An atlas of conserved transcription factor binding sites reveals the cell type-resolved gene regulatory landscape of flowering plants

Leo A Baumgart[1]*, Abraham Morales-Cruz[1]*, Sharon I Greenblum[1]*, Peng Wang[1]*, Yu Zhang[1]*, Lin Yang[2], Cindy Chen[1], David J Dilworth[1], Alexis C Garretson[34], Nicolas Grosjean[1], Guifen He[1], Emily Savage[1], Yuko Yoshinaga[1], Ian K Blaby[1], Chris G Daum[1], Ronan C O'Malley[12]**

* Equal-first authors
** Corresponding author

1 U.S. Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
2 Department of Human Genetics, University of Chicago, Chicago, 60637, Illinois, USA
3 The Jackson Laboratory, Bar Harbor, ME
4 Tufts University Graduate School of Biomedical Sciences, Boston, MA

Summary

Transcription factors (TFs) play a central role in regulating gene expression, a process fundamental to cellular function. Characterizing transcription factor binding sites (TFBSs) is essential for understanding TF functions and identifying TF target genes, but many predicted TFBSs lack clear biological significance. We generated an atlas of TFBSs across ten flowering plants and demonstrated that conserved TFBSs are strongly enriched for genomic and epigenomic signatures of functional regulatory elements. We uncovered conserved properties of TFBSs, including TF family-specific clustering of binding sites in distinct promoter and gene body locations. By integrating this atlas with single nuclei RNA and chromatin profiles, we elucidated conserved TF activity underlying development and stress responses and annotated key cell type-specific regulatory pathways. Finally, we found that TF ortholog binding specificity is preserved over long evolutionary times, suggesting that divergence of regulatory sequences, rather than TFs, is a primary driver of regulatory evolution.

Introduction

Transcription factors (TFs) regulate gene expression by binding to genomic regions known as transcription factor binding sites (TFBSs). When these sites are part of cis-regulatory elements (CREs)—regions of DNA that modulate expression of nearby genes—they play a central role in coordinating gene expression programs. Identifying TFBSs and determining their contribution to CRE function is essential for understanding the genetic programs that drive growth, development, and responses to environmental signals.

A variety of methods have been developed to map TFBSs, each offering distinct strengths and limitations. In vivo approaches such as chromatin immunoprecipitation sequencing (ChIP-seq) provide valuable tissue-specific context for TF binding but are more technically demanding and limited in scalability[1]. Additionally, as they capture only the sites actively bound in the sampled tissue, they do not provide a full profile of all potential binding sites[1]. In vitro techniques like SELEX[2], which uses synthetic DNA pools to determine consensus TF binding motifs, are highly scalable. However, using these short motifs to infer binding sites in large eukaryote genomes can lead to high false-positive rates[3]. DNA affinity purification sequencing (DAP-seq) is an in vitro method for profiling TF binding to purified genomic DNA,

which easily scales to enable mapping of all potential TFBSs within a native genomic sequence context and generally has shown better agreement with ChIP-seq in terms of overlap and number of predicted sites[4–8]. However, TF binding maps produced by DAP-seq, while comprehensive, do not reflect the in vivo cellular context and native chromatin state, and as a result include both functionally important and less significant or non-functional sites[5,6].

A persistent challenge across all methods is identifying the TFBSs that drive essential regulatory functions and control genes involved in core TF processes. One promising avenue leverages comparative genomics to identify the TFBSs that have been retained across evolutionary lineages, as these sites are often critical for survival or underlie key lineage-specific traits[9–14]. Studies following this approach have yielded valuable insights, yet many were restricted to only a small number of transcription factors, as scalability of most existing technologies is limited. To address this gap, we developed a multiplexed version of DAP-seq (multiDAP) for eukaryotes. MultiDAP enables simultaneous mapping of TFBSs across pooled genomic DNA samples from multiple species. This high-throughput method accelerates the discovery of conserved TFBSs, providing insights into their evolutionary conservation and regulatory roles across diverse genomes.

