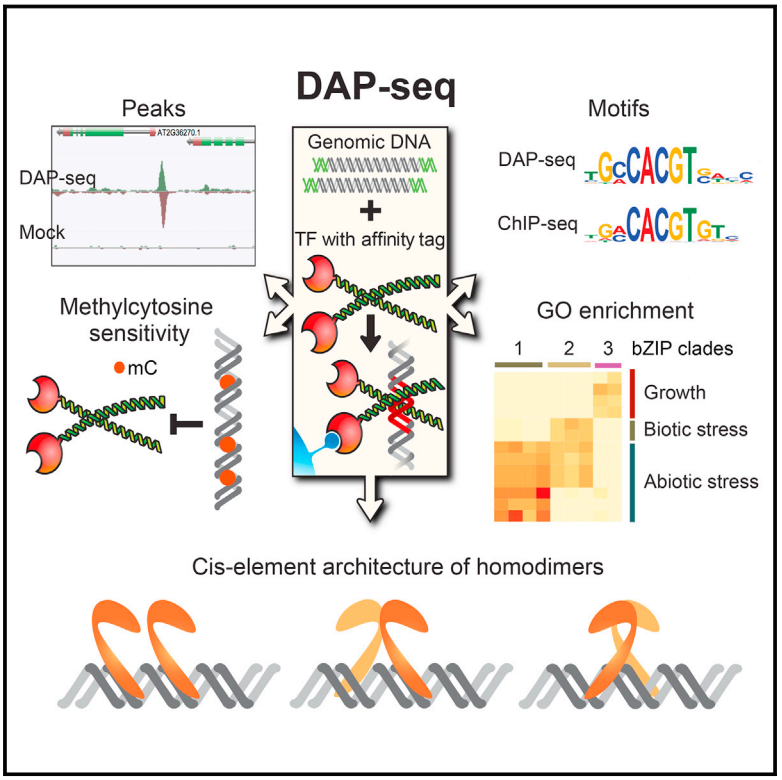
Resource

Cell

Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape

Graphical Abstract



Authors

Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, ..., Mary Galli, Andrea Gallavotti, Joseph R. Ecker

Correspondence

ecker@salk.edu

In Brief

A new method for pinpointing transcription factor binding sites in the *Arabidopsis* genome and their responsiveness to DNA methylation demonstrates the impact of tissue-specific DNA chemical modifications on gene regulation, potentially for any organism.

Highlights

- 2.7 million binding targets for hundreds of TFs define the *Arabidopsis* cistrome
- Methylation sensitivities of 76% of TFs surveyed shape the *Arabidopsis* epicistrome
- Strong enrichment of relevant gene functions is predicted for TF target genes
- Auxin response factor motif architecture promotes cooperative binding

Accession Numbers

GSE60143

O'Malley et al., 2016, Cell *165*, 1280–1292
May 19, 2016
http://dx.doi.org/10.1016/j.cell.2016.04.038

CellPress

Cell

Resource

Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape

Ronan C. O'Malley,[1,2,5] Shao-shan Carol Huang,[1,2,5] Liang Song,[2] Mathew G. Lewsey,[2,6] Anna Bartlett,[1] Joseph R. Nery,[1] Mary Galli,[1,4] Andrea Gallavotti,[4] and Joseph R. Ecker[1,2,3,*]

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd., La Jolla, CA 92037, USA
[2]Plant Biology Laboratory, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd., La Jolla, CA 92037, USA
[3]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd., La Jolla, CA 92037, USA
[4]Waksman Institute, Rutgers University, Piscataway, NJ 08854-8020, USA
[5]Co-first author
[6]Present address: Centre for AgriBioscience, School of Life Science, Department of Animal, Plant, and Soil Science, La Trobe University, Bundoora, VIC 3086, Australia
*Correspondence: ecker@salk.edu
http://dx.doi.org/10.1016/j.cell.2016.04.038

SUMMARY

The cistrome is the complete set of transcription factor (TF) binding sites (*cis*-elements) in an organism, while an epicistrome incorporates tissue-specific DNA chemical modifications and TF-specific chemical sensitivities into these binding profiles. Robust methods to construct comprehensive cistrome and epicistrome maps are critical for elucidating complex transcriptional networks that underlie growth, behavior, and disease. Here, we describe DNA affinity purification sequencing (DAP-seq), a high-throughput TF binding site discovery method that interrogates genomic DNA with in-vitro-expressed TFs. Using DAP-seq, we defined the *Arabidopsis* cistrome by resolving motifs and peaks for 529 TFs. Because genomic DNA used in DAP-seq retains 5-methylcytosines, we determined that >75% (248/327) of *Arabidopsis* TFs surveyed were methylation sensitive, a property that strongly impacts the epicistrome landscape. DAP-seq datasets also yielded insight into the biology and binding site architecture of numerous TFs, demonstrating the value of DAP-seq for cost-effective cistromic and epicistromic annotation in any organism.

INTRODUCTION

Comprehensive identification of transcription factor binding sites (TFBS) in a genome, the cistrome, is essential for characterizing regulatory elements and TF function. Chromatin immunoprecipitation sequencing (ChIP-seq) is a powerful approach for TFBS discovery (Kheradpour and Kellis, 2014; Stamatoyannopoulos et al., 2012). However, ChIP-seq experiments have been generally limited in scale as they are difficult to execute, dependent on antibody quality, and challenging for rare or lowly expressed proteins (Kidder et al., 2011). As a result, binding site information is available for relatively few TFs and substantial TFBS coverage is only available for humans and several model organisms. Methods such as DNase hypersensitivity (DHS) assay or ATAC-seq offer more facile approaches for annotating genome-wide regulatory elements across many organisms and cell types (Buenrostro et al., 2015; Sullivan et al., 2014; Thurman et al., 2012). However, without comprehensive knowledge of TF sequence specificity, the targeting TFs of the identified regions cannot be readily verified.

In contrast to ChIP-seq, in vitro mapping of TFBS provides a scalable alternative to rapidly and inexpensively interrogate large numbers of TFs. The two most commonly used in vitro methods are systematic evolution of ligands by exponential enrichment (SELEX) (Jolma et al., 2010) and protein binding microarrays (PBM) (Berger and Bulyk, 2009). In both methods synthetic DNA oligomers are enriched with an affinity-tagged TF and the preferred binding sequences are used to derive binding motifs. Both methods can resolve a large number of TF motifs, which can then be used to predict TFBS genome-wide. However, these assays employ synthetic DNA that lacks genomic DNA properties known to impact TF binding, including primary sequence context and chemical modifications, such as the widespread and tissue-specific 5-methylcytosine found in plants and animals. Efforts have been made to build synthetic oligomer pools that reflect relevant *cis*-element sequence (Levo and Segal, 2014) or incorporate methylation (Mann et al., 2013), but complex variation in nucleotide sequence and DNA methylation patterns (Schmitz et al., 2013) makes it extremely challenging to fully reproduce native nuclear DNA patterns by synthesis.

Genomic DNA (gDNA) is the native substrate for a TF and therefore ideal for an in vitro TF interaction assay. Unlike synthetic oligomers, gDNA encodes primary sequence and cell-, tissue-, and organism-specific methylation patterns that may impact TF binding. Moreover, as gDNA from different tissue/cell types and species can be easily obtained, the impact of sequence and methylation variation can be experimentally determined. Previous TF:DNA binding assays using naked gDNA were effective in identifying motifs and in vivo binding sites (Guertin et al., 2012; Liu et al., 2005; Rajeev et al., 2014), but this approach has not been applied for global TFBS characterization or to investigate the impact of primary sequence and DNA methylation on in vivo TF binding.

We developed DNA affinity purification sequencing (DAP-seq), a high-throughput assay that uses in-vitro-expressed TF to interrogate naked gDNA fragments to establish binding locations (peaks) and sequence motifs. We demonstrated the ultra-high-throughput capability of the assay by creating a cistrome map for *Arabidopsis thaliana*, consisting of peaks and motifs for 529 (30%) *Arabidopsis* TFs. These datasets include 2.7 million experimentally determined genomic-context TFBS covering 11 Mb (9.3%) of the genome, predicting thousands of target genes enriched in known and new functions. Comparison of DAP-seq and ChIP-seq datasets showed that DAP-seq peaks predicted in vivo TF binding better than motif inference. This improved predictive power can be partially explained by the ability of the assay to directly capture the impact of primary sequence and DNA methylation on binding affinities at individual TFBS. Globally, 76% of *Arabidopsis* TFs surveyed were sensitive to methylation in their motifs. By testing gDNA libraries in which methylcytosines were removed by PCR (ampDAP-seq), we identified ~180,000 TFBS occluded by leaf DNA methylation (the *Arabidopsis* epicistrome). Finally, we showed that closely spaced motifs significantly affected TF binding by developing a model for cooperative auxin response factor (ARF) homodimer binding to complex motif repeats. In total, ~2,300 individual DAP-seq experiments are reported, with all motifs, peaks, and TF-methylation sensitivities publicly available on our Plant Cistrome Database (http://neomorph.salk.edu/PlantCistromeDB).

RESULTS

DAP-Seq

DAP-seq is an in vitro TF-DNA binding assay that allows low-cost and rapid generation of genome-wide binding site maps for a large number of TFs, while capturing gDNA properties that impact binding in vivo. A DAP-seq gDNA library is prepared by attaching a short DNA sequencing adaptor onto purified and fragmented gDNA (Figure 1A; DAP library). In a separate reaction, an affinity-purified TF is prepared by in vitro expression, bound to ligand-coupled beads, and washed to remove non-specific cellular components (Figure 1B). The gDNA library is added to the affinity-bound TF and the unbound DNA is washed away (Figure 1C). The bound fraction is eluted, amplified with PCR primers to introduce an indexed adaptor, and the DNA is sequenced. By mapping the reads to a reference genome, enriched loci (peaks) can be used to identify TFBS and motifs. For example, inspection of DAP-seq peaks for the bZIP TF ABI5 revealed enrichment at a known regulatory site that contains two adjacent G-box motifs (CACGTG) (Xu et al., 2014), where a ChIP-seq peak was also found (Figure 1D). The DAP-seq-derived motif matched the motifs derived from both ChIP-seq and PBM (Weirauch et al., 2014), although the DAP- and ChIP-seq motifs shared more sequence similarity at the edges (Figure 1E).

To measure the impact of DNA modifications on TF binding, we implemented a modified version of DAP-seq, ampDAP-seq, which uses a DNA library in which the DNA modifications are removed by PCR (Figure 1A). Together with DNA chemical modification maps, i.e., base-resolution methylomes (Schmitz et al., 2013), the comparison of DAP-seq and ampDAP-seq data allows for a global assessment of the effects of DNA modifications on TF binding.

The *Arabidopsis* Cistrome

To create a comprehensive catalog of *Arabidopsis* motifs and genomic TF binding locations, DAP-seq experiments were carried out on 1,812 TFs comprising 80 families (Pruneda-Paz et al., 2014) (Tables S1A and S1B). Using a computational pipeline that identified highly enriched motifs from the strongest peaks (Supplemental Experimental Procedures; Machanick and Bailey, 2011; Guo et al., 2012), we characterized peaks for 1,055 TFs and derived motifs for 529 TFs. The dataset provided coverage for 52 of the 66 families with more than two members (Figure S1A) and identified a total of ~2.7 million TFBS covering 11 Mb (9%) of the genome. Reproducibility was high, with replicate correlations between 0.71 and 0.99 (Figure S1B). The entire set of motifs (Figure 2A) and peaks (Figure 2B), which we collectively term the *Arabidopsis* cistrome, can be viewed and downloaded (http://neomorph.salk.edu/PlantCistromeDB).

We investigated properties of the assay (reproducibility and protein expression levels) and TF family features that may predict the failure or success for a particular TF (Supplemental Experimental Procedures). Overall, technical issues could explain ~10% of failures, and thus some TFs produced peak datasets in retesting (Table S2). Generally, the rescue rate of failed TFs in a retest was related to the overall success rate of the family. For example, retesting 87 failed MADS-box TFs did not produce a single successful DAP-seq dataset, while recovery rates were higher than average in the more successful bZIP and NAC families (Table S2). This suggests that family-specific properties strongly affect the ability to produce a protein with DNA binding activity and may be influenced by protein stability in the assay conditions or a requirement for a protein partner, cofactor, or post-translational modification for activity.

Comparing DAP-seq-derived motifs to curated motif databases (Transfac, JASPAR, and AGRIS), we found most DAP-seq motifs were highly similar to published data (Table S1C). For TFs that were also present in two large-scale *Arabidopsis* PBM datasets (118 from CIS-BP [Weirauch et al. 2014] and 24 from PBM [Franco-Zorrilla et al. 2014]; Figure 2C; Table S1D), we found quantitative agreement between the DAP-seq and PBM-derived PWM (Figure S1C), although the DAP-seq PWM contained a higher number of informative positions (information content $\geq$ 0.8 bits; Figure 2D; 4.8 bp for CIS-BP versus 6.8 bp for DAP-seq), and predicted many fewer TFBS (Figure 2E; 122,200 for CIS-BP versus 11,900 for DAP-seq). From DAP-seq and ChIP-seq comparisons of TFs from three different families, the average number of TFBS identified by DAP-seq peaks was similar to the average number of in vivo binding sites recovered (12,352 in DAP-seq versus 8,372 in ChIP-seq; see "DAP-seq Captures TF Binding Sites Identified by ChIP-seq").

To investigate overall motif relationships, we clustered the PWM of the 529 TFs and observed that related paralogs targeted similar motifs (Figure S2A). Applying a dynamic tree cut (Langfelder et al., 2008) to the clustering dendrogram, we identified 85 motif types (Figure S2A). At the family level, motif clusters

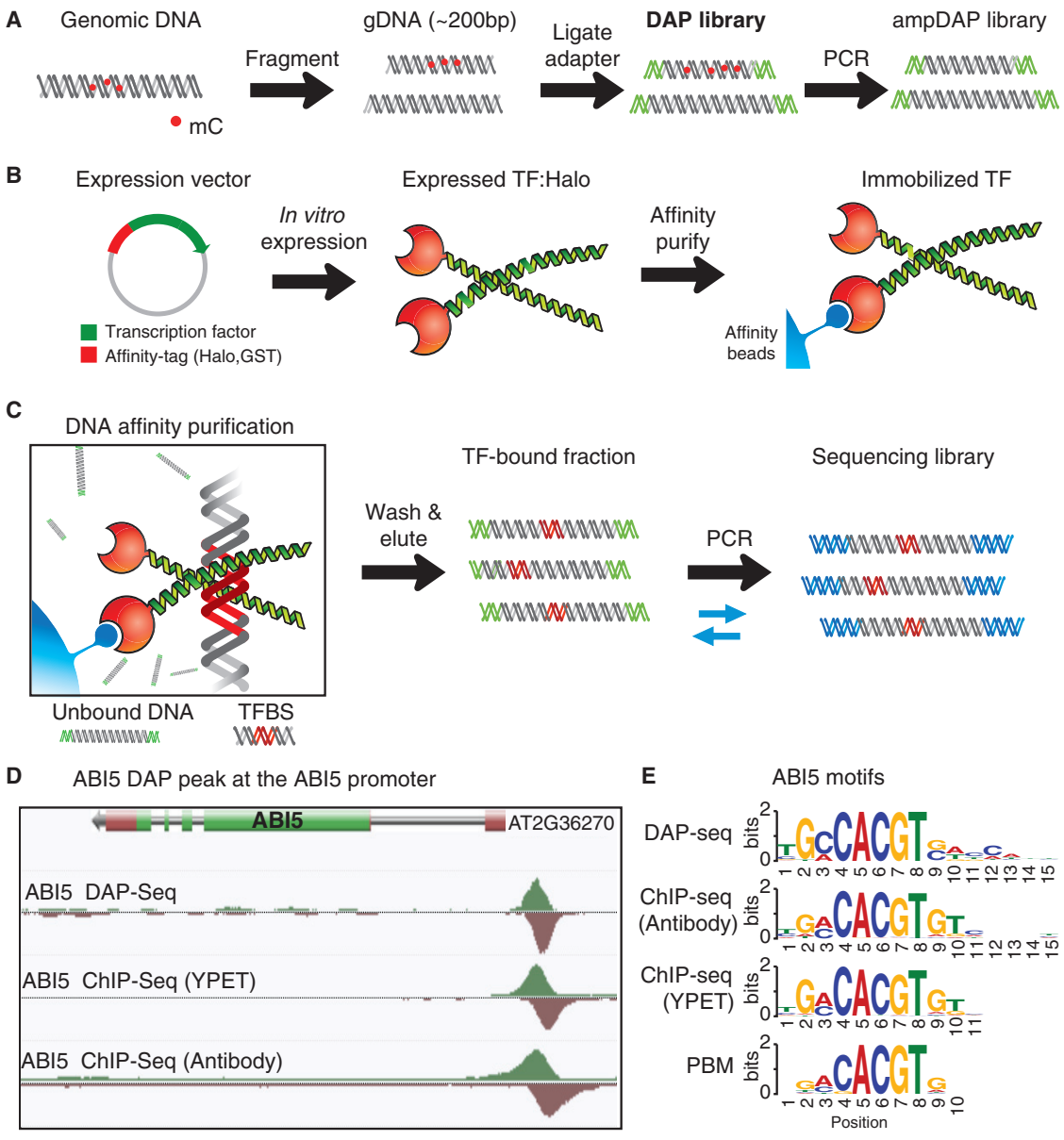


Figure 1. Genome-wide TFBS Discovery by DAP-Seq

(A) Preparation of DAP- and ampDAP-seq libraries.

