genes in the heat stress response in Arabidopsis. Plant molecular biology *60*, 759-772.

Sgrò, C.M., Wegener, B., and Hoffmann, A.A. (2010). A naturally occurring variant of Hsp90 that is associated with decanalization. Proceedings of the Royal Society B: Biological Sciences *277*, 2049-2057.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L*., et al.* (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. Nat Biotechnol *31*, 686-688.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shen, X., Pettersson, M., Rönnegård, L., and Carlborg, Ö. (2012). Inheritance Beyond Plain Heritability: Variance-Controlling Genes in <italic>Arabidopsis thaliana</italic>. PLoS Genet *8*, e1002839.

Shin, J., Kim, K., Kang, H., Zulfugarov, I.S., Bae, G., Lee, C.H., Lee, D., and Choi, G. (2009). Phytochromes promote seedling light responses by inhibiting four negatively-acting phytochrome-interacting factors. P Natl Acad Sci USA *106*, 7660-7665.

Shindo, C., Aranzana, M.J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., and Dean, C. (2005). Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. Plant Physiol *138*, 1163-1173.

Shiu, S.H., and Bleecker, A.B. (2001). Plant receptor-like kinase gene family: diversity, function, and signaling. Sci STKE *2001*, re22.

Shpak, E.D., Berthiaume, C.T., Hill, E.J., and Torii, K.U. (2004). Synergistic interaction of three ERECTA-family receptor-like kinases controls Arabidopsis organ growth and flower development by promoting cell proliferation. Development *131*, 1491-1501.

Shpak, E.D., McAbee, J.M., Pillitteri, L.J., and Torii, K.U. (2005). Stomatal patterning and differentiation by synergistic interactions of receptor kinases. Science *309*, 290-293.

Sieber, P., Wellmer, F., Gheyselinck, J., Riechmann, J.L., and Meyerowitz, E.M. (2007). Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. Development *134*, 1051-1060.

Sieburth, L.E., and Meyerowitz, E.M. (1997). Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically. The Plant Cell Online *9*, 355-365.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J*., et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

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

Sollars, V., Lu, X., Xiao, L., Wang, X., Garfinkel, M.D., and Ruden, D.M. (2003). Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. Nature genetics *33*, 70-74.
Soule, M.E. (1965). Phenetics of natural populations I. Phenetic relationships of insular populations of the side-blotched lizard. Evolution *21*, 584-591.

Soule, M.E. (1967). Phenetics of natural populations. II. Asymmetry and evolution in a lizard. American Naturalist *101*, 141-159.

Spartz, A.K., Lee, S.H., Wenger, J.P., Gonzalez, N., Itoh, H., Inzé, D., Peer, W.A., Murphy, A.S., Overvoorde, P.J., and Gray, W.M. (2012). The SAUR19 subfamily of SMALL AUXIN UP RNA genes promote cell expansion. The Plant Journal.

Specchia, V., Piacentini, L., Tritto, P., Fanti, L., D'Alessandro, R., Palumbo, G., Pimpinelli, S., and Bozzetti, M.P. (2010). Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. Nature *463*, 662-665.

Spiker, S. (1985). Plant Chromatin Structure. Annu Rev Plant Phys *36*, 235-253.

Stergachis, A., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E., Akey, J.A*., et al.* (2013a). Exonic transcription factor binding directs codon choice and impacts protein evolution. Science (New York, NY) *in press*.

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S.L., Vernot, B., Cheng, J.B., Thurman, R.E., Sandstrom, R*., et al.* (2013b). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell *154*, 888-903.

Stupar, R.M., Lilly, J.W., Town, C.D., Cheng, Z., Kaul, S., Buell, C.R., and Jiang, J. (2001). Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci U S A *98*, 5099-5103.

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*., et al.* (2014). Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep.

Taipale, M., Jarosz, D.F., and Lindquist, S. (2010). HSP90 at the hub of protein homeostasis: emerging mechanistic insights. Nat Rev Mol Cell Biol *11*, 515-528.

Takahashi, K.H., Okada, Y., and Teramura, K. (2011a). Genome-wide deficiency mapping of the regions responsible for temporal canalization of the developmental processes of Drosophila melanogaster. J Hered *102*, 448-457.

Takahashi, K.H., Okada, Y., Teramura, K., and Tsujino, M. (2011b). Deficiency mapping of the genomic regions associated with effects on developmental stability in Drosophila melanogaster. Evolution *65*, 3565-3577.

Talbot, D., Collis, P., Antoniou, M., Vidal, M., Grosveld, F., and Greaves, D.R. (1989). A Dominant Control Region from the Human Beta-Globin Locus Conferring Integration Site-Independent Gene-Expression. Nature *338*, 352-355.

Teves, S.S., and Henikoff, S. (2011). Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. Genes Dev *25*, 2387-2397.

The_Arabidopsis_Genome_Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. Current opinion in plant biology *4*, 75-85.

Theissen, G., and Melzer, R. (2007). Molecular Mechanisms Underlying Origin and Diversification of the Angiosperm Flower. Annals of Botany *100*, 603-619.

Thomas, G.H., and Elgin, S.C. (1988). Protein/DNA architecture of the DNase I hypersensitive region of the Drosophila hsp26 promoter. The EMBO Journal *7*, 2191-2201.

Thomas, S., Li, X.Y., Sabo, P.J., Sandstrom, R., Thurman, R.E., Canfield, T.K., Giste, E., Fisher, W., Hammonds, A., Celniker, S.E*., et al.* (2011). Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome biology *12*, R43.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B*., et al.* (2012a). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B*., et al.* (2012b). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell *10*, 1453-1465.