By applying our optimized version of multiDAP to a wide range of flowering plant species, we efficiently generated a large-scale atlas of conserved TFBSs for a diverse array of representative plant TFs. The scale of this atlas enabled us to explore key patterns of TF binding including positional and epigenetic preferences, as well as motif conservation across TF families and TF orthologs. Furthermore, by integrating our conserved TFBS atlas with single nuclei RNA and ATAC sequencing (snRNA-seq and snATAC-seq), we devise an approach to effectively probe TF activity across cell types and environmental perturbations. The discoveries facilitated by this resource provide the foundation for a deeper understanding of core TF-target gene relationships, evolutionary histories, and the DNA regulatory landscape underlying growth, development, and environmental response programs, as well as other agriculturally important traits.

Results

Genome-wide TF binding maps across plants

In this study we optimized multiDAP for large eukaryotic genomes and applied it to investigate the function of 360 TFs across Brassicaceae, and then extended our approach to ten plant species representing a period of 150 million years of divergence across flowering plants[15] (**Figure 1A**, **Table S1**). This group of 10 species included the model plant *A. thaliana* plus three other members of the Brassicaceae family: *Arabidopsis lyrata*, *Capsella rubella*, *Brassica oleracea* (wild cabbage), four additional members of the eudicots: *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Fragaria vesca* (wild strawberry), *Populus trichocarpa* (black cottonwood), and two members of the monocots in the grass family Poaceae: *Oryza sativa* (rice) and *Sorghum bicolor* (**Table S2**). As our goal was to capture all TFBSs genome-wide, we used PCR-amplified genomic DNA fragments in the multiDAP assay to remove any native DNA methylation that could impact TF binding. This allowed us to produce comprehensive maps of all potential TFBSs across the genomes of ten species, independent of tissue- or condition-specific DNA modifications.