(B) Expression and capture of affinity-tagged TFs.

(C) gDNA is bound to immobilized TFs, eluted, and sequenced.

(D) ABI5 DAP- and ChIP-seq peaks at a known regulatory element in the ABI5 promoter.

(E) Motifs derived from DAP- and ChIP-seq match a published ABI5 motif (Weirauch et al., 2014).

See also Table S2.

from the large and functionally diverse bZIP (Figure 3A) and NAC families (Figure S2B) closely reflected TF phylogeny (Corrêa et al., 2008; Olsen et al., 2005), indicating target sequences are conserved for close paralogs. Binding peaks of these TFs showed a range of enrichment in conserved non-coding regions (Haudry et al., 2013) (Figure S2C). Although the 529 DAP-seq motifs provided a global description of motif types, it was biased toward larger and more tractable families, such as bZIP, NAC, and WRKY, while some families, such as MADS and C3H, were underrepresented (Figure S1A). For a more balanced analysis, a subset of 57 TFs were identified (Figure 3B) that spanned the space of motif diversity (Figure 3C) and captured about 50% of motif types (Figure S2D). They were also selected based on published literature regarding consensus motifs and functions to highlight the known and new properties predicted by DAP-seq (Table S1C).

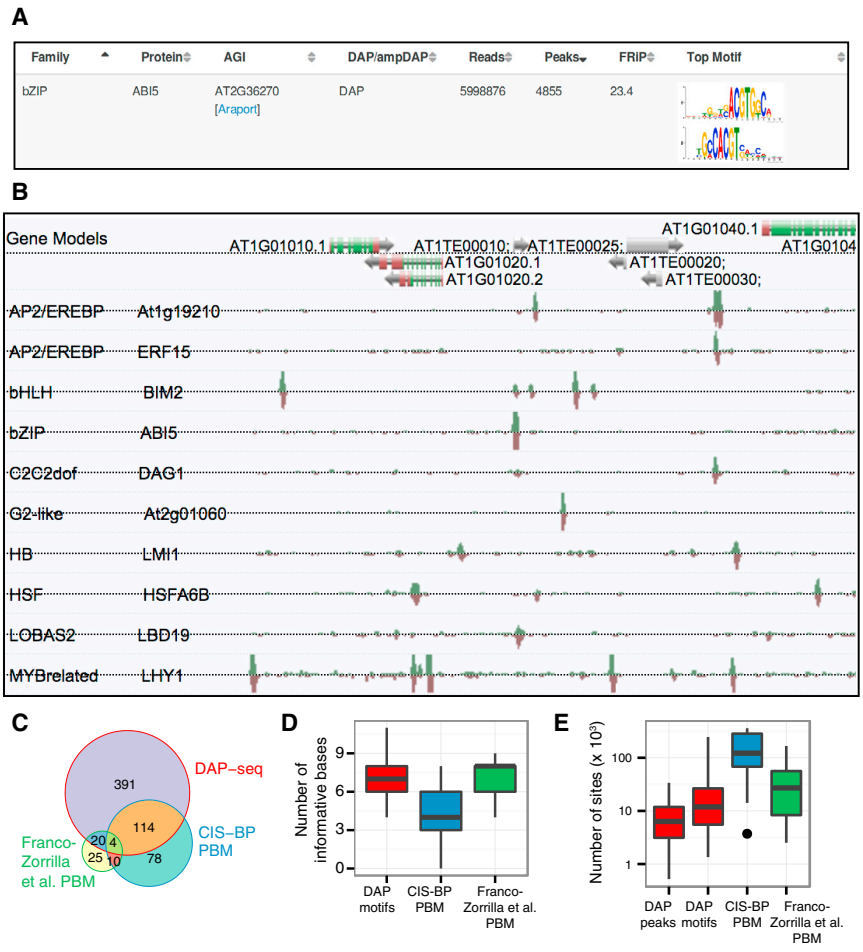


Figure 2. A Genome-wide Atlas of *Arabidopsis* TFBS Motifs and Binding Locations

(A) Web portal of TF binding motifs from 529 DAP-seq and 343 ampDAP-seq experiments.

(B) Sample screen shot of genome browser with DAP-seq peaks for selected TFs.

(C) Overlap between TFs from the DAP-seq, CIS-BP PBM (Weirauch et al., 2014), and PBM datasets (Franco-Zorrilla et al., 2014).

(D) Number of informative bases (information content $\geq$ 0.8 bits) in DAP-seq and PBM motifs.

(E) Number of TFBS predicted by peaks (DAP-seq) or motifs (DAP-seq, CIS-BP, and PBM [Franco-Zorrilla et al., 2014]).

See also Figure S1 and Tables S1 and S2.

Several new motif types identified in the DAP-seq dataset included members of the C2H2, GRF, and AP2-EREBP family. The discovery of a long poly-A motif for VRN1 and REM19, closely related ABI3-VP1 paralogs, was surprising as previous electrophoretic mobility shift assay experiments found no DNA sequence preference for VRN1, although this was likely because a poly-A oligomer was not tested (Berke and Snel, 2015). Notably, the motif captured for VRN1 (29 bp) was twice as long as REM19 (15 bp), which could be explained by the presence of tandem B3 DNA-binding domains in VRN1 compared to only one copy in REM19. VRN1 and REM19 are master regulators of cold-induced flowering and were recently proposed to be components of the plant Polycomb Repressive Complex PRC1 (Berke and Snel, 2015), suggesting VRN1/REM19 may target the PRC1 to poly-A motifs to repress flowering.

To better understand the genome-wide binding profiles of the different TF families, we computed the enrichment/depletion of binding sites of the 57 representative TFs relative to gene features (Figure S6A) and observed overall distributions similar to those identified by PBM (Weirauch et al., 2014). While substantial positional heterogeneity existed, there was global preference across TF families for enrichment at promoters and 5′ UTR and moderate depletion in coding regions. Enrichment/depletion at long non-coding RNA promoters was weaker and showed patterns different from protein coding genes, suggesting distinct modes of regulation.

Target genes predicted for the 57 representative TFs were strongly enriched ($1 \times 10^{-4} < p < 1 \times 10^{-64}$) in gene ontology (GO) terms that agreed with known functions and indicated potential new functions (Figure S3; Table S1C). By removing generic and redundant GO terms, we could highlight a set of TFs whose target genes predicted functions that were pertinent to all organismal biology (Figure 4, asterisks; related citations in Table S1C). We noted the largest split in TF functions was between those involved in hormone and endogenous response pathways (Figure 4, black bar) and those involved in intrinsic pathways (gray bar). Within this larger division, we observed six specific functional categories: hormone-regulated development (Figure 4, box 1) and growth (Figure 4, box 2), defense (Figure 4, box 3), cell division (Figure 4, box 4), metabolism and nutrition (Figure 4, box 5), and intrinsically regulated growth (Figure 4, box 6).

TFs enriched for GO terms related to hormone-regulated development (Figure 4, box 1) included two ARFs (ARF2 and MP/ARF5), master regulators of auxin hormone responses, and a Homeobox (HB) family TF (LMI1) also known to play a role in auxin responses. TFs enriched for innate immune response (Figure 4, box 3) included two master regulators of plant defense (WRKY40 and TGA5). Factors enriched in cell-cycle function (Figure 4, box 4) included the E2F family (E2FA and DEL2), direct regulators of DNA replication, and Growth-regulating Factor 6 (AtGRF6), which belongs to a family modulating cell-cycle progression and growth. The metabolism and nutrition category (Figure 4, box 5) contained very specific functions for several TFs that were consistent with the literature, such as the role of MYB61 in phenylpropanoid regulation and lignification. Finally, hormone (Figure 4, box 2) and intrinsic growth (Figure 4, box 6) both contained NAC TFs, important regulators of growth. These

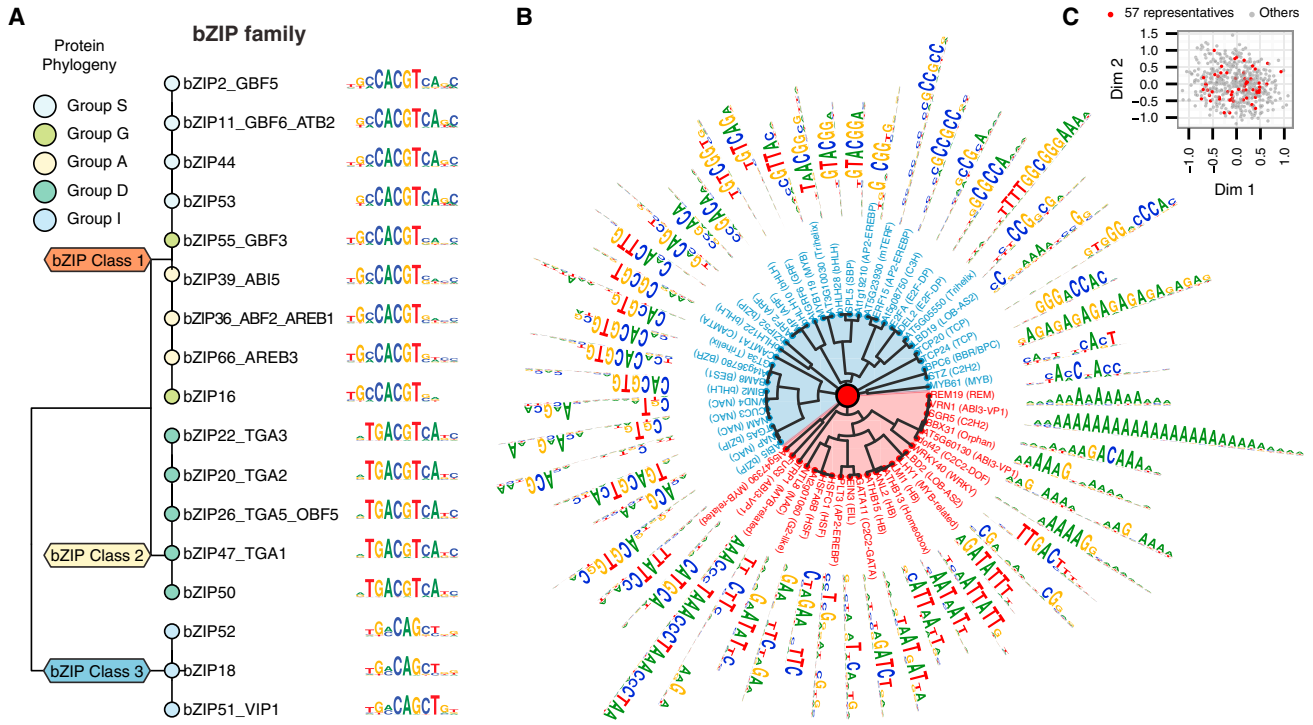


Figure 3. The Global Diversity of *Arabidopsis* TF Motifs

(A) bZIP family motifs from DAP-seq clustered by motif similarity.

(B) 57 TF motifs with GC-rich clusters in blue and AT-rich clusters in red.

(C) Multidimensional scaling plot of the full set of 529 TFs highlighting the 57 representative motifs.

See also Figure S2 and Table S1.

functions are consistent with known roles of NAP in hormone-regulated growth and defense, and VND4 in vascularization, an intrinsically regulated process. In summary, many of the predicted functions of the representative TFs are consistent with known functions.

New functions for many TFs were also predicted (Figure 4, arrows; related citations in Table S1C). For the heat-shock factor HSFA6B, we saw enrichment for high heat responses as expected, but also observed enrichment in mitotic functions (cell cycle and DNA replication; Figure 4, box 4). While plant HSFs have not been implicated in mitosis, a recent study of the human HSF1 indicates that this family may directly regulate cell division in proliferating cancer cells (Mendillo et al., 2012). For the C2H2 TF STZ, where mutants have enhanced tolerance to salt and abiotic stresses, we also saw enrichment in mitosis-related functions (Figure 4, box 4). As C2H2 family members from both plants and animals are known to regulate the cell cycle and DNA replication (Staudt et al., 2006; Welch et al., 2007), STZ enrichment for mitotic functions suggests that its stress response phenotype may involve direct regulation of cell division. Finally, bHLH122, known to be important for abiotic drought responses, targeted genes in immune processes (Figure 4, box 3), suggesting that it may also play a role in biotic defense. Overall, GO analysis of DAP-seq-derived target genes revealed TF functions consistent with the literature and identified new possible TF functions.

DAP-Seq Captures TF Binding Sites Identified by ChIP-Seq

To examine the relevance of in-vitro-derived DNA binding profiles compared to those from in vivo experiments, we performed ChIP-seq experiments for three *Arabidopsis* TFs from unrelated families: ABI5 (bZIP family), ATHB5 (HB family), and ANAC055 (NAC family). The bZIP family is found in all eukaryotes, while the NAC and HB families are plant specific. All three families have functions in plant hormone and growth regulation, although at different stages. The bZIP family in plants includes master regulators of salicylic and abscisic acid (ABA) hormone responses (Finkelstein and Lynch, 2000). ANAC055 is downstream of ABA and jasmonic acid signaling pathways and affects abiotic growth responses (Bu et al., 2008). HB family members play important roles in water stress and interact directly with auxin regulation (Ré et al., 2014). Three independent ChIP-seq experiments were performed on ABI5: two with an anti-ABI5 antibody in dark- and light-grown seedlings (ABI5 Ab etiolated and light) and one with an anti-GFP antibody in light grown seedlings containing a recombineered YPET-tagged ABI5 fusion protein (ABI5 YPET light). ChIP-seq for ANAC055 and ATHB5 used the same YPET-tagging strategy as the ABI5 YPET experiment.