Tominaga-Wada, R., Iwata, M., Sugiyama, J., Kotake, T., Ishida, T., Yokoyama, R., Nishitani, K., Okada, K., and Wada, T. (2009). The GLABRA2 homeodomain protein directly regulates CESA5 and XTH17 gene expression in Arabidopsis roots. Plant J *60*, 564-574.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat Biotechnol *28*, 1089-1095.

van Zanten, M., Tessadori, F., Peeters, A.J.M., and Fransz, P. (2012). Shedding light on large-scale chromatin reorganization in Arabidopsis thaliana. Mol Plant *5*, 583-590.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inze, D., and De Veylder, L. (2005). Genome-wide identification of potential plant E2F target genes. Plant Physiol *139*, 316-328.

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A.C., Hilbert, J.-L., Bartel, D.P., and Crété, P. (2004). Endogenous trans-Acting siRNAs Regulate the Accumulation of Arabidopsis mRNAs. Molecular Cell *16*, 69-79.

Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J.A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat Methods *11*, 66-72.

Waddington, C.H. (1942a). Canalization of development and the inheritance of acquired characters. Nature *150*, 563-565.

Waddington, C.H. (1942b). Canalization of development and the inheritance of aquired characters. Nature, 563-565.

Wang, D.Y., Kumar, S., and Hedges, S.B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc Biol Sci *266*, 163-171.

Wang, M.Y., Zhao, P.M., Cheng, H.Q., Han, L.B., Wu, X.M., Gao, P., Wang, H.Y., Yang, C.L., Zhong, N.Q., Zuo, J.R*., et al.* (2013). The cotton transcription factor TCP14 functions in auxin-mediated epidermal cell differentiation and elongation. Plant Physiol *162*, 1669-1680.

Weigel, D., and Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. Genome Biol *10*, 107.

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

Weirauch, M., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H., Lambert, S., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J., Galli, M., Lewsey, M., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F., Ratsch, G., Larrondo, L., Ecker, J.R., Hughes, T. ((in press, September 11, 2014)). Determination and inference of eukaryotic transcription factor sequence specificity. Cell.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S*., et al.* (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol *31*, 126-134.

Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Biochem *52*, 25-73.

West, A.G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. Genes Dev *16*, 271-288.

Western, T.L., Skinner, D.J., and Haughn, G.W. (2000). Differentiation of mucilage secretory cells of the Arabidopsis seed coat. Plant Physiol *122*, 345-356.

Whitacre, J.M. (2012). Biological robustness: paradigms, mechanisms, and systems principles. Front Genet *3*, 67.

Windsor, J.B., Symonds, V.V., Mendenhall, J., and Lloyd, A.M. (2000). Arabidopsis seed coat development: morphological differentiation of the outer integument. Plant J *22*, 483-493.

Wollmann, H., Mica, E., Todesco, M., Long, J.A., and Weigel, D. (2010). On reconciling the interactions between APETALA2, miR172 and AGAMOUS with the ABC model of flower development. Development *137*, 3633-3642.

Wu, C. (1984). Two protein-binding sites in chromatin implicated in the activation of heat-shock genes. Nature *309*, 229-234.

Wu, C., M. Bingham, P., Livak, K.J., Holmgren, R., and Elgin, S.C.R. (1979a). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell *16*, 797-806.

Wu, C., Wong, Y.-C., and Elgin, S.C.R. (1979b). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell *16*, 807-814.

Yadav, V., Mallappa, C., Gangappa, S.N., Bhatia, S., and Chattopadhyay, S. (2005). A basic helix-loop-helix transcription factor in Arabidopsis, MYC2, acts as a repressor of blue light-mediated photomorphogenic growth. Plant Cell *17*, 1953-1966.

Yamamoto, Y.Y., Ichida, H., Abe, T., Suzuki, Y., Sugano, S., and Obokata, J. (2007). Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. Nucleic Acids Res *35*, 6219-6226.

Yang, X., Zhang, W., Dong, M., Boubriak, I., and Huang, Z. (2011). The achene mucilage hydrated in desert dew assists seed cells in maintaining DNA integrity: adaptive strategy of desert plant Artemisia sphaerocephala. PLoS One *6*, e24346.

Yang, X.J., Dong, M., and Huang, Z.Y. (2010). Role of mucilage in the germination of Artemisia sphaerocephala (Asteraceae) achenes exposed to osmotic stress and salinity. Plant Physiol Bioch *48*, 131-135.

Yeyati, P.L., Bancewicz, R.M., Maule, J., and van Heyningen, V. (2007). Hsp90 selectively modulates phenotype in vertebrate development. PLoS Genet *3*, e43.

Zeller, G., Clark, R.M., Schneeberger, K., Bohlen, A., Weigel, D., and Ratsch, G. (2008). Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Research *18*, 918-929.

Zhan, S., Horrocks, J., and Lukens, L.N. (2006). Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. Plant J *45*, 347-357.

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

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

Zhang, W.L., Zhang, T., Wu, Y.F., and Jiang, J.M. (2014). Open Chromatin in Plant Genomes. Cytogenet Genome Res *143*, 18-27.

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, S., Zhao, M., Shi, T., Shi, H., An, F., Zhao, Q., and Guo, H. (2009). EIN3/EIL1 cooperate with PIF1 to prevent photo-oxidation and to promote greening of Arabidopsis seedlings. P Natl Acad Sci USA *106*, 21431-21436.

Zhong, X., Hale, C.J., Law, J.A., Johnson, L.M., Feng, S., Tu, A., and Jacobsen, S.E. (2012). DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. Nat Struct Mol Biol *19*, 870-875.