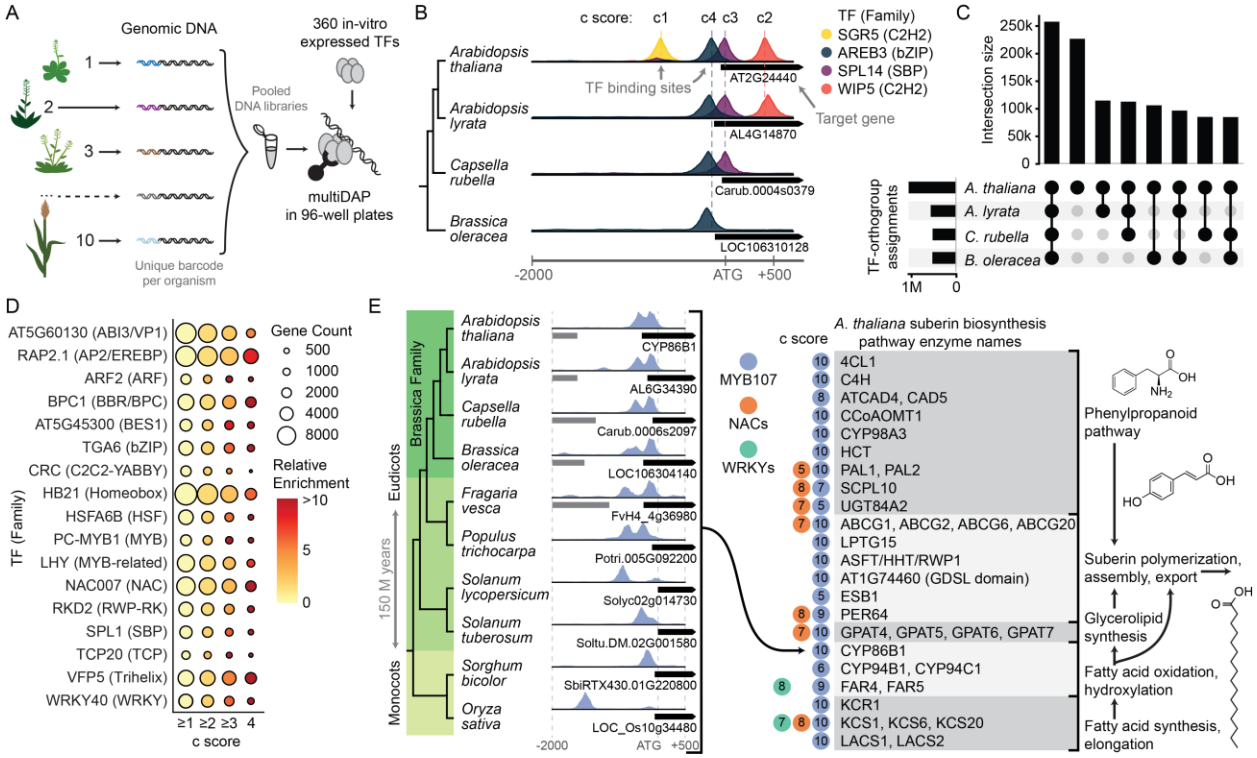


Figure 1. Plant multiDAP identifies conserved TFBSs across 150 million years of divergence. See also Figure S1 and Table S2. (A) Experimental overview of multiDAP in plants. (B) Example of peak assignment to genes and conservation scores (c scores). (C) Number of TF target genes shared between *A. thaliana* and other Brassicaceae species. (D) Overrepresentation test for conserved sites. All enrichments with c scores ≥ 2 were significant at $p < 0.01$. (E) Left side shows the phylogenetic relationships between the 10 species used in this study. Track plots show as an example the target gene CYP86B1, which encodes an enzyme in the suberin pathway in *A. thaliana,* and its orthologs all targeted by MYB107. Right side shows key enzymes in suberin biosynthesis, which are ultra-conserved targets of MYB107 and other TFs in the NAC and WRKY families. The color of the circles and inset numbers next to each enzyme name indicate the TFs and the c score, respectively.

TFs bind in promoters of a core set of conserved target genes

Our next goal was to assess target gene conservation by determining whether a TF binds to orthologous genes in multiple species. First we assigned all protein-coding genes from all species to orthogroups (an assignment based on shared amino acid sequences)[16]. Then, for each *A. thaliana* target gene with a binding site for a particular TF, we counted the number of species in which an orthologous gene also had a TFBS for the same TF. Thus, for each *A. thaliana* TF-target gene relationship, we generated a conservation score (c score) (**Figure 1B**). Target genes scored as c1 represent genes where a binding site for a given TF could only be found in *A. thaliana*, whereas target genes scored as c4 represent genes where binding sites for a given TF were found in *A. thaliana* plus orthologous genes from three other species. Our approach allowed us to detect conservation even across highly diverged species where intergenic sequence alignments cannot reliably identify conserved TFBSs[10,12,17].

Our atlas assigned c scores to 1.1 million total TF-orthogroup interactions represented in the *A. thaliana* genome. Of these, 24% were conserved across all four Brassicaceae family genomes (c4), with the next largest category at 21% which were private to *A. thaliana* (c1) (**Figure 1C**, see **Data File S2** for all Brassicaceae target genes). By comparing the number of TF-target genes shared between two or more (c2+), three or more (c3+) or all four species (c4)

to a background frequency calculated from overlap between sets of randomly generated TF target orthogroups, we found that all conserved TF target gene sets were significantly larger than expected by random chance. Furthermore, the enrichment and significance increased as the c score increased (**Figure 1D**). This indicates that there is a large set of core TF target genes that is evolutionarily maintained across the Brassicaceae.

We next investigated the degree to which these core target genes are conserved across flowering plants by examining the binding of 74 representative TFs in the genomes of 10 plants. Although running the multiDAP assay with 10 genomes simultaneously introduced a slight decrease in signal for some TFs, overall binding profiles remain similar (**Figure S1A**). The 150 million year evolutionary divergence between the monocots and eudicots[15] within this set of species enabled identification of a smaller ultra-conserved TF target gene set. Among the target genes conserved in all 10 species (c10 target genes), we observed many that are consistent with previously elucidated TF functions and capture ancient processes shared across flowering plants. For example, we found that the MYB family TF MYB107, along with several NAC and WRKY family TFs, targets a core set of genes that are consistent with their known function as regulators of suberin production[18–20]. MYB107 target genes with high c scores include key enzymes at all major parts of the suberin biosynthesis pathway and the upstream phenylpropanoid and fatty acid synthesis pathways, as well as transporters involved in exporting suberin monomers across the cell membrane (**Figure 1E**)[18,21,22]. In addition, we observed many conserved yet previously unknown MYB107 target genes with predicted functions consistent with a role in suberin synthesis and related pathways, which may serve as attractive targets for future characterization. Many other examples of TFs with ultra-conserved target genes consistent with known function include Dehydraration Response Element 1A (DREB1A) in drought response[23], PC-MYB1 (aka MYB3R1) in cell division[24], CAMTA1/EICBP.B in stress hormone responses[25], and GTL1 in root hair development[26] (see **Data File S3** for all c10 TF target genes). In summary, we identified a large number of conserved TF target genes in the Brassicaceae family, as well as a smaller ultra-conserved set of TF target genes that point to core functional regulons across flowering plants.