Genome-wide comparison of the three TFs showed that DAP-seq peaks captured significant fractions of ChIP-seq peaks (36% to 81%; $p \leq 1 \times 10^{-5}$; Figure 5A, blue bars). Ranking ChIP-seq peaks by motif scores in the peak, we observed

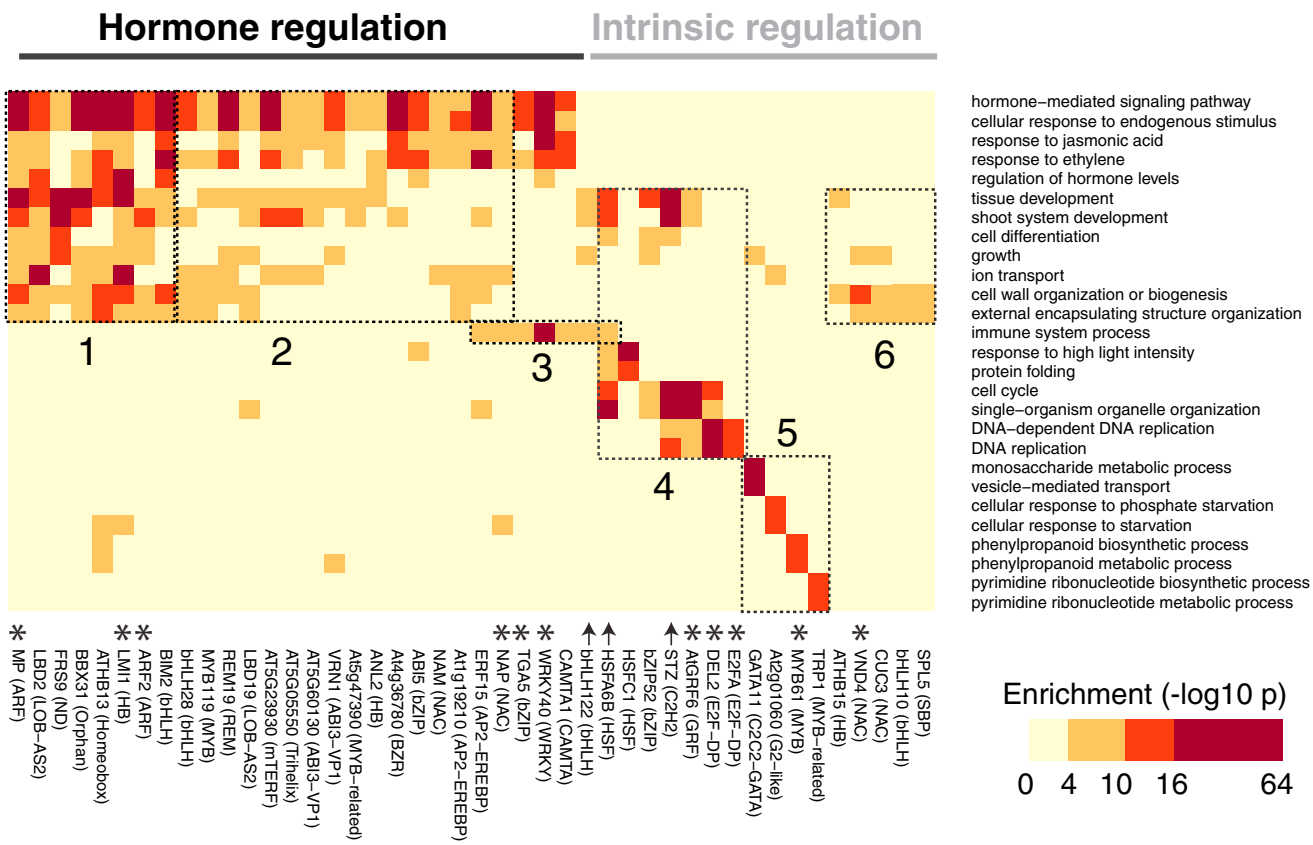


Figure 4. Critical Biological Processes Are Enriched in DAP-Seq Target Genes

Target genes predicted for 44 diverse TFs (subset of the 57 representatives) are enriched for functional terms associated with basic cellular properties. See also Figure S3 and Table S1.

increased overlap with DAP-seq peaks as motif score increased; 69% to 97% of ChIP-seq peaks that ranked in the top 25% by motif score overlapped with DAP-seq peaks (Figure 5A, red bars). This result suggests that DAP-seq preferentially captures in vivo binding sites associated with high scoring motifs. To confirm this, we compared the motif scores at peaks present in both ABI5 DAP-seq and ChIP-seq (DAP-ChIP) to those unique to one of the datasets (DAP-only and ChIP-only). Overall, we found that DAP-ChIP and DAP-only peaks contained high-scoring motifs, while the motif scores under ChIP-only peaks were only slightly elevated over background (Figure 5B). As a substantial fraction of ChIP-seq peaks do not contain a detectable target motif sequence, it was suggested that only a portion of the ChIP-seq peaks are from direct TF binding (Worsley Hunt and Wasserman, 2014). Our results indicate that DAP-seq may preferentially capture direct in vivo binding targets and thus can provide valuable binding affinity measurements at these sites.

Having identified that indirect binding may explain a large portion of the ChIP-only sites, we investigated whether chromatin properties could explain why certain strong binding sites detected by DAP-seq were not observed in ChIP-seq (DAP-only). Chromatin accessibility is known to influence TF binding affinities in vivo, and although this property cannot be directly measured by DAP-seq, integration with DHS datasets (Sullivan et al., 2014; Zhang et al., 2012) can provide information regarding in vivo site availability (Guertin et al., 2012). Within DHS regions 15% to 64% of DAP-seq peaks overlapped with a ChIP-seq peak (Figure 5C), significantly higher than the 5% to 28% of all DAP-seq peaks overlapping ChIP-seq peaks. As a single tissue captures only a subset of open chromatin states, we sequentially added DHS sites from four tissue types and found each tissue-specific DHS set overlapped with a unique set of DAP-seq peaks (Figure 5D). As these DHS experiments were performed on whole organs, many tissue-, cell-, and condition-specific DHS regions may still be unidentified, and the chromatin-free TF binding profiles from DAP-seq provide a valuable dataset for characterizing open chromatin regions.

We were interested to determine how well in vitro binding captured by DAP-seq peak signal could predict in vivo binding sites compared to conventional motif matching approaches. Using (1) PWM from published PBM, (2) PWM from DAP-seq, and (2) DAP-seq peak signal strength, we established ranked lists of binding sites inside DHS for comparison to the ABI5 YPET ChIP-seq experiment. We computed precision-recall metrics with increasing thresholds on each ranked list: precision is the fraction of predicted sites captured by ChIP-seq, and recall is

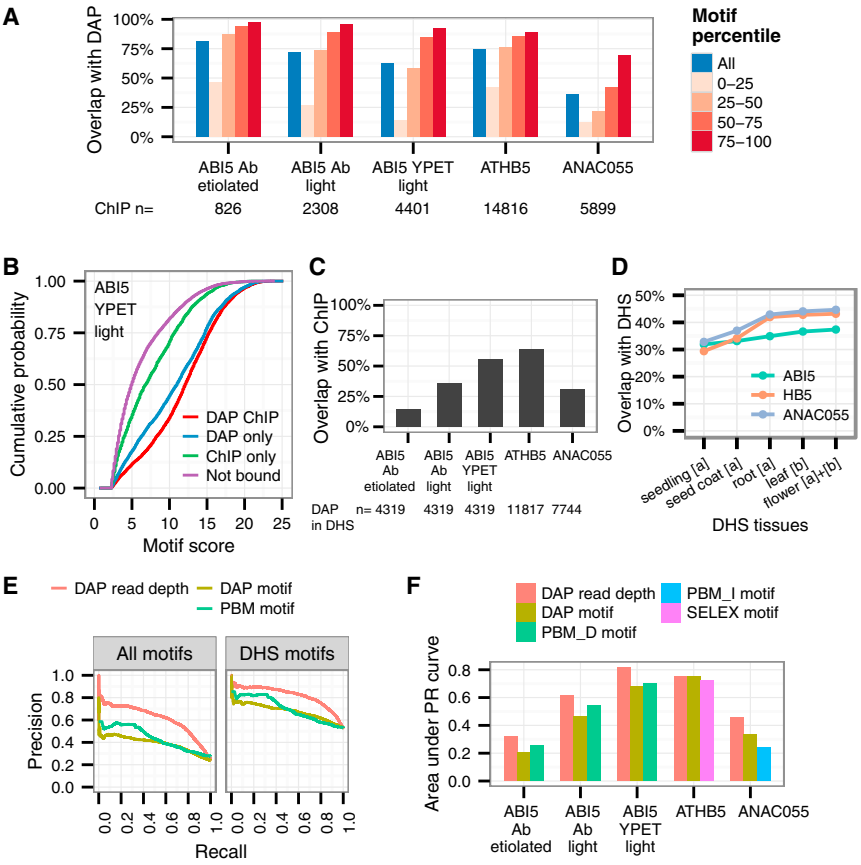


Figure 5. Concordance of In Vitro and In Vivo Binding Sites for Multiple TF Families

(A) Percent overlap of ChIP-seq peaks with DAP-seq peaks (blue), which increase for peaks associated with higher motif scores (red).

(B) Empirical cumulative distribution of motif scores shows shared ChIP- and DAP-seq peaks (DAP-ChIP) contain higher scoring motifs than do ChIP-only peaks, in which motif scores are similar to the motifs not bound in either assay.

(C) Percent DAP-seq peaks in DHS that overlap with ChIP-seq peaks.

(D) Using DHS data from multiple sources ([a] Sullivan et al., 2014; [b] Zhang et al., 2012) increases coverage of DAP-seq peaks.

(E) Precision (y axis) and recall (x axis) curve shows DAP-seq read depth (signal) predicts in vivo ABI5 binding sites better than mapping DAP-seq and PBM-derived motifs to genome, for all motifs (left) and motifs in DHS (right).

(F) By area under the precision-recall (PR) curve as in (E), all ChIP-seq datasets are most accurately predicted by DAP-seq read depth. PBM_D, motif directly determined by PBM. PBM_I, motif inferred by PBM based on DNA binding domain similarity. See also Figure S4.

the fraction of ChIP-seq bound sites captured by predicted sites. We found DAP-seq binding signal achieved 14%–17% higher precision than PWM matching (Figure 5E). For the other four ChIP-seq datasets, DAP-seq binding signal also outperformed PWM predictions, except for ATHB5 (Figure 5F). These results indicated that a direct biochemical interaction assay better predicted in vivo binding compared to motif inference, suggesting that the DAP-seq experiments measure the impact of genomic properties that influence TF binding in vivo.

We examined several primary sequence properties of genomic DNA known to impact in vivo binding, including motif clusters (Pott and Lieb, 2015) and TF sensitivity to DNA methylation in motifs (Domcke et al., 2015), by restricting predictions to peaks containing a single motif with no strong motifs within 100 bp, or to peaks containing motifs with no methylcytosine (Figure S4A). We observed improved performance of motif inference relative to DAP signal for ABI5, but not for ANAC055, suggesting these two TFs have different binding environment requirements and DAP-seq signal may achieve better predictive power by directly capturing environment features other than core recognition sequence.

To more thoroughly investigate this hypothesis, we constructed two random forest (RF) models using both motif and environment features. The first model used DAP signal as the motif feature, and the second used motif match score. Both included the same environment features for motif clusters, cytosine methylation in the motif, and predicted DNA shape parameters for sequences flanking the motifs (Zhou et al., 2015). As expected, adding environment features improved the accuracy for predicting in vivo binding for both types of motif features (Figure S4B), but the importance of the environment features was markedly different for each TF (Figure S4C). The motif score RF model for ANAC055 heavily relied on shape features, while the motif cluster feature and the motif methylation feature were more important for ABI5 YPET ChIP-seq. In contrast, these environment features were less important in DAP-seq signal RF models, suggesting DAP-seq natively captured the TF-specific effects of motif environment.

Cooperative Binding of ARF Homodimers at Phased Motif Repeats

Next, we explored TF-specific effects of motif clustering and how they impact the plant cistrome landscape. Even with the higher resolution of DAP-seq compared to ChIP-seq, for many TFs we observed strong binding at closely spaced motif clusters where multiple binding events were resolved as a single peak (Figures S5A and S5B). Although not surprising for TFs known to target repeat sequences (FRS9 and TRP1), this was also observed for many non-repeat binding TFs (ERF15, BIM2, and ABI5). Several TFs were at the opposite extreme, where strong DAP-seq peaks contained much less than one motif on average (STZ, NAP, ARF2, and MP/ARF5; Figure S5B). Unexpectedly, this group included two ARFs (ARF2 and MP/ARF5) with only 0.1–0.2 motifs per peak despite evidence that they bind to motif repeats in vitro and in vivo (Boer et al., 2014; Ulmasov et al., 1997). This suggests that direct examination is needed to

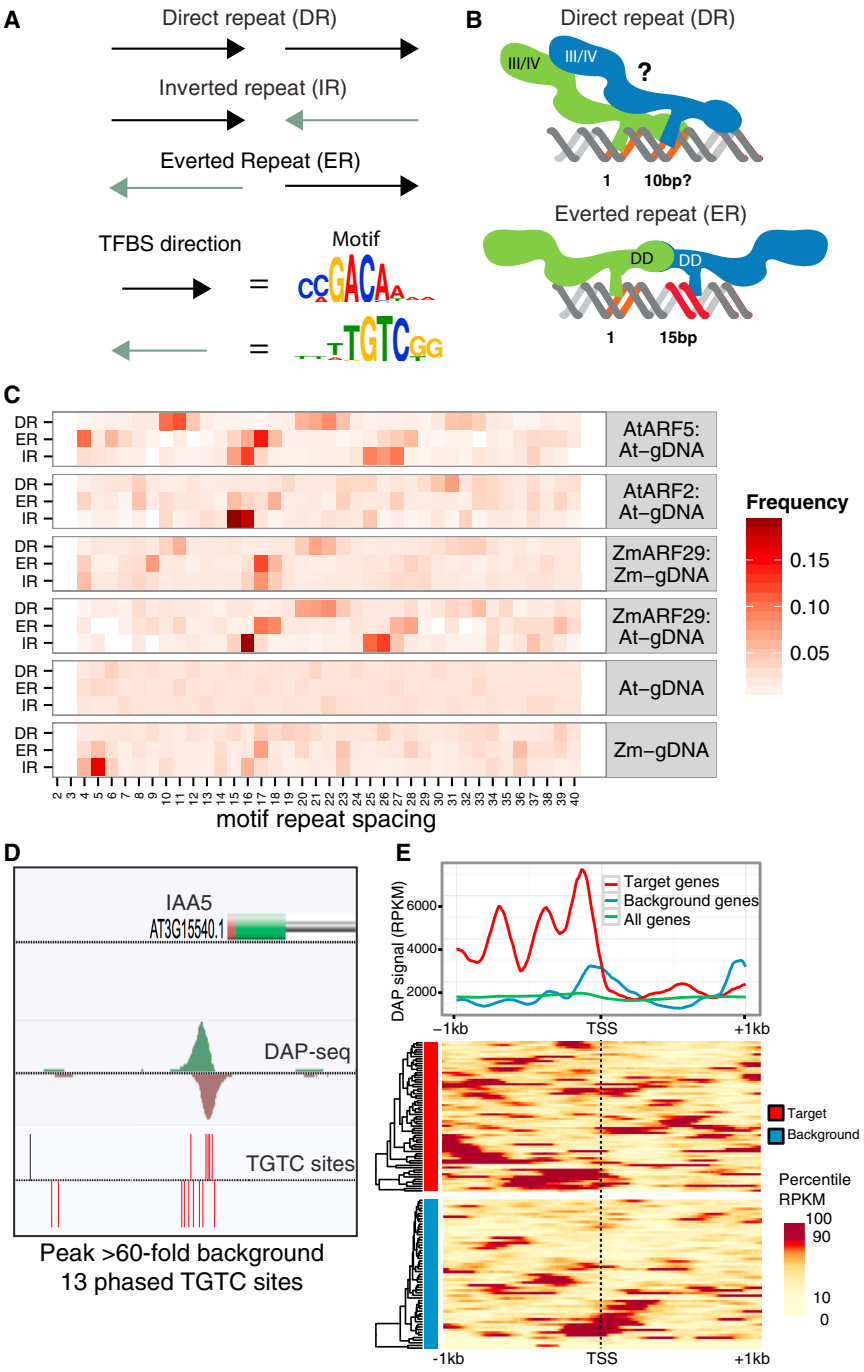


Figure 6. The ARF Family Preferentially Binds to Phased Motif Clusters that Are Enriched in Target Gene Promoters

(A) Three possible orientations of an ARF motif repeat.