Conserved TFBSs and target genes exhibit signatures of CREs

To further validate that these conserved TFBSs represent functional regulatory TF-target gene interactions, we investigated whether TF target gene sets are enriched for Gene Ontology (GO) functional terms. Within the Brassicaceae dataset, we found that while c1 target gene sets are rarely enriched for GO terms, conserved *A. thaliana* target gene sets with c scores of c2 or greater are significantly enriched for functional GO terms that are consistent with known TF functions (**Figure 2A**, see also **Table S3**). Examples include GO terms related to organ development for CRC[27], cell division for MYB3R-4[28] and DEL2[29], heat stress for HSFA6B[30], suberin synthesis for MYB107[18], and response to pathogens and stress related hormones for TGA6[31], WRKY18[32] and NAC007 (aka VND4)[33]. The number of significantly enriched GO terms increases as the conservation level increases (**Figure 2B**). In the ten species dataset, we also found many examples where the most conserved c10 target genes are strongly enriched for the expected functional annotations (**Table S4**).

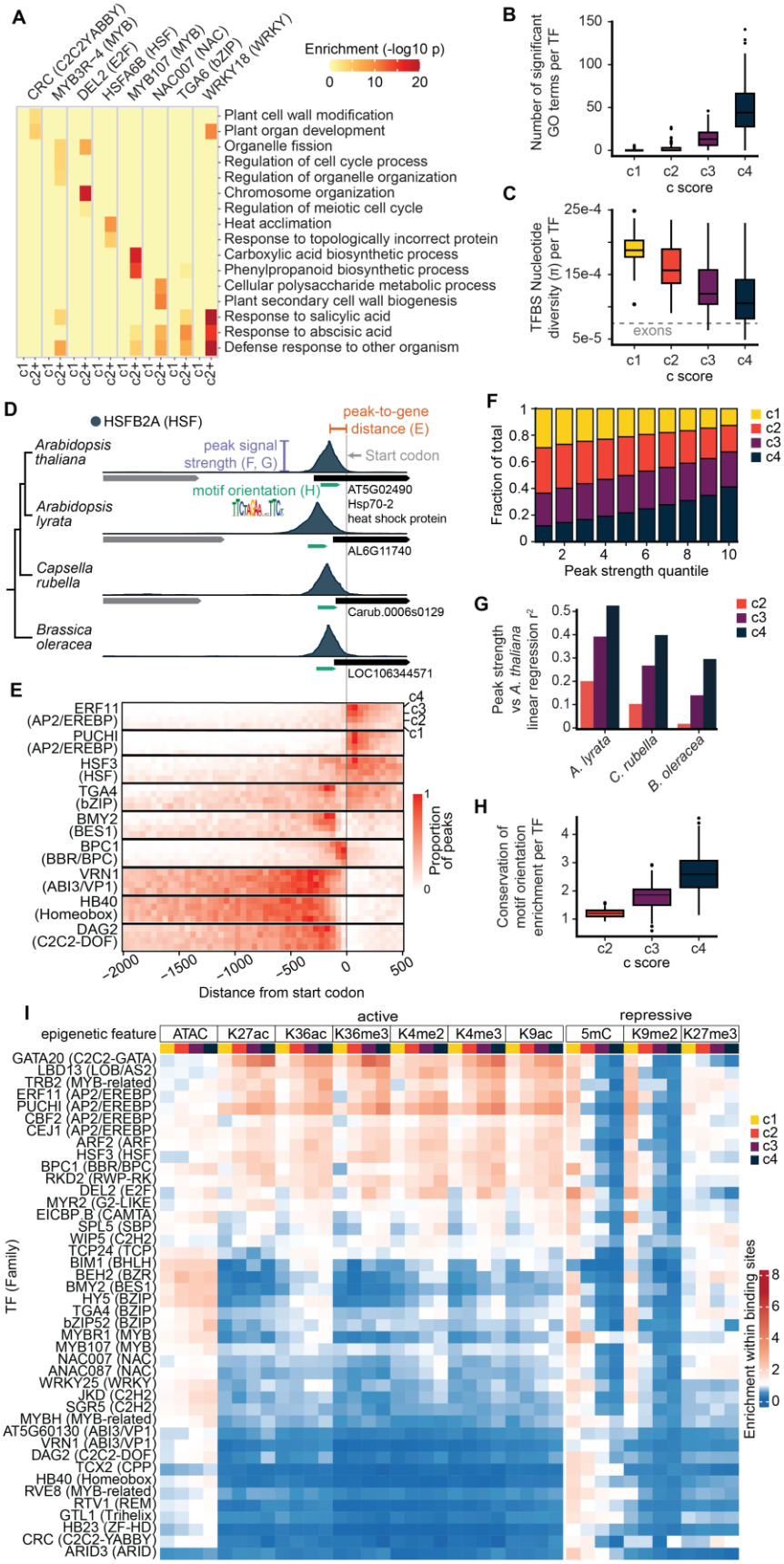


Figure 2. Features and functions of TFBSs across conservation levels. See also Figures S2 and S3 and Tables S1, S2, and S3. Boxplots show median +/- interquartile range. (A) GO term enrichment of non-conserved target genes (c1) in comparison with conserved target genes (c2+). (B) Number of GO terms significantly enriched in target genes across 244 TFs. (C) Average nucleotide diversity around TFBSs of 244 TFs in the 1001 *A. thaliana* genomes. (D) Overview of DAP-seq peak features measured at different c scores. (E) Heatmap of TFBS frequencies for example TFs, relative to the start codon stratified by c score. (F) Stackplot showing the relationship between peak strength in quantiles and the fraction of total peaks at each c score. (G) Correlation of peak strength in *A. thaliana* vs. other species. (H) Conservation of motif orientation in other species relative to *A. thaliana* across 244 TFs. (I) Enrichment of epigenetic markers of target genes from a subset of TFs tested.