(B) ARF homodimers could be stabilized at a DR by an interaction of the III/IV domain (top) (Nanao et al., 2014) and at an ER by the dimerization domain (bottom) (Boer et al., 2014).

(C) Relative frequencies of DAP-seq peaks at DR, ER, and IR pairs for *Arabidopsis* (AtARF5 and AtARF2) and maize (ARF5/ZmARF29) proteins interrogating *Arabidopsis* (At-gDNA) or maize (Zm-gDNA) DAP-seq libraries.

(D) A cluster of 13 phased TGTC sites (red ticks) in the promoter of the ARF5 target IAA5. Black ticks are non-phased TGTC sites.

(E) DAP-seq signal at the TSS (x axis) of ARF5 direct target genes, non-auxin-responsive background genes, and all genes.

See also Figure S5.

understand the more complex binding site architecture required for strong binding for some families. We explored this hypothesis in more depth focusing on the ARF family.

ARFs are important regulators of many basic plant processes, and multiple lines of evidence indicate that homodimerization, and possibly hetero- and multimerization, are important for ARF binding and function (Farcot et al., 2015). ARF DNA binding is known to strongly prefer direct (DR) and everted repeats (ER) of the well-characterized auxin response element (AuxRE: TGTCTC), although no binding at inverted repeats (IR) was reported (Figure 6A). The spacing between individual AuxREs is important as the DR was bound only when tested with spacing of 10–12 bp (i.e., between the two T's in bold **T**GTCTC-N_{4-6}-**T**GTCTC), and the ER AuxRE with spacing of 15–18 bp (**T**GTCGG-N_{6-9}-CC**G**ACA) (Ulmasov et al., 1997). A crystal structure of an ARF bound to an ER AuxRE showed that the 15–18 bp spacing allowed the bound ARF homodimer to stabilize through interaction of the ARF dimerization domain (DD) (Boer et al., 2014). Similarly, the DR spacing preference may be explained by stabilizing interactions through a second dimerization domain, the III/IV domain (Figure 6B) (Nanao et al., 2014). Importantly, although substantial evidence supports a role for motif dimers in ARF binding and auxin response regulation, no comprehensive model yet exists to explain or predict ARF binding site preferences (Farcot et al., 2015).

To refine the model of motif repeat orientation and spacing for ARF homodimer binding, we first identified genome-wide tandem motifs in the three repeat types above (Figure 6A). Since the **TGTC**GG motif reported by both DAP-seq and PBM was present in only ~30% of strong DAP-seq peaks and was distinct from the AuxRE sequence **TGTC**TC, we used the consensus sequence TGTC as our motif model. We extracted all instances of inverted, everted, and direct TGTC repeats in the genome (IR-TGTC, ER-TGTC, or DR-TGTC), recorded the distance between each pair in the repeat, and tabulated the number of strong DAP-seq peaks found at each repeat type as a function

of spacing (Figure 6C). For ARF5/MP, DR-TGTC binding preferentially occurred at three spacing groups: 10–12, 20–23, and 31–34 bp. For ER-TGTC, we observed three spacing groups at 4–8, 15–18, and 25–28 bp. Importantly, our results exactly matched the known spacing of 10–12 bp for DR (Ulmasov et al., 1997) and 15–18 bp for ER (Boer et al., 2014). We also identified novel binding events at IR repeats, which showed similar spacing preferences to those seen for ER-TGTC repeats but had only two spacing groups (15–16 and 25–27 bp). To explain homodimer binding at this third repeat type, we propose a model in which a third isoform of the ARF5 homodimer, with interactions between positively and negatively charged sides of the III/IV dimerization domain (Nanao et al., 2014), stabilizes the complex at specific spacing of the IR-TGTC (Figure S5C). To summarize, we observed three repeat-specific patterns of ARF binding that may be explained by three different ARF dimerization models. The multiple spacing groups for each repeat type and the flexibility within each group suggest that dimers can be stabilized by protein interactions spanning multiple helical turns as long as the interacting protein domains are in phase relative to the DNA helix.

Although the two functionally distinct ARF family members ARF2 and ARF5 had similar binding motifs (Figure 3B), their genome-wide binding correlation was only 0.09, much lower than the typical range of 0.6 to 0.8 for family members with very similar motifs such as those in the bZIP (Figure S5D) and NAC families (Figure S5E). Analysis of repeat spacing preferences for ARF2 revealed a more restricted pattern dominated by the IR-TGTC with a narrower range of flexibility within a spacing group compared to ARF5 (Figure 6C). Therefore, the low genome-wide binding correlation between ARF2 and ARF5 may be explained, in part, by the divergence of preferred spacing groups and motif repeat types, which, in turn, may be due to differences in the protein dimerization properties of the two phylogenetically distinct ARF proteins.

As the ARF family traces its origins back to the first land plants, we investigated whether the maize and *Arabidopsis* ARF binding properties have diverged in the 140–150 million years since their last common ancestor (Finet et al., 2013). Testing a maize co-ortholog of ARF5 (ZmARF29; Galli et al., 2015) on maize gDNA (Zm-gDNA) by DAP-seq, we observed similar, but not identical, motif spacing preferences with two dominant spacing groups in maize compared to the eight more distributed groups in *Arabidopsis* (Figure 6C). To determine if the ZmARF29 protein or the maize gDNA influenced the spacing differences, we assayed ZmARF29 using *Arabidopsis* gDNA (At-gDNA). The resultant ZmARF29:At-gDNA pattern was more similar to maize than to *Arabidopsis*, indicating that the spacing divergence is primarily due to ARF5 dimerization properties. Together, the ARF2/ARF5 and maize/*Arabidopsis* comparisons illustrate how natural variation of homodimer interactions can impact TF binding properties and thus the global TFBS landscape. The ZmARF29 experiments also demonstrate that the DAP-seq assay works in a large, repeat-rich genome (~2.5 Gb), similar in size to mammalian genomes.

To evaluate the in vivo relevance of our spacing model, we identified a set of 69 ARF5 target genes that rapidly respond to ARF5-specific repression with IAA19/BODENLOS and auxin treatments (Schlereth et al., 2010). 64% of these ARF5 targets contained a DAP-seq peak in their promoter, 3-fold enrichment over expectation (Figure S5F; $p < 1 \times 10^{-10}$). For example, the promoter of the ARF5 target IAA5 contained 13 phased TGTC sites in a ~400 bp DAP-seq peak that showed 60-fold enrichment over background (Figure 6D). We plotted the average DAP-seq read depth in 2-kb regions centered on the TSS of the 69 target genes and 62 non-auxin-responsive genes (Supplemental Experimental Procedures) and observed very strong binding primarily in the target gene promoters. Strikingly, there was a strong phased signal with a period of ~300 bp beginning ~150 bp upstream of the TSS (Figure 6E). Although DAP-seq was carried out on naked gDNA, the phasing pattern of ARF5 binding in target gene promoters resembled in vivo nucleosome phasing patterns found in active eukaryotic gene promoters (+1, −1 nucleosome, etc., locations) (Struhl and Segal, 2013). This suggests that the rapid responses of these ARF target genes may be due, in part, to high ARF occupancy relative to the preferred nucleosome positions characteristic of an active promoter.

In summary, our results support a model in which three flexible ARF homodimer isoforms bind to three distinct motif-repeat types spanning multiple helical turns, and that spacing preferences affect both ARF paralog and ortholog binding specificity. Moreover, enrichment of phased clustered repeats in the promoters of ARF5 target genes suggests that promoter location of ARF5 regulatory elements may play an important role in regulation of auxin-responsive genes.

The Epicistrome

Arabidopsis thaliana leaf nuclear DNA contains 5-methylcytosine at ~11% of cytosines (Schmitz et al., 2013; see the Supplemental Experimental Procedures), an important epigenomic feature for gene silencing. Several examples demonstrate that TF DNA methylation-sensitivities can impact in vivo TF binding, but the global impact on the cistrome has not yet been established in any organism. To determine how DNA methylation affected binding, we used base-pair methylation maps from *Arabidopsis* leaf DNA (Schmitz et al., 2013) to quantify DAP-seq and ChIP-seq binding at high-scoring motifs that contained 5-methylcytosine. As plant DNA methylation is equally distributed between two mutually exclusive patterns (Cokus et al., 2008; Lister et al., 2008), we classified these motifs into two categories: (1) motifs in mC-all regions identified by dense methylation in all contexts (CHH, CHG, and CG, where H is A, C, or T) associated with silenced genes and transposons (Figure 7A, inset), and (2) motifs in mCG-only regions exclusively methylated in the CG context, more sparsely distributed, and enriched in expressed genes (Figure 7B, inset). As a control, we identified a set of motifs that neighbored a methylated region (within 200 bp), but themselves did not contain methylation.

By calculating the ratio of the ChIP-seq or DAP-seq binding strength (read depth) at methylated and unmethylated motifs, we observed strong binding inhibition for ABI5 both in vitro and in vivo (Figures 7A, 7B, and S6B). Across all TF families both mC-all and mCG-only methylation impacted binding (Figures 7A and 7B), although inhibition by mC-all methylation was more pronounced, possibly due to the higher methylation

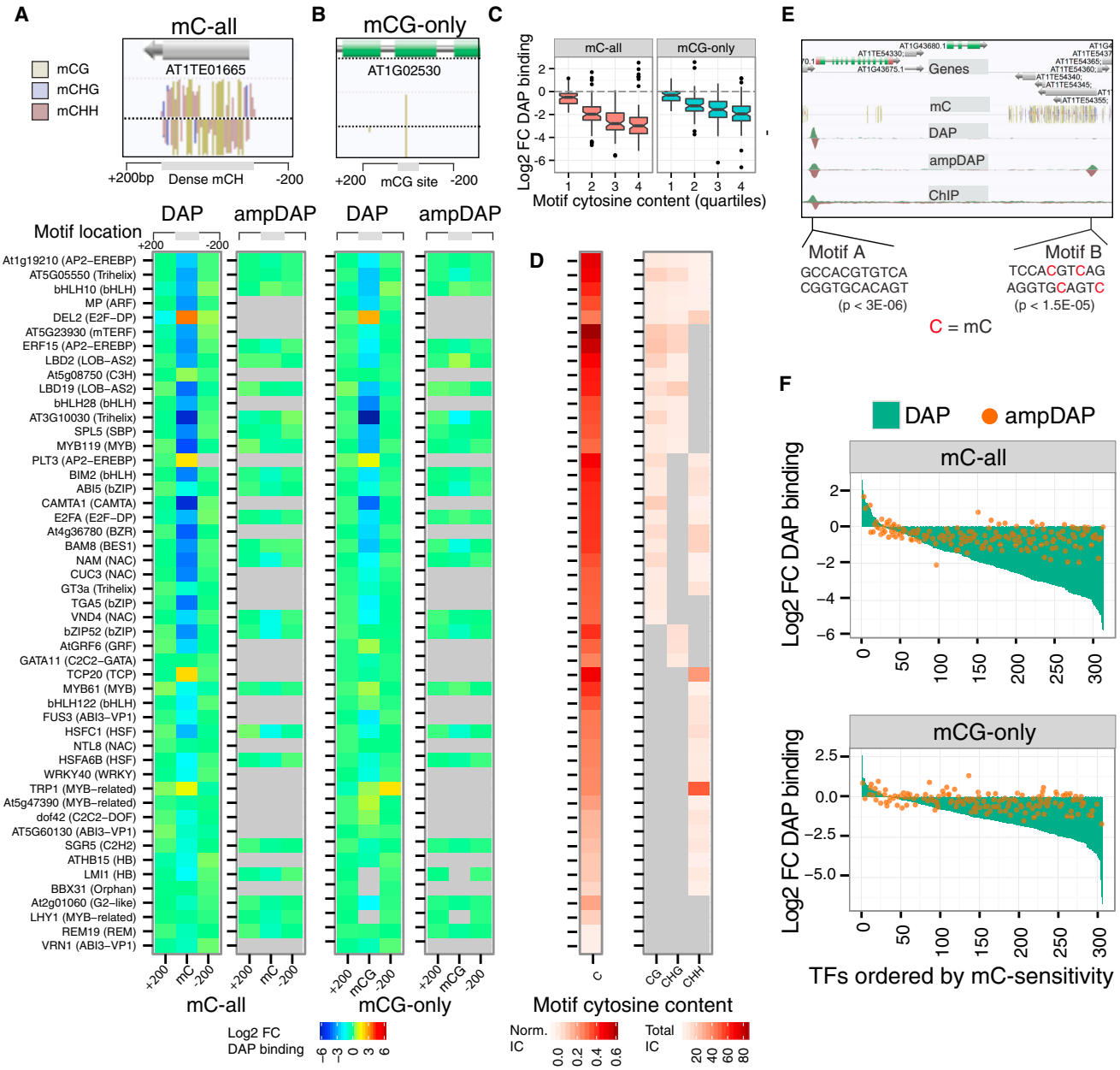


Figure 7. Motif Methylation Impacts Binding For 76% of TFs Surveyed

(A) Inset: mC-all regions contain dense methylation in all cytosine contexts. Left: binding fold change (FC) at motifs containing relative to motifs neighboring (within 200 bp) an mC-all site. Right: relative ampDAP-seq binding at the same motifs. Gray boxes indicate TFs with too few (<25) methylated motifs to score or a failed experiment.

(B) Inset: an isolated mCG-only site. Left: binding FC at motifs containing relative to motifs neighboring (within 200 bp) an mCG-only site. Right: relative ampDAP-seq binding at the same motifs.

(C) TF methylation sensitivity is correlated with cytosine content of the motif, defined as the informative content (IC) of cytosines, divided by total IC of the motif.

(D) Cytosine content (left) and informative CG, CHG, and CHH content for each motif (right) of TFs in (A and B).

(E) Genome browser showing DAP-, ampDAP-, and ChIP-seq peaks at methylated and unmethylated ABI5 motifs.

(F) Waterfall plot of log2 relative binding at methylated motifs for 349 TFs in DAP-seq and 219 TFs in ampDAP sets. In total, 248 of 327 TFs (76%) that had sufficient motif instances for quantification were found to be methylation-sensitive.

See also Figure S6 and Table S3.

density in these regions (Figure 7A; inset). For the entire set of 327 TFs that had sufficient motif instances for quantification, mC-all inhibition was seen for 72% (234) of TFs, weak to no binding inhibition for 24% (79), while 4.3% (14) preferentially bound methylated motifs (Figure 7F). Interestingly, E2F family member DEL2, with specific roles in DNA replication, preferentially bound

to methylated motifs, suggesting a possible relationship between this epigenetic mark and central regulators of cell division (Harashima et al., 2013).

To independently confirm the effect of methylation on TF binding, we used the modified DAP-seq assay, ampDAP-seq, where PCR replaces the 5-methylcytosines in the gDNA library with unmethylated cytosines (Figure 1A). ampDAP-seq of the 529 TFs resulted in motifs and peaks for 343 TFs. To ensure even comparison, we analyzed 219 TFs that had greater than 5% reads in peaks in both DAP- and ampDAP-seq (Figure S1D). DNA methylation sensitivities detected by DAP-seq were absent in the methylation-free ampDAP-seq datasets (Figure 7F), supporting our conclusion that 5-methylcytosine (or, although less likely, another chemical modification) was responsible for the observed TF binding changes (Figures 7A, 7B, and 7F; Table S3). Importantly, our ampDAP-seq data also provided the methylation-free binding strength of 178,135 TFBS normally occluded by leaf methylation, the *Arabidopsis* epicistrome.

We found that the cytosine content of a TF's PWM correlated with its binding sensitivity to 5-methylcytosine (Figures 7C and 7D), with a few exceptions, such as TCP20, MYB61, and bHLH122, suggesting the relationships for some TFs between motif cytosine content and methylation sensitivity are more complex. The *Arabidopsis* methylome is established by distinct DNA methyltransferases, each with a preference for one of three cytosine contexts CG, CHG, and CHH (Law and Jacobsen, 2010; Zemach et al., 2013). Comparing the CG, CHG, and CHH content in the motifs (Figure 7D) to the methylation sensitivities revealed three general rules: (1) TFs with strong CG or CHG in their motifs were strongly inhibited in both mC-all and mCG-only regions, (2) TFs with only CHH in their PWM were generally insensitive to methylation, and (3) motifs containing multiple cytosine contexts typically showed very high methylation inhibition. These general rules suggest that regulatory relationships could potentially exist between specific DNA methyltransferases and TF families.

One possible mechanism for the role of DNA methylation in gene and transposon silencing is through exclusion of TF binding at regulatory sites. Consistent with this model, the loss of methylation in DNA methyltransferase mutants results in increased expression of thousands of transposons and genes (Zhang et al., 2006). However, since cytosine methylation is also required for targeting of silencing-related chromatin modifications (Law and Jacobsen, 2010), it is difficult to delineate the contributions of individual epigenomic features to gene silencing in vivo. In this regard, the less dense mCG-only methylation provides a valuable complement for analyzing the effects of methylation on binding in vivo since it has not been associated with silencing. We compared the degree of binding reduction between in vitro DAP-seq and in vivo ChIP-seq using the ABI5 datasets. We observed examples of strong ampDAP-seq peaks at high-scoring ABI5 motifs with no equivalent peaks for either DAP-seq or ChIP-seq in both mC-all (Figure 7E) and mCG-only regions (Figure S6C). The extent of reduced binding genome-wide at methylated motifs was similar in the DAP-seq and ChIP-seq datasets at both mC-all and mCG-only sites (Figures 7A, 7B, and S6B). While these observations do not directly demonstrate that motif methylation contributes in vivo to TF exclusion, our findings are consistent with this model.

DISCUSSION

The in vivo protein-DNA interaction landscape is affected by multiple factors including primary sequence, DNA modifications, and chromatin accessibility, along with stabilizing and destabilizing interactions between proteins associated with the DNA (Lelli et al., 2012; Levo and Segal, 2014). Our in vitro DAP-seq assay offers a simple method to examine TF binding to its cognate target (gDNA) in a chromatin-free context, while maintaining important information related to primary genome sequence and DNA methylation. The assay's high-throughput capability allowed us to create a comprehensive atlas of the *Arabidopsis* cistrome consisting of 529 TFs targeting 2.7 million binding sites. Furthermore, by integrating DAP-seq TFBS, methylome maps, and direct measurements of binding in the absence of methylation in ampDAP-seq, we have performed the largest analysis to date for evaluating the relationship between TFs and methylated DNA. ampDAP-seq of 219 TFs identified ~180,000 TFBS occluded by leaf DNA methylation, characterizing an *Arabidopsis* epicistrome atlas. The precise base at which methylation affects binding cannot be easily isolated in mC-all regions as multiple 5-methylcytosines are often found both in and proximal to a motif. However, the same trends of binding changes were observed at the sparsely methylated mCG-only sites, and these binding changes correlated with motif cytosine content and context. Therefore, we propose that DNA methylation at high information positions in the motif may directly affect the interaction between TF and genomic DNA and contribute to the observed TF methylation sensitivity. Finally, by demonstrating the utility of these datasets to generate biological insights (GO enrichment and ARF motif architecture), we believe DAP-seq will be a powerful tool for understanding regulatory DNA functions in eukaryotic genomes. With the availability of hundreds of sequenced genomes and methylomes of wild accessions, these cistrome and epicistrome maps provide a valuable resource to evaluate the impact of natural genetic and epigenomic variation on transcriptional networks controlling plant adaptation.

Our analysis of ARF *cis*-element architecture shows how genome-wide DAP-seq datasets can be used to characterize regulatory sequence in a native genomic context. Evidence demonstrating preferential binding of ARFs to DR, ER, and IR supports a model in which three distinct ARF homodimer isoforms can form stable protein-protein interactions across multiple turns of the DNA helix. Considering that (1) ARF5 homodimers may be able to associate with DNA in three distinct isoforms, (2) multimeric binding sites are associated with very strong DAP-seq peaks, and (3) ARF5 direct target genes contain multimeric binding sites (e.g., 13 TGTCs in IAA5 promoter), we propose that ARF5 multimerization on genomic DNA could play a functional role in regulating auxin transcriptional response in vivo. Although the experiments presented here did not test nucleosome or TF heterodimer cooperativity, the method may be extended to test *cis*-element architecture associated with heterodimers and higher-order chromatin complexes. Such

assays will be useful for studying heterodimer binding properties important in the biological functions of the ARF and other TF families.

EXPERIMENTAL PROCEDURES

DAP-Seq and ChIP-Seq Experiments

For DAP-seq, gDNA was extracted from young *Arabidopsis* leaves, fragmented, and ligated with a truncated Illumina TruSeq adaptor. Separately, HALO-tagged TFs were expressed in an in vitro wheat germ system. HALO-TFs were immobilized on Magne HALO-Tag beads, washed, and incubated with the DNA library. After bead washing, DNA was eluted and amplified with indexed TruSeq primers. Sequencing was performed on an Illumina HiSeq 2500 with 100-bp SR reads. For ABI5, ANAC055, and HB5 ChIP-seq, the YPET or wild-type lines were germinated and grown for 36 hr under dark or long day light conditions. ChIP-seq was carried out as previously described with minor modifications (Chang et al., 2013).

The ampDAP-Seq DNA library was prepared by PCR amplification of a standard DAP-seq gDNA library using Phusion High-Fidelity DNA Polymerase (NEB; 15 ng of DNA in a 50 μl reaction) and the A and B adaptor oligos (25 μM each; Supplemental Experimental Procedures) with the cycling conditions below: 2 min at 95°C, 30 s at 98°C, 10 cycles of 15 s at 98°C, 30 s at 60°C, 2 min at 72°C, and a final extension time of 10 min at 72°C, followed by a hold at 4°C. The DNA was purified by Sera-Mag beads (Thermo) and resuspended in 30 μl elution buffer. Following the DAP binding protocol, the recovered DNA was PCR amplified for 20 cycles using the same conditions as DAP-seq using the full-length Illumina primers.

DAP-Seq and ChIP-Seq Data Processing

Reads were mapped to the TAIR10 genome for *Arabidopsis* and B73_v2 for maize. DAP-seq peaks were called by the GEM peak caller (Guo et al., 2012) and ChIP-seq peaks by MACS2 (Zhang et al., 2008) with the IDR pipeline for replicated samples (Li et al., 2011). Motif discovery was performed using the MEME-ChIP suite (Machanick and Bailey, 2011). Binding signals were calculated by deepTools (Ramírez et al., 2014), and GO enrichment was calculated by g:Profiler (Reimand et al., 2011).

See the Supplemental Experimental Procedures for additional details.

ACCESSION NUMBERS

The accession number for the raw and processed data of DAP-seq and ChIP-seq reported in this paper has been uploaded to GEO: GSE60143.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.04.038.

AUTHOR CONTRIBUTIONS

R.C.O. and J.R.E. designed the experiments. R.C.O., A.B., S.C.H., L.S., M.G.L., M.G., and A.G. performed the experiments. S.C.H. and R.C.O. established the sample processing and bioinformatics pipelines and carried out the computational analyses. J.R.N. carried out the DNA sequencing. R.C.O., S.C.H., M.G., and J.R.E. prepared the manuscript.

ACKNOWLEDGMENTS

We thank Andrew Kuruzar for assistance with images (p40design@gmail.com). We thank Chris Benner, Ian Quigley, Debra Fulton, Robert Schmitz, and Yue Zhao for their critical reading of the manuscript. We thank Rosa Castanon for sharing biological materials. A.G. acknowledges funding from NSF (IOS-0820729/IOS-1114484). This work was supported by grants from the NSF (MCB1024999) and the Gordon and Betty Moore Foundation (GBMF3034) (to J.R.E.). J.R.E. is an Investigator of the Howard Hughes Medical Institute.

Received: November 11, 2015
Revised: February 26, 2016
Accepted: April 11, 2016
Published: May 19, 2016

REFERENCES

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. *4*, 393–411.

Berke, L., and Snel, B. (2015). The plant Polycomb repressive complex 1 (PRC1) existed in the ancestor of seed plants and has a complex duplication history. BMC Evol. Biol. *15*, 44.

Boer, D.R., Freire-Rios, A., van den Berg, W.A.M., Saaki, T., Manfield, I.W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, J.M., de Vries, S.C., Solano, R., et al. (2014). Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. Cell *156*, 577–589.

Bu, Q., Jiang, H., Li, C.-B., Zhai, Q., Zhang, J., Wu, X., Sun, J., Xie, Q., and Li, C. (2008). Role of the Arabidopsis thaliana NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. Cell Res. *18*, 756–767.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.-S.C., Schmitz, R.J., Urich, M.A., Kuo, D., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife *2*, e00675.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Corrêa, L.G.G., Riaño-Pachón, D.M., Schrago, C.G., dos Santos, R.V., Mueller-Roeber, B., and Vincentz, M. (2008). The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. PLoS ONE *3*, e2944.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Farcot, E., Lavedrine, C., and Vernoux, T. (2015). A modular analysis of the auxin signalling network. PLoS ONE *10*, e0122231.

Finet, C., Berne-Dedieu, A., Scutt, C.P., and Marlétaz, F. (2013). Evolution of the ARF gene family in land plants: old domains, new tricks. Mol. Biol. Evol. *30*, 45–56.

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

Franco-Zorrilla, J.M., López-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. USA *111*, 2367–2372.

Galli, M., Liu, Q., Moss, B.L., Malcomber, S., Li, W., Gaines, C., Federici, S., Roshkovan, J., Meeley, R., Nemhauser, J.L., and Gallavotti, A. (2015). Auxin signaling modules regulate maize inflorescence architecture. Proc. Natl. Acad. Sci. USA *112*, 13372–13377.

Guertin, M.J., Martins, A.L., Siepel, A., and Lis, J.T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet. *8*, e1002610.

Guo, Y., Mahony, S., and Gifford, D.K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput. Biol. *8*, e1002638.

Harashima, H., Dissmeyer, N., and Schnittger, A. (2013). Cell cycle control across the eukaryotic kingdom. Trends Cell Biol. *23*, 345–356.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. *45*, 891–898.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.

Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol. *12*, 918–922.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719–720.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. *11*, 204–220.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. Annu. Rev. Genet. *46*, 43–68.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779.

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, X., Noll, D.M., Lieb, J.D., and Clarke, N.D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Res. *15*, 421–427.

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

Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.

Mendillo, M.L., Santagata, S., Koeva, M., Bell, G.W., Hu, R., Tamimi, R.M., Fraenkel, E., Ince, T.A., Whitesell, L., and Lindquist, S. (2012). HSF1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. Cell *150*, 549–562.

Nanao, M.H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., Lainé, S., Wang, S., Hagen, G., Li, H., et al. (2014). Structural basis for oligomerization of auxin transcriptional regulators. Nat. Commun. *5*, 3617.

Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. *10*, 79–87.

Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nat. Genet. *47*, 8–12.

Pruneda-Paz, J.L., Breton, G., Nagel, D.H., Kang, S.E., Bonaldi, K., Doherty, C.J., Ravelo, S., Galli, M., Ecker, J.R., and Kay, S.A. (2014). A genome-scale resource for the functional characterization of Arabidopsis transcription factors. Cell Rep. *8*, 622–632.

Rajeev, L., Luning, E.G., and Mukhopadhyay, A. (2014). DNA-affinity-purified chip (DAP-chip) method to determine gene targets for bacterial two component regulatory systems. J. Vis. Exp. *Jul 21*, 89.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Ré, D.A., Capella, M., Bonaventure, G., and Chan, R.L. (2014). Arabidopsis AtHB7 and AtHB12 evolved divergently to fine tune processes associated with growth and responses to water stress. BMC Plant Biol. *14*, 150.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. *39*, W307–W315.

Schlereth, A., Möller, B., Liu, W., Kientz, M., Flipse, J., Rademacher, E.H., Schmid, M., Jürgens, G., and Weijers, D. (2010). MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. Nature *464*, 913–916.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., and Ecker, J.R. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al.; Mouse ENCODE Consortium (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Staudt, N., Fellert, S., Chung, H.R., Jäckle, H., and Vorbrüggen, G. (2006). Mutations of the Drosophila zinc finger-encoding gene vielfältig impair mitotic cell divisions and cause improper chromosome segregation. Mol. Biol. Cell *17*, 2356–2365.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.

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. *8*, 2015–2030.

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. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Ulmasov, T., Hagen, G., and Guilfoyle, T.J. (1997). ARF1, a transcription factor that binds to auxin response elements. Science *276*, 1865–1868.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Welch, D., Hassan, H., Blilou, I., Immink, R., Heidstra, R., and Scheres, B. (2007). Arabidopsis JACKDAW and MAGPIE zinc finger proteins delimit asymmetric cell division and stabilize tissue boundaries by restricting SHORT-ROOT action. Genes Dev. *21*, 2196–2204.

Worsley Hunt, R., and Wasserman, W.W. (2014). Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol. *15*, 412–428.

Xu, D., Li, J., Gangappa, S.N., Hettiarachchi, C., Lin, F., Andersson, M.X., Jiang, Y., Deng, X.W., and Holm, M. (2014). Convergence of Light and ABA signaling on the ABI5 promoter. PLoS Genet. *10*, e1004197.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell *153*, 193–205.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell *126*, 1189–1201.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.1–R137.9.

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

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. USA *112*, 4654–4659.