We hypothesized that TFBSs with shared functional roles across species in the Brassicaceae family would also be subject to constraint on sequence evolution in the *A. thaliana* accessions sequenced as part of the 1001 Genomes Project[34]. We found evidence of negative selection pressure acting to preserve the sequences immediately underlying peak summits, and decreasing nucleotide diversity (π) at peaks with higher c scores (**Figure 2C**), with the nucleotide diversity underlying c4 peaks being significantly lower than c1, c2, and c3 peaks (t-test, $p < 0.05$), and nearly as low as that found within protein-coding exonic regions.

Conservation of TFBS location, strength, and orientation

As conserved binding sites are under negative selection, we investigated whether other quantifiable properties of binding sites, including location relative to genes, binding strength, and motif orientation, also show conservation (**Figure 2D**). Previous studies, including our own, have demonstrated that different TF families have distinct distributions of binding sites relative to transcription start sites[5,35]. We were interested to further explore these preferred location patterns at different conservation levels, specifically considering that the more conserved binding sites showed evidence of enrichment for CREs. Because transcription start sites are not comprehensively annotated in most plant genomes, we analyzed the distribution of TFBS locations relative to target gene start codons. In the four species Brassicaceae dataset, we observed TF families showing strong preference for binding sites immediately upstream of start codons (e.g. bZIP, BBR/BPC, BES1), families with strong preference downstream of start codons (e.g. AP2/EREBP), and others showing a more diffuse pattern across promoter regions (e.g. C2C2-DOF, ABI3/VP1, Homeobox) (**Figure 2E**, **Figure S2A**). In all three groups, the distinct patterns of TF binding typically become more enriched among TFBSs with higher c scores (**Figure S2B**). Finally, a common trend across all TF families is that binding sites are rarely found overlapping the start codon itself. These TF binding distributions and preferences are apparent in the genomes of all four Brassicaceae species (**Figure S2C**).

We were surprised to observe that 84 TFs from 17 families have more than a third of their c4 TFBSs in coding sequences (**Table S5**). This included all members of the AP2/EREBP family with for example 73.6% of the 2,811 c4 sites for ERF11 and 90.3% of 1,523 c4 sites for PUCHI found in coding sequences (**Figure 2E**). These AP2/EREBP c4 coding sequence binding sites were associated with functionally relevant target genes. For example, ERF11 target genes with c4 binding sites in coding sequences showed strong enrichment for “response to abiotic stimulus” ($p=7.6e-10$), “ethylene-activated signaling pathway” ($p=1.5e-4$), “response to water deprivation” ($p=2.1e-2$). Moreover, the AP2/EREBP TF family members frequently targeted themselves and other family members through these c4 coding sequence TFBSs, with 48% of AP2/EREBP TFs containing an AP2/EREBP TFBS within their own coding sequence. This interfamilial self- and cross-regulation occurred 4.45-fold more frequently than a permuted background ($p < 0.001$; **Figure S2D**), reflecting a known feature of the AP2/EREBP family[36]. Together, these observations support the conclusion that the AP2/EREBP and other transcription factor families may depend on TFBSs within coding sequences for regulatory functions.

We next hypothesized that high affinity TFBSs are enriched for the most important functional TF-target gene interactions. Peak strength, measured as fold-enrichment of signal over background, varied by three orders of magnitude and was reproducible between replicates

(**Figure S2E**). We binned *A. thaliana* DAP-seq peaks into deciles by peak strength and observed that the highest decile contained 2.8-fold more c4 peaks than the lowest decile (hypergeometric test $p < 0.001$) (**Figure 2F**). We also found that peak strength in *A. thaliana* is predictive of peak strength in other Brassicaceae species, with higher correlations for the most conserved sites (**Figure 2G**, **Figure S2F**). In more distantly related species, the same trend is still apparent, though the predictive value of *A. thaliana* peak strength is reduced (**Figure S3A**). These results show that while high affinity TFBSs in one species are predictive of conservation, the binding affinity at individual target genes may evolve rapidly.

TFs with non-palindromic DNA binding motifs have a defined binding orientation relative to each of their target genes. We hypothesized that a binding site that is conserved across species is often derived from a single original DNA sequence and would therefore tend to retain the same motif orientation as the ancestral sequence. Our analysis revealed a positive correlation between c score and preservation of binding orientation across Brassicaceae (**Figure 2H**), with c4s showing a 1.4- and 2.1-fold enrichment over c3s and c2s, respectively, among TFs with significant conservation of binding site orientation (binomial test $p < 0.01$). Additionally, certain animal and plant TFs are known to require a specific motif orientation to function[38,39]. We found evidence of previously described preferred binding orientations for the BBR/BPC family TF BPC1[39], as well as for members of the C2H2, MYB, and Trihelix families for which a preferred binding orientation has not been reported. This trend strengthens with higher c scores (**Figure S3B**), supporting a role for a specific motif orientation in the regulatory function for these TFs

Overall, the observed conserved patterns and trends in TFBS location, binding affinity, and orientation among sites with higher c scores suggest that these conserved sites are under selective pressure across species and represent functional CREs.

The TF epigenomic environment is predicted by preferred binding location

Epigenomic properties influence CRE activity in different tissues and cell types, as chromatin accessibility[40], histone modifications[41], and DNA methylation[42] can activate or repress gene expression in plants. To determine to what degree the conserved binding sites for each TF overlapped with either active or repressive chromatin marks, we selected published data sets from *A. thaliana* seedlings profiling chromatin features including accessible chromatin regions (ACRs)[43], activating histone 3 (H3) lysine (K) modifications H3K27ac[44], H3K36ac[45], H3K36me3[46], H3K4me2[47], H3K4me3[47], and H3K9ac[48], repressive H3 modifications H3K9me2[47] and H3K27me3[49], and DNA methylation[42].

Conserved binding sites for most TFs showed enrichment in ACRs with some of the strongest enrichment seen for TFs in the bHLH, bZIP1, and TCP families (**Figure 2I**, **Figure S3C**). Since ACRs are known to be concentrated near transcription start sites[43], we were interested to see whether more distal TFBSs were also enriched in ACRs. We split *A. thaliana* promoter regions into 250 bp bins and found that conserved TFBSs in all promoter bins were enriched in ACRs, while c1 TFBSs were not (**Figure S3D**). Many TF families also showed enrichment of TFBSs at active histone marks, including members of the Trihelix, BBR/BPC, and C2C2-GATA families, with greater enrichment for more conserved TFBSs. However, TFs in several families (e.g. C2C2-DOF, Homebox, and MYB-related) were depleted for active chromatin marks at all c scores. TF families could be grouped into three broader levels of