Cell

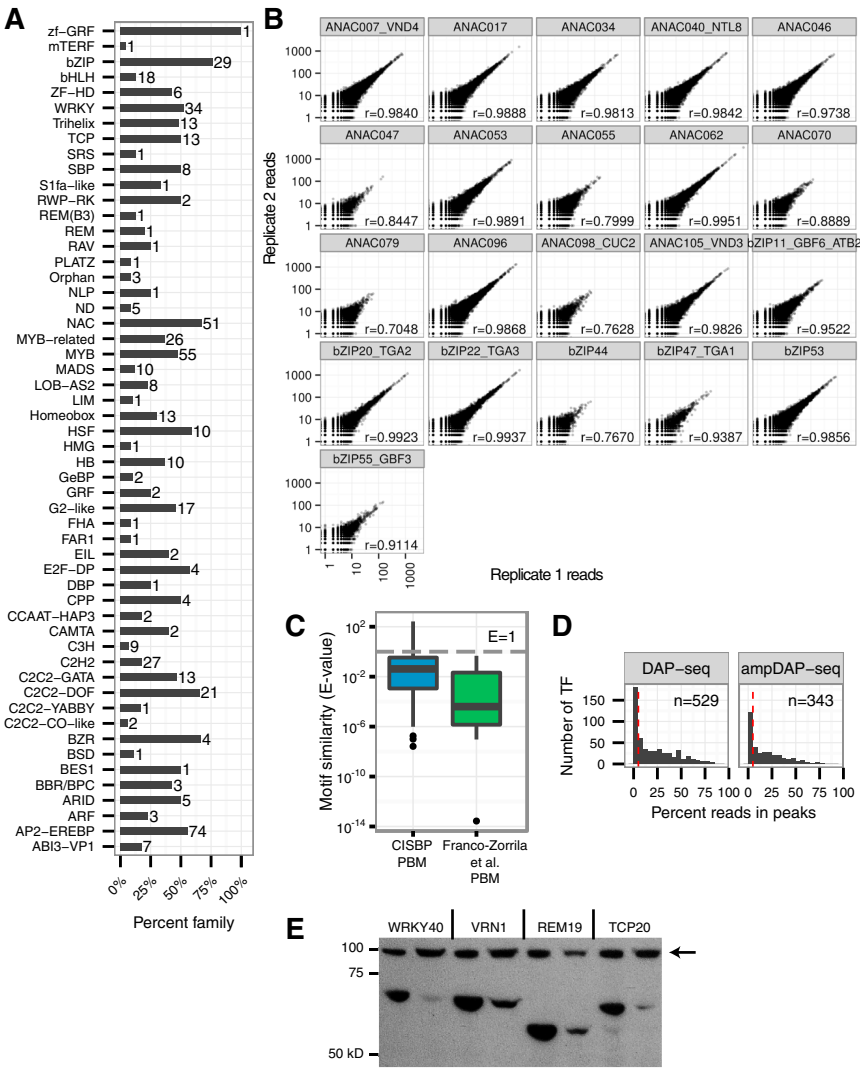


Figure S1. Global Success Rates and Quality Control of the *Arabidopsis* Cistrome and Epicistrome Datasets by DAP-Seq, Related to Figure 2

(A) Percentage of TFs with motifs successfully identified by DAP-seq in each family of the *Arabidopsis* clone collection, with the numbers of TFs next to the bar.

(B) Replicate reproducibility of DAP-seq. Scatter plots of sequencing reads in peaks between duplicate DAP-seq experiments for 21 TFs, with Pearson's correlation shown in lower right corner.

(C) Distribution of TOMTOM E-values comparing DAP-seq motifs to PBM motifs from the same TF using two large published PBM collections for *Arabidopsis*: CIS-BP (Weirauch et al., 2014), and Franco-Zorrilla et al. (2014).

(D) Histogram of percent Reads in Peaks (RiP) metric for all DAP-seq and ampDAP-seq experiments. Red vertical line indicates 5% RiP.

(E) Representative western blot for verification of protein expression in wheat germ extract. For each TF, samples on left are 10% of DAP-seq protein expression reaction; samples on right are 10% of supernatant after binding to HaloTag beads. A non-specific band of ~100 kD was detected by the polyclonal anti-HALO antibody in all samples (arrow).

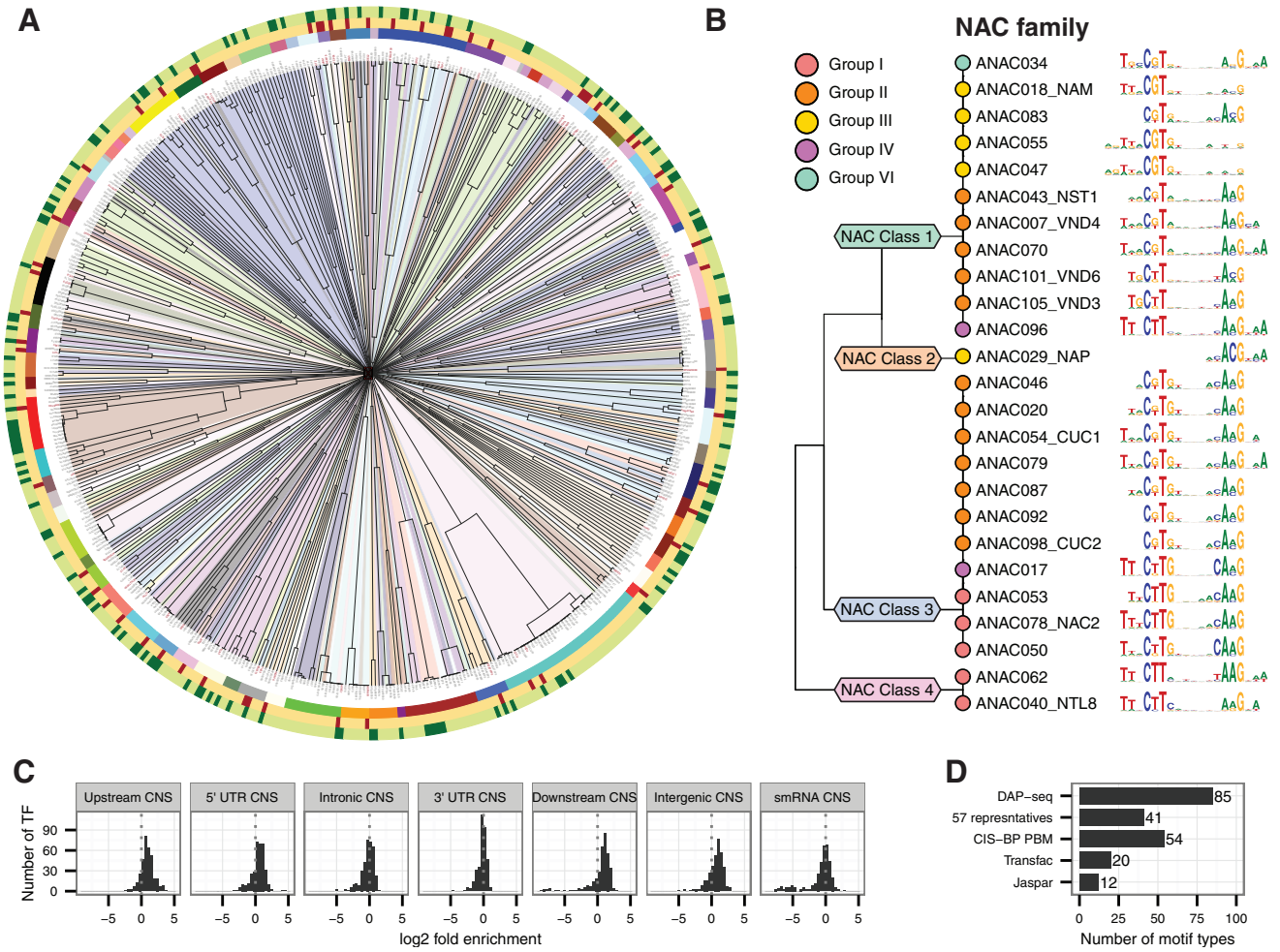


Figure S2. Motif Clustering Identified Motif Types for All Motifs and Motifs within TF Family, Related to Figure 3

(A) Hierarchical clustering dendrogram for all 529 TF motifs. Background colors represent TF families. Outer tracks, from inside to outside, are: i) motif types from applying dynamic tree cut to the clustering dendrogram; ii) the 57 selected representatives (dark red) and others (yellow); iii) the TF has a published CisBP/Transfac/Jaspar motif (dark green) or novel motif from DAP-seq (light green).

(B) Motif clustering dendrogram for NAC family members.

(C) Histogram of log2 fold enrichment of DAP-seq peak overlap with conserved non-coding sequences (Haudry et al., 2013).

(D) Number of motif types resulting from dynamic tree cut on the 529 motif clustering dendrogram and the subsets that have been represented in various motif databases.

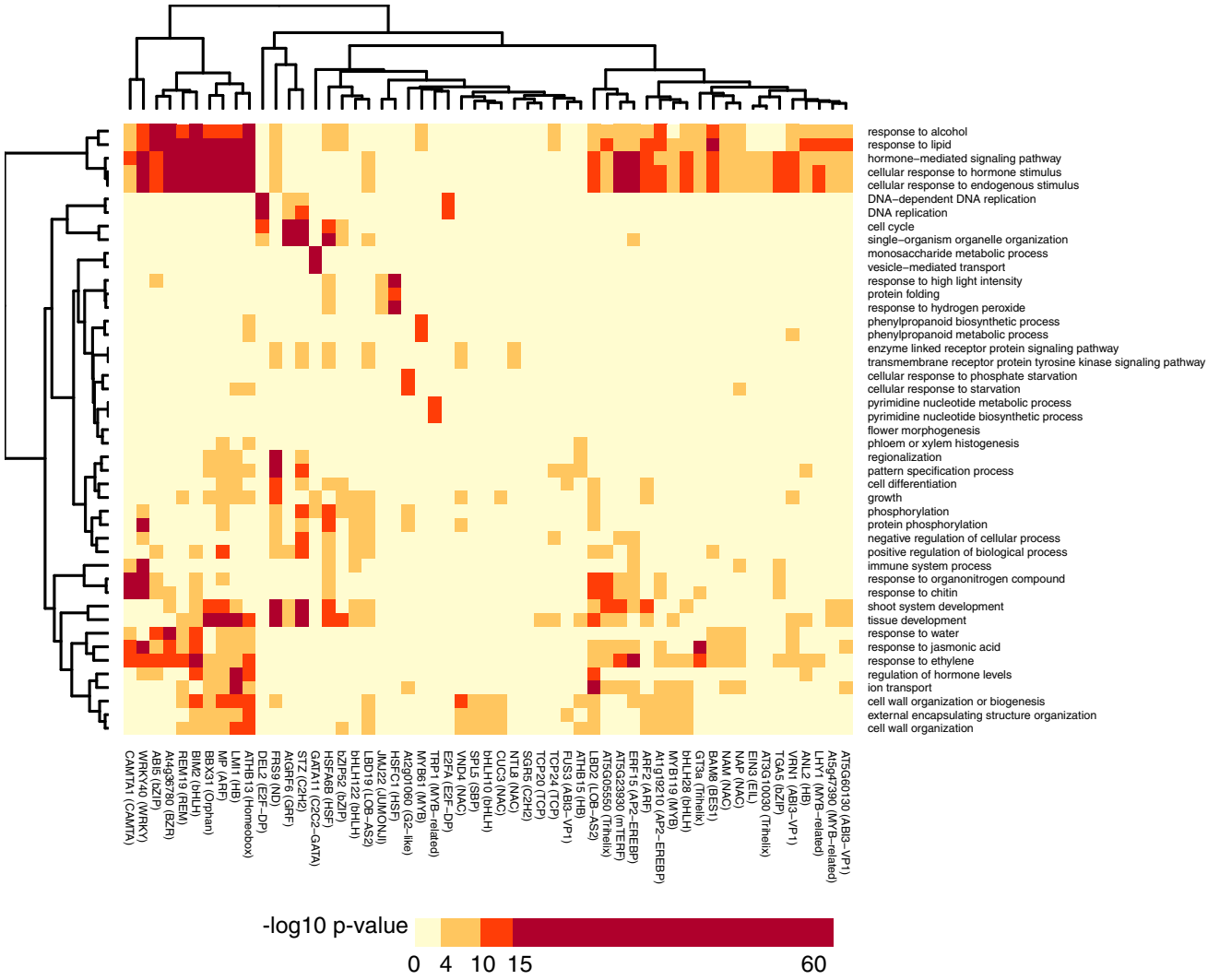


Figure S3. Specific GO Terms Were Enriched for Target Genes of a Diverse Set of TFs, Related to Figure 4

Top two enriched GO terms associated with DAP-seq target genes for each of the 56 representative TFs (excluding one out of the 57 that had less than 5% reads in peaks) were selected and their enrichment calculated for association with each target gene set.

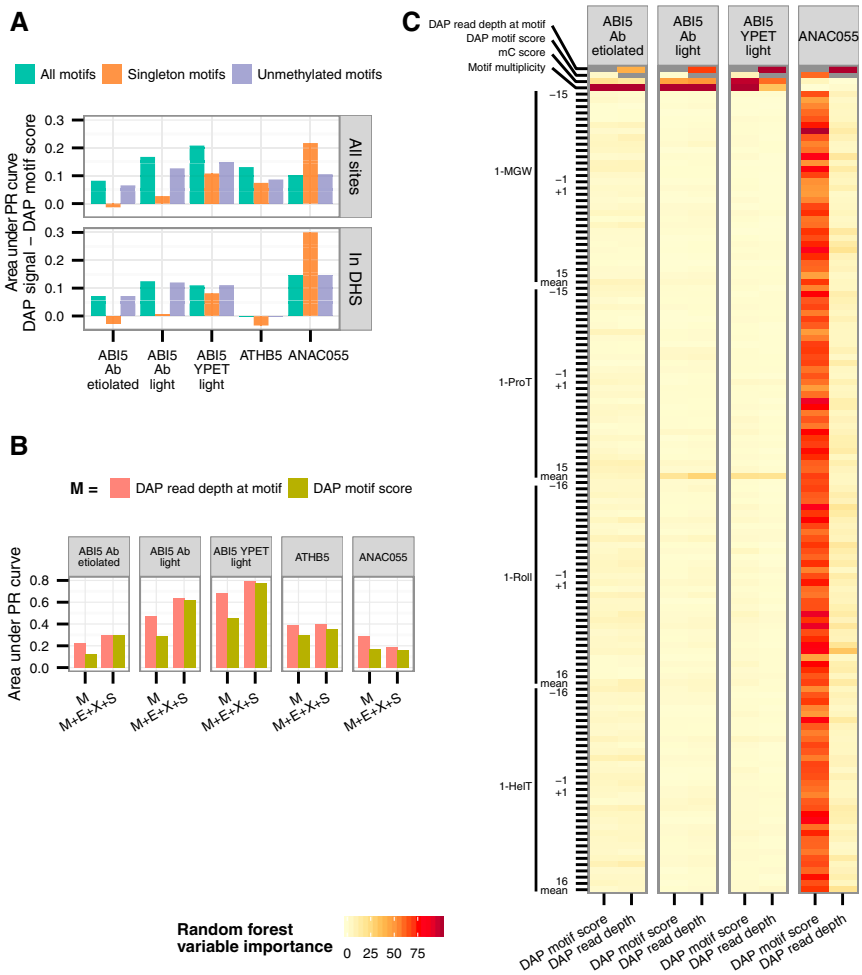


Figure S4. Comparison of Motif Score and DAP-Seq Signal Models for Predicting ChIP-Seq Binding Sites, Related to Figure 5

(A) Restricting predictions to singleton motifs and unmethylated motifs changes the differences in predictive power between DAP signal (read depth) at motif and motif score, measured by area under the precision-recall (PR) curve as computed in Figure 5. By restricting predictions to peaks containing a single motif with no strong motifs within 100 bp to remove the effect of motif clusters, the performance of motif inference relative to DAP signal was improved for ABI5 but worsened for ANAC055. By restricting predictions to motifs with no methylcytosines to remove the effect of methylation, we observed improved performance of motif inference relative to DAP signal for ABI5 in the absence of DHS filter, but not for ANAC055.

(B) Area under PR curve metrics for models of only motif feature M, where M is DAP signal at motif (read depth) or motif score, and random forest models of motif and environment features M+E+X+S, where M is DAP signal or motif score, E is methylation level in motif, X is number of motifs within 100bp (motif multiplicity), S is a set of DNA shape features (1-MGW, 1-ProT, 1-Roll and 1-HelT) within 17 bp flanking the motif.

(C) Scaled feature importance of random forest models with either DAP signal at motif (read depth) or motif score.

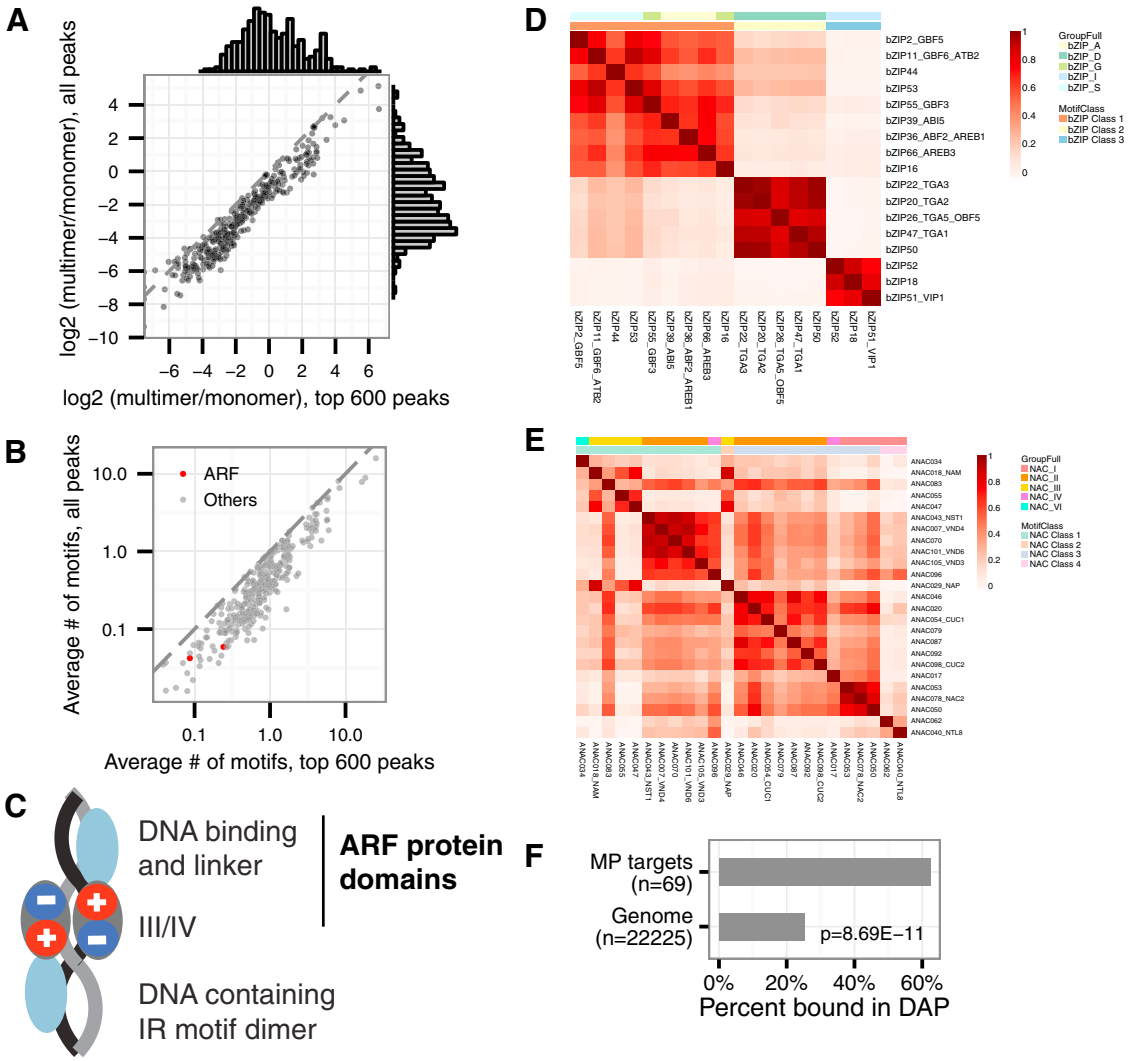


Figure S5. DAP-Seq Refined Binding Model for ARF Family TFs, Related to Figure 6

(A) Ratio of peaks with multiple motifs and peaks with single motifs, for top 600 peaks of each TF (x axis) and all peaks (y axis).

(B) Average number of motifs per peak for top 600 peaks of each TF (x axis) and all peaks (y axis).

(C) Model for the ARF homodimer at an inverted TFBS repeat. We postulate that the positively and negatively charged domains of the III/IV domain (Nanao et al., 2014) could allow for a stabilizing interaction between a pair of oppositely oriented III/IV domains over the top of the DNA helix.

(D and E) bZIP (D) and NAC (E) pairwise Pearson correlation coefficients of DAP-seq read depth in union of peaks for all factors in each family.

(F) ARF5 DAP-seq peaks are found significantly more frequently in the promoters of ARF5 target genes than in gene promoters genome-wide.

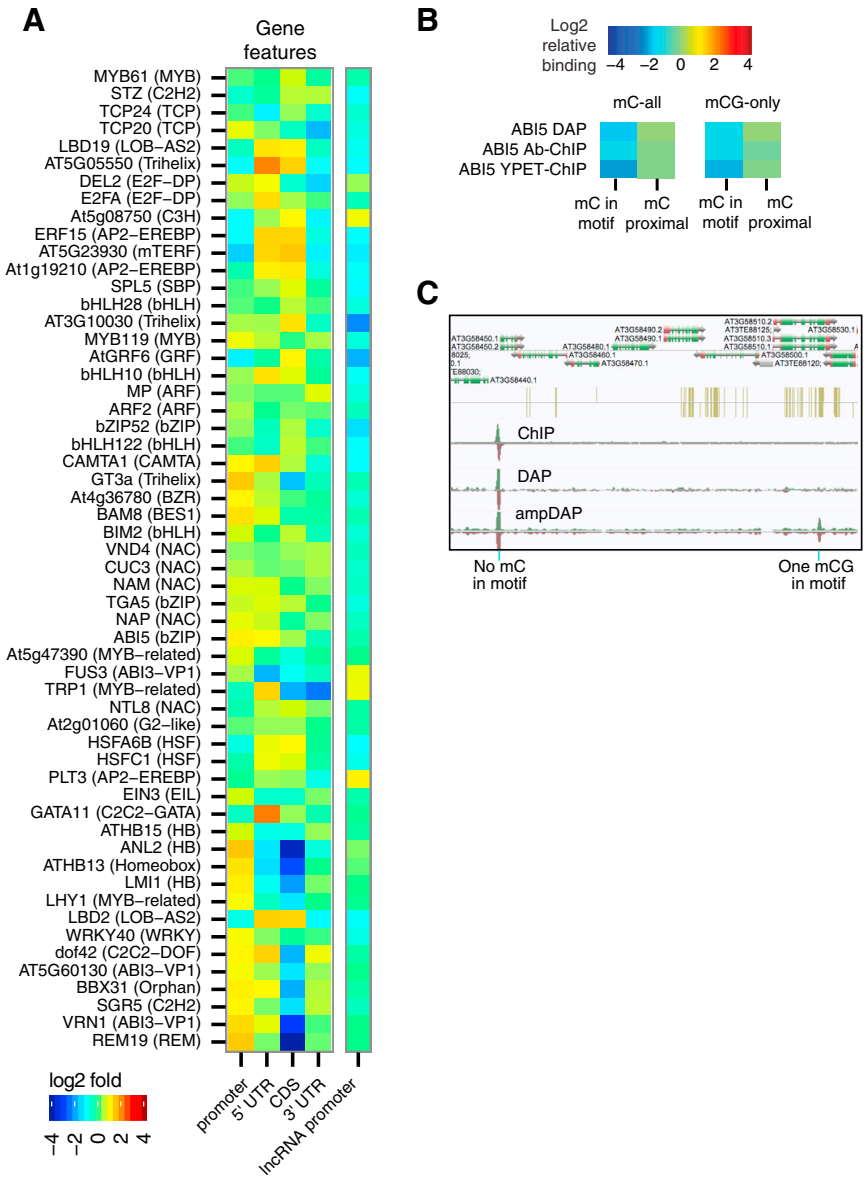


Figure S6. DAP-Seq Binding Events Are Generally Enriched at Gene Promoters and Inhibited by DNA Methylation, Related to Figure 7

(A) Association of DAP-seq peaks with TAIR10 protein coding gene features and lncRNA promoters (Xie et al., 2014), in terms of fold enrichment/depletion relative to random sampling computed by the Genome Association Tester (Heger et al., 2013).

(B) ABI5 DAP-seq and ChIP-seq binding at motifs containing methylcytosines in the mC-all and mCG-only categories, relative to motifs without methylcytosines and not in densely methylated regions.

(C) Genome browser screen shot of representative DAP-seq and ampDAP-seq peaks at methylated motifs.

Cell, Volume *165*

Supplemental Information

Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape

Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G. Lewsey, Anna Bartlett, Joseph R. Nery, Mary Galli, Andrea Gallavotti, and Joseph R. Ecker

SUPPLEMENTAL INFORMATION

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

DAP-seq Genomic DNA Library Preparation

The DAP genomic DNA (gDNA) library was prepared as a standard high-throughput gDNA sequencing library for the Illumina platform. First, gDNA (5-10µg in 130µL elution buffer (EB): 10mM Tris-Cl, pH 8.5) was fragmented to an average of 200bp using a Covaris S2 and manufacturer recommended settings. The fragmented gDNA was then purified using Sera-Mag beads (Thermo) at a 1:2 DNA to beads ratio (130µL DNA, 260µL beads). The beads were incubated with the gDNA for 10 minutes, placed on a magnet to immobilize the beads, and the supernatant was removed. The beads were then washed twice with 500µL 80% ethanol and allowed to dry. Once dry they were resuspended in 100µL EB, incubated for 10 minutes, placed on the magnet, and the DNA-containing supernatant was transferred to a new tube. 5µg (in 34µL) of DNA was end repaired in a 50µL reaction using the End-It DNA End-Repair Kit (Epicentre), and incubated at 22°C for 45 minutes. The reaction was precipitated with isopropanol, resuspended in 32µL elution buffer, and used in a 50µL A-tailing reaction using dATP and Klenow Fragment 3'->5' exo- (NEB) incubated at 37°C for 30 minutes. A second isopropanol precipitation was performed and DNA was resuspended in 10µL EB. The DNA was ligated in a 50µL ligation reaction using T4 DNA Fast Ligase (Promega)) at 22°C for 50 minutes followed by heat inactivation at 70°C for 10 minutes. Following ligation, a Sera-Mag bead purification was performed at a 1:1 ratio and the resulting DAP-seq library was resuspended in 30µL EB. The adapters sequences are truncated Illumina TruSeq adapters; the TruSeq Universal and Index adapters correspond to the DAP-seq Adapter A: CACGACGCTCTTCCGATCT and Adapter B: GATCGGAAGAGCACACGTCTG. In the PCR after the DNA affinity-precipitation step, the full-length Illumina adapter sequences including the flow cell attachment and index information are introduced in the PCR primers.

DNA Affinity Purification Sequencing (DAP-seq)

A collection of 1,812 Gateway compatible full-length Arabidopsis TF-ORFs were assembled from Pruneda-Paz et al. and AtORFeome2.0 (Arabidopsis Interactome Mapping Consortium, 2011). 27 families present in Pruneda-Paz et al. that were not annotated as DNA-binding proteins were excluded from our collection. Clones were recombined using LR clonase (Life Technologies) into the pIX *in vitro* expression vector (Arabidopsis Interactome Mapping

Consortium, 2011)) modified to contain an N-terminal HALO-Tag (Promega). pIX-HALO-ORF plasmid DNA was extracted using the Qiaprep 96-Turbo DNA extraction kit, quantified with the Quantifluor dsDNA system (Promega), and normalized prior to expression. HALO fusion proteins were expressed using the TNT SP6 Coupled Wheat Germ Extract System (Promega) following the manufacturer's specifications for expression in a 50μL reaction containing on average ~800ng DNA with a 2hr incubation at 30°C. Typical reactions yields ranged from 50-500ng of protein as measured relative to purified Halo-GST protein standards detected with anti-Halo antibody (Promega) by western blot. Expressed proteins were directly captured using Magne HaloTag Beads (Promega; 10μL per expression reaction). Proteins were incubated with the beads on a rotator for 1hr at RT. The beads were then washed three times with 125μL of wash/bind buffer (PBS with 0.005% NP-40).

The protein-bound beads were then incubated with 50ng of adapter-ligated gDNA fragments on a rotator for 1hr at RT in 50μL wash/bind buffer. Beads were washed again three times using the same wash buffer to remove unbound DNA fragments. The HaloTag beads were then resuspended in 30uL EB and heated to 98°C for 10 minutes to denature the protein and release the bound DNA fragments into solution. The supernatant was transferred to a new well. For HaloTag protein reactions, 25μL were used in a 50μL PCR reaction using Phusion polymerase and the same cycling conditions as described above for ampDAP-seq for 20 cycles. PCR primers consisted of the full-length Illumina TruSeq Universal primer (5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) and an Illumina TruSeq Index primer (5'GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCT GCTTG) where NNNNNN represents the 6 base pair sequence index used for sample identification. The PCR product was precipitated with 95% ethanol, washed with 70% ethanol, and resuspended in 21μL EB. DNA concentrations were determined using a Qubit (Life Technologies). Multiple sets of indexed DAP-seq sequencing libraries could be combined in a single flow cell lane. Most datasets described here were sequenced as pools of 48 DAP-seq libraries per flow cell lane (Illumina 2500). We found that ~2-3 million reads per experiment produced a high signal-to-noise DAP-seq datasets for Arabidopsis. The correct DAP-seq library concentration to achieve a specific read count can be calculated based on library fragment size. Negative control mock DAP-seq libraries were prepared as described above without the addition of protein to the beads.

Expression and binding-efficiency of Halo-fusion proteins were confirmed by western

blot for a subset of TFs from multiple families. All samples tested showed protein expression at the expected size (Figure S1E). We observed no noticeable difference in expression or Halo-Tag binding efficiency in samples that generated high quality DAP-seq results and those that failed to pass our quality filters, indicating that the failure of certain TFs was due to factors other than lack of expression of full length protein in the wheat germ extract.

We retested a set of 201 TFs to determine the reproducibility rate of DAP-seq using TFs that either succeeded or failed in the initial experiments (Table S2). A set of 32 TFs that were successful when initially tested, reproduced at a rate of 88% indicating a technical failure rate estimate of ~12%. We believe this technical failure is likely due to operational errors associated with the scale of the experiment (201 TFs tested) and may be lower when assaying a smaller number of TFs. However as our goal was to establish a technical failure rate of the high throughput approach, we believed it important to perform our quality control in a large-scale experiment. We also retested 168 factors that failed in the first experiment to determine the rate of rescue of failed TFs. We examined several families that showed either a high (bZIP, NAC) or low (MADS, GRF, C2H2) initial success rate to determine if family success rates influence the rate of recovery in retests. Repeated attempts resulted in successful DAP-seq datasets for only ~6% of TFs tested, compared to an 88% reproducibility rate for initially successful TFs (Table S2). Moreover, families with high initial success rates (bZIP, NAC) showed higher rescue, 18% and 13% respectively, than those with lower initial success rates. This is illustrated by the failure to rescue any of the 87 MADS family members in a retest (Table S2) suggesting that some TF families may require alternative conditions for optimal DNA-binding.