enrichment for active histone marks. These groups appeared to largely coincide with the TF groupings defined above based on preferred TFBS locations (see **Figure 2E** and **Figure S2A**). To quantify the relationship between preferred TF binding location and preferred epigenetic environment, we plotted the median location of c4 binding sites for each TF against their enrichment for each active chromatin mark and found strong correlations for most marks, including H3K36ac (Spearman r=0.84), H3K9ac (r=0.89), and H3K4me2 (r=0.96) (**Figure S3E**). Constructing similar plots for ACR enrichment revealed that TFs enriched for binding immediately upstream of genes showed the highest ACR enrichment (**Figure S3D**).

In contrast to active chromatin marks, the silencing marks of 5mC DNA methylation and H3K9me2 showed strong enrichment at c1 sites and depletion at c3 and c4 sites for most TFs. These patterns suggest that these two marks, associated with inactive heterochromatin[50], may silence spurious transcription at non-conserved *A. thaliana* TFBSs. We observed more variable patterns observed for the repressive mark H3K27me3, which is consistent with its regulation by a distinct pathway[51] and its dynamic, multifaceted roles in development and environmental responses[41].

In summary, our findings reveal that conserved TFBSs overall are enriched for epigenomic marks indicative of active CREs and depleted for repressive marks, supporting their role as functional regulatory sites. Furthermore, the level of enrichment for specific activating marks varies across TF families and is strongly correlated with binding location preference, which underscores a previously unrecognized association between certain TF families and distinct chromatin states. Overall, this highlights the power of using conserved TFBSs to identify CREs and TF target genes.

TF binding specificity is widely conserved across plants

Having thus far focused exclusively on TFs encoded by *A. thaliana*, we next examined orthologous TFs from other plant species to investigate divergence of their preferred motifs and genome-wide binding properties. We selected 35 *A. thaliana* TFs representing different families and motif types and identified the closest ortholog of each in five additional species for a total of 175 orthologous TFs. Initially we used multiDAP to map the binding of 60 of these 175 orthologous TFs across the genomes of all ten plant species (see **Table S1**) and observed both high correlation in binding profiles among orthologous TF pairs (**Figure 3A**) and nearly identical motifs (**Figure 3B**) even for the most distantly related TFs from *O. sativa*. We then verified that this finding can be generalized more broadly across plant TFs by mapping the binding of the full set of 175 orthologs in duplicate multiDAP experiments with genomic DNA from the two most divergent species, *A. thaliana* and *O. sativa*. We again observed high correlation within all groups of orthologs (**Figure 3C**). Of the total 175 orthologous TFs tested, 108 performed well enough to generate high quality binding maps in multiDAP experiments: 21 from *A. lyrata*, 22 from *C. rubella*, 21 from *B. oleracea*, 21 from *S. lycopersicum*, and 23 from *O. sativa*. Although slight variations existed between the genome-wide binding signal of some pairs of orthologous TFs, correlations remained high (mean signal correlation=0.89, standard deviation=0.1, n=1431). The slightly lower correlation between orthologs compared to the 0.94 correlation previously measured between direct technical replicates of the same TF can at least partially be attributed to differences in performance of individual TFs in the multiDAP assay (**Figure S4A**). Furthermore, we did not observe any trend between signal correlation and either evolutionary

distance or TF amino acid sequence similarity (**Figure S4B**). We found high signal correlation even between pairs of *A. thaliana* and *O. sativa* orthologs, which on average share only 52% amino acid identity (mean signal correlation=0.86, standard deviation=0.09, n=195). This demonstrates that the binding of *A. thaliana* TFs in a different plant genome can serve as a good predictor for the binding of the corresponding native TFs and suggests that divergence of TF binding specificity may not be the primary driver of regulatory evolution.