DAP-seq datasets for ARF5, and ZmARF5 were performed using *E. coli* expressed GST-fusions generated during the initial phases of DAP-seq development. TFs were recombined into pDEST15 and expressed as follows: GST-HAT2 fusion protein was expressed in a 10mL culture of One Shot BL21 Star (DE3) cells using 1mM IPTG to induce protein expression for 5hr at room temperature. After expression, the cells were pelleted, frozen overnight, and lysed using 1mL B-PER II Bacterial Protein Extraction Reagent (Thermo Scientific) mixed with 15mg lysozyme (Thermo Scientific) and 5µL protease inhibitors (Sigma). The GST-ARF5 and GST-ZmARF5 fusion proteins were expressed in 500ml of BL21 DE3 codon plus cells (Stratagene) with 0.4mM IPTG at 23°C for 4 hours. Cells were lysed similar to that described above with an additional sonication step. GST-fusion proteins were column purified using Glutathione Sepharose 4B (GE Healthcare) and ~2ug of purified protein was used in the affinity capture reactions as described for Halo-fusion proteins with the following

modifications: GST-fusion proteins were captured using Glutathione-Superflow Resin (Clontech; 300μL per pellet) and washed with 1mL of GST wash buffer (25mM Tris pH 7.5, 150mM NaCl, 1mM EDTA). The protein-bound resin was incubated with 800ng (or 5ug for maize) of adapter-ligated gDNA fragments, washed, and resuspended in 200μL EB prior to DNA elution. 100μL were used in a 200μL PCR reaction split across two wells. PCR primers and conditions were identical to those described above for Halo-fusions. The full-length maize ARF5 co-ortholog ZmARF29 (GRMZM2G086949; (Xing et al., 2011)) was PCR amplified from B73 cDNA (mixed stage ear).

Generation of Transgenic Lines

Recombineered gene tagging of ABI5 (AT2G36270), HB5 (AT5G65310) and ANAC055 (AT3G15500) were carried out as described previously (Alonso and Stepanova, 2014). An YPET-6xHis-3xFLAG tag was fused in frame right before the stop codon of the *ABI5*, *HB5* and *ANAC055* genes in the transformation-competent bacterial artificial chromosome (TAC) clone JAtY64K17, JAtY58F22 and JAtY59D10, respectively. The resulting constructs were introduced into wild-type *A. thaliana* (accession Col-0) by the floral dip transformation method (Bent and Clough, 1998). Single-insertion transgenic lines were selected by the Chi-square test on plates containing Linsmaier and Skoog medium (Caisson labs, USA) with 0.8% agar and 15 μg/ml glufosinate ammonium (Sigma-Aldrich, USA) from plants in the T2 generation. The expression of the tagged transcription factors was confirmed by western blotting. Homozygous transgenic lines were selected from the subsequent generation and used for bulking seeds. The ABI5-YPET line using in our ChIP-seq experiments exhibited mild over-expression of ABI5, but the plant responded to ABA treatment phenotypically and transcriptionally.

ChIP-seq

To ChIP ABI5, seedlings were germinated and grown on nylon mesh in hydroponics for 36 hours under long day light conditions, followed by 5 μM ABA treatment for 4 hours. To ChIP HB5, seedlings were grown in dark for 3 days. ChIP-Seq was carried out as described previously with minor modifications (Chang et al., 2013). Briefly, harvested seedlings were cross-linked by 1% formaldehyde solution (Sigma-Aldrich, USA) under vacuum for 20 minutes. After nuclei isolation, chromatin was sonicated to 100-400bp fragments. Native ABI5 in wild-type Col-0 plants was immunoprecipitated by a rabbit polyclonal ABI5 antibody (cat # ab98831, Abcam, USA). Tagged ABI5 or HB5 proteins in the transgenic lines were immunoprecipitated by

a rabbit polyclonal GFP antibody (cat # A11122, Thermo Fisher Scientific, USA). Col-0 was immunoprecipitated with either rabbit IgG or the GFP antibody as mock IP for the native ABI5 and tagged transcription factors, respectively. After elution and reverse crosslinking, ChIP DNA was used to generate sequencing libraries according to the Illumina ChIP-Seq instructions. The libraries were sequenced on an Illumina HiSeq 2500 with 100bp SR according to the manufacturer's instructions (Illumina, USA).

To ChIP ANAC055, etiolated seedlings expressing recombineered ANAC055 were grown in the dark on agar plates, as described previously (Chang et al., 2013). Plates were sprayed with a control solution of 0.5x Linsmaier and Skoog medium containing 0.05% (w/v) ethanol and seedlings harvested two hours later. Cross-linking, nuclei extraction and immunoprecipitation were conducted as for tagged ABI5, but using polyclonal anti-GFP antibodies supplied by David Drechsel (Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany). Control ChIP experiments were conducted using wild-type *A. thaliana* treated in the same manner. Immunoprecipitation was conducted using rabbit whole IgG (Covance) for these experiments. ChIP-seq libraries were generated and sequenced as for ABI5.

DAP-seq Data Processing

Reads were mapped to the TAIR9/10 genome sequence using bowtie2 (Langmead and Salzberg, 2012) version 2.0-beta7 with default parameters and post-processing to filter out reads mapped to multiple locations. Peak calling was done using GEM peak caller version 2.5 (Guo et al., 2012) with the TAIR9/10 genome sequence and following parameters "--f SAM --k_min 6 --kmax 20 --k_seqs 600 --k_neg_dinu_shuffle" limited to nuclear chromosomes only. For factors with technical replicates, GEM was called with the replicate mode. Quality control metrics were computed by the R package ChIPQC (Carroll et al., 2014). Read coverage across the genome was computed in deepTools (Ramírez et al., 2014) by extending the reads to 150bp (DNA fragment size of the assay) and normalized by the number of reads mapped to nuclear chromosomes to base pair FPKM values. Association of DAP-seq peaks with gene features and repeat regions was computed by the Genome Association Tester (Heger et al., 2013). The peak calling and analysis pipeline was created in the snakemake framework (Köster and Rahmann, 2012).

Correlation between Factors and Factor Replicates

A consensus peak set for each TF family was generated by merging peak summits from all the TFs in the family that were within 100bp of one another. Average FPKM values of DAP-seq read counts were calculated for 100bp up- and downstream of each region in the consensus and used in computing the Pearson correlation coefficients between replicate libraries of the same factor or between (merged) libraries of different factors.

Motif Discovery, Clustering and Scoring

Since the kmer search approach used by GEM tends to discover short motifs, we used the meme-chip tool (Machanick and Bailey, 2011) in MEME suite 4.10.1 to allow identification of longer motifs. For each set of GEM-called TF binding events, we retrieved 200bp sequences surrounding the top 600 events and ran meme-chip with default parameters. The top motif from each factor was selected and evaluated for central enrichment. Motif PWM models were clustered by functionalities provided in the MotIV BioConductor package (Mahony and Benos, 2007; Mahony et al., 2007; Mercier et al., 2011) and visualized by the motifStack package. The PWM were used to score for motif matches in the TAIR10 genome sequence by the motif search tool FIMO (Grant et al., 2011), with a zero-order background model of TAIR10 nucleotide composition. To compare the number of motif matches from DAP-seq and PBM PWM models, a threshold of 65% maximum score for each motif was used for all the factors.

Comparison of DAP-seq and ChIP-seq

ChIP-Seq reads were mapped to the TAIR9/10 genome sequence using bowtie2 (Langmead and Salzberg, 2012) version 2.0-beta7 with default parameters. Since the GEM peak caller reports individual point source binding events but the ChIP-seq peaks have a broad appearance, to identify ChIP-seq peaks we used MACS2 peak callers. For ABI5 ChIP-seq (Ab etiolated, Ab light, YPET light) we used MACS version 2.0.10 with the parameters "--gsize 1.19e8 --nomodel –shiftsize 150 –keep-dup auto –call-summits --bdg" and a mock IP sample as control. For consistency MACS2 was also used to call peaks for DAP-seq of these factors, with parameters "--gsize 1.19e8–keep-dup auto –call-summits –bdg --SPMR". The peaks reported (default minimum q-value cutoff 0.05) were further filtered by a minimum fold enrichment threshold of 3 for DAP-seq, 2.5 for ABI5 YPET, and 2 for the other ChIP-seq experiments. For HB5 and ANAC055 ChIP-seq, each with two replicates, we used the IDR pipeline with the MACS peak caller (Li et al., 2011). Peak comparison calculations were done using BEDTools 2.19.1 (Quinlan and Hall, 2010), BEDOPS 2.4.1 (Neph et al., 2012) and UCSC genome browser

utilities (Kuhn et al., 2013). ChIP peaks were divided into equal-sized quartiles by the minimum p-value of the motifs in each peak and percentage of peak overlap with DAP peaks was computed for each quartile. To calculate overlapping peak categories for the empirical cumulative distribution curves, a union peak set was first created by merging peak regions from the two assays. Each merged region was designated as “DAP-ChIP” if it contained peak regions from both assays; “DAP-only” or “ChIP-only” if it contained regions only from the respective assay. A peak was considered to be in a DHS if it overlapped with a DHS region by more than 50%. For comparisons of DAP and motif-based predictions, motifs discovered from the top 600 DAP-seq peaks and motifs from PBM or SELEX assay were used to scan the TAIR10 genome sequence with p-value threshold of 1E-4. A motif match was considered bound in DAP-seq or ChIP-seq if it completely fell within peak regions of either assay, and was in DHS if it fell inside a DHS region. DAP signals at motif matches were computed by averaging the normalized DAP-seq read depth (in FPKM) over regions in -10bp to +10bp flanking regions of the motif match, with non-covered bases counted as zero. Precision-recall curves and AUC were computed by the R PRROC package (Keilwagen et al., 2014).

The random forest classifiers including motif and environment features for predicting in vivo binding sites of ABI5 and ANAC055 from ChIP-seq can be represented as:

$$Y \sim M + E + X + S$$

where Y is a binary variable indicating whether a motif is bound in ChIP-seq, M is motif feature (described below), E is the level of methylation inside the motif, X is the number of motifs of the same factor found within 100bp and S is a set of four first order DNA shape features for 17-bp immediately flanking both sides of each motif (30 features for minor groove width, 30 features for propeller twist, 32 features for roll and 32 features for helix twist), computed by the DNAshapeR package (Chiu et al., 2015). To compare motif scores and DAP-seq signal, two random forest classifiers were trained for each ChIP-seq dataset, consisting of the same motif environment features (E, X, S) but differed by the motif feature M: in the first one M was the motif score for motifs with p-value threshold of 1.5E-4 and in the second one M was the DAP-seq read depth computed as above on the same set of motifs as in the first model. Training and testing of the classifiers were performed by functionalities provided by the R caret package (Kuhn, 2008) calling the randomForest package (Liaw and Wiener, 2002). For training, 50% of the data were used in 5-fold cross-validation repeated 5 times, using the one standard error rule for selecting tuning parameters and ROC as the performance metric. Precision-recall metrics

were computed using the remaining 50% of the data. Scaled variable importance were retrieved by the caret package and plotted as heatmap.

GO Enrichment and Gene Feature Overlap Analysis

Peak summits called by GEM were associated with the closest TAIR10 gene model using the BioConductor package ChIPpeakAnno (Zhu et al., 2010). Enriched GO terms were identified by the g:Profiler web service accessed via the R API (Reimand et al., 2011) limited to the Biological Process ontology, maximum size of functional category of 1500, no hierarchical filtering and corrected for multiple hypothesis testing by the default method g:SCS. The two most significantly enriched terms were plotted in the heatmap by the aheatmap function in the R package NMF (Gaujoux and Seoighe, 2010). Enrichment/depletion of DAP-seq peaks with gene features and repeat/non-repeat regions were computed by the Genome Association Tester (Heger et al., 2013).

Identification of mCG-only and mC-all Sites and Comparison of mC to DAP-derived Motifs

For the methylation analysis we used methylcytosine calls from *Arabidopsis thaliana* Col-0 leaf from the GEO accession number GSM1085222 (Schmitz et al., 2013). The same seed stock used to create this leaf methylome map was used for all the DAP-seq leaf gDNA libraries. A 5'-methylcytosine site was called at any cytosine with read coverage of three or more and a ratio of mC over total read depth (methylation site-frequency) greater than 15%. Using this threshold, 11% of all cytosines with sufficient coverage were found to be methylated in the Col-0 leaf tissue. To identify the mC-all regions, we examined 100bp on either side of each methylcytosine. If the site was a non-mCG then the site was considered mC-all. If the site was mCG but an additional non-mCG site was found in the 30bp window the region was also designated as an mC-all. All mCG sites that did not neighbor mC-all regions were designated as the mCG-only set. A motif match was considered methylated if the total methylation level in the motif match region (mC/C) exceeded 0.66. This high threshold ensured that most gDNA fragments associated with the motif matches will contain a 5'-methylcytosine to allow us to measure the impact of the chemical mark on TF binding. Unmethylated motif matches (mC/C of 0.0) are used to establish background control levels.

Quantification of Cytosine Content and Classification of Cytosine Context for Motifs

Normalized cytosine content of a motif was calculated by dividing the sum of cytosine information content (IC) across all positions in the motif by the total IC of the motif. To assign motifs to CG, CHG and CHH contexts, we first computed distributions of IC for each nucleotide A, C, G, and T in all positions in all motifs. Nucleotides at each position were considered *strong* if their IC at this position was within the top 15% of the IC distribution for this nucleotide. A motif was considered to contain CG if it contained neighboring strong C and strong G. Similarly, a motif contained CHG if it contained neighboring strong C, strong A/C/T and strong G, while CHH-containing motifs were classified as containing neighboring strong C, strong A/C/T and strong A/C/T. Both the motif and its reverse complement were scored. All IC calculations were computed by functionalities in the R package motifStack with TAIR10 nucleotide frequencies as background.

Microarray Analysis to Identify ARF5/MP Targets

Series matrix file of accession GSE13881 was downloaded from NCBI Gene Expression Omnibus and analyzed as described in (Schlereth et al., 2010). Briefly, the normalized intensity values were log2 transformed and probe sets with low variance across samples (the lowest 0.5 percentile in terms of interquartile range) were removed. A linear model was fitted for each gene, and moderated t-statistics were calculated by an empirical Bayes method (Ritchie et al., 2015) for two comparisons 1) between *mp* and GR-*bdl* –DEX and 2) between GR-*bdl* +DEX and –DEX. A probe set was differentially expressed if the adjusted p-value was lower than 0.001 and fold change was at least two. Sixty-nine genes that were differentially expressed in both comparisons were considered ARF5/MP targets. A set of 62 genes that had p-values greater than 0.6 in both comparisons were considered background genes.

SUPPLEMENTAL REFERENCES

Alonso, J.M., and Stepanova, A.N. (2014). A Recombineering-Based Gene Tagging System for Arabidopsis. In Methods in Molecular Biology, (New York, NY: Springer New York), pp. 233–243.

Bent, A.F., and Clough, S.J. (1998). Agrobacterium Germ-Line Transformation: Transformation of Arabidopsis without Tissue Culture. In Plant Molecular Biology Manual, (Dordrecht: Springer Netherlands), pp. 17–30.

Carroll, T.S., Liang, Z., Salama, R., Stark, R., and Santiago, I. de (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in Genetics *5*, 1–11.

Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.-S.C., Schmitz, R.J., Urich, M.A., Kuo, D., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife Sciences *2*, e00675.

Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2015). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics 1–3.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics *11*, 367.

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

Guo, Y., Mahony, S., and Gifford, D.K. (2012). High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput. Biol. *8*, e1002638.

Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046–2048.

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. PLoS ONE *9*, e92209.

Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in Bioinformatics *14*, 144–161.

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

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics *5*, 1752–1779.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News *2*, 18–22.

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

Mahony, S., and Benos, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. *35*, W253–W258.

Mahony, S., Auron, P.E., and Benos, P.V. (2007). Inferring protein DNA dependencies using motif alignments and mutual information. Bioinformatics *23*, i297–i304.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., and Gottardo, R. (2011). An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-Seq.

PLoS ONE *6*, e16432.

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. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

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

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. *39*, W307–W315.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47–e47.

Schlereth, A., Möller, B., Liu, W., Kientz, M., Flipse, J., Rademacher, E.H., Schmid, M., Jürgens, G., and Weijers, D. (2010). MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. Nature *464*, 913–916.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., et al. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 42, D98-D103.

Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237.