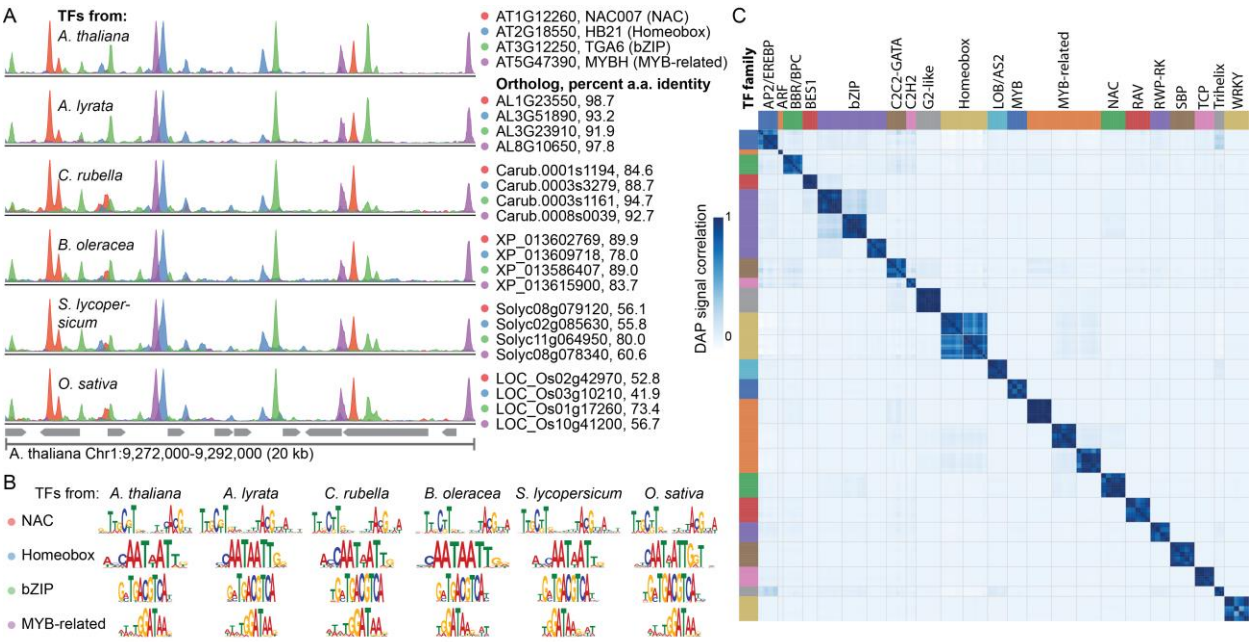


Figure 3. Comparison of binding profiles between *A. thaliana* TFs and the TF orthologs from other species. See also Figure S4. (A) An example 20 kilobase region of the *A. thaliana* genome shows similar TF-binding profiles of four *A. thaliana* TFs and their corresponding orthologs from five other species. Numbers after TF IDs indicate percent amino acid identity compared to the *A. thaliana* ortholog. (B) The similarity of consensus motifs from TFs in panel A (C) All-vs-all genome-wide correlations of DAP-seq binding profiles between TFs from six species. Gray grid lines delineate individual groups of orthologous TFs and colored bars indicate TF families.

A multi-species multi-tissue single nuclei gene expression atlas

As conserved TFBSs have the genetic and epigenetic hallmarks of functional CREs, TF binding at these sites is most likely to elicit a change in expression of the associated target gene. We therefore explored whether tracking the expression of each TF's conserved target genes would allow us to chart TF activity across cell and tissue types. To do so, we generated snRNA-seq atlases for multiple tissues (seedling root and shoot, mature leaf, and flower bud) of all four Brassicaceae species using the 10x Genomics Chromium platform. Each atlas included multiple biological replicates, for example, our *A. thaliana* seedling atlas included root (n=4), shoot (n=2) and whole seedling (n=4) transcriptomes. Across the species and tissue types, we profiled 163,982 high-quality single nuclei transcriptomes, with a median of 899 unique transcripts and 742 genes per nucleus (**Table S6**). Using published reference atlases[52–54] and marker gene lists[55], we assigned our *A. thaliana* nuclei to 61 unique cell type and developmental stage labels (**Figure 4A, Figures S5A-S5C)**, including a small number of seedling labels that were not present in published protoplast-based atlases, but were supported by integration of an

external seedling root nuclei dataset[56]**.** Cross-species data integration with SAMap[57] was used to lift-over labels from the *A. thaliana* atlases to the other three species, (**Figure 4B)** resulting in 57 labels with at least 10 nuclei from all four species (**Data File S4)**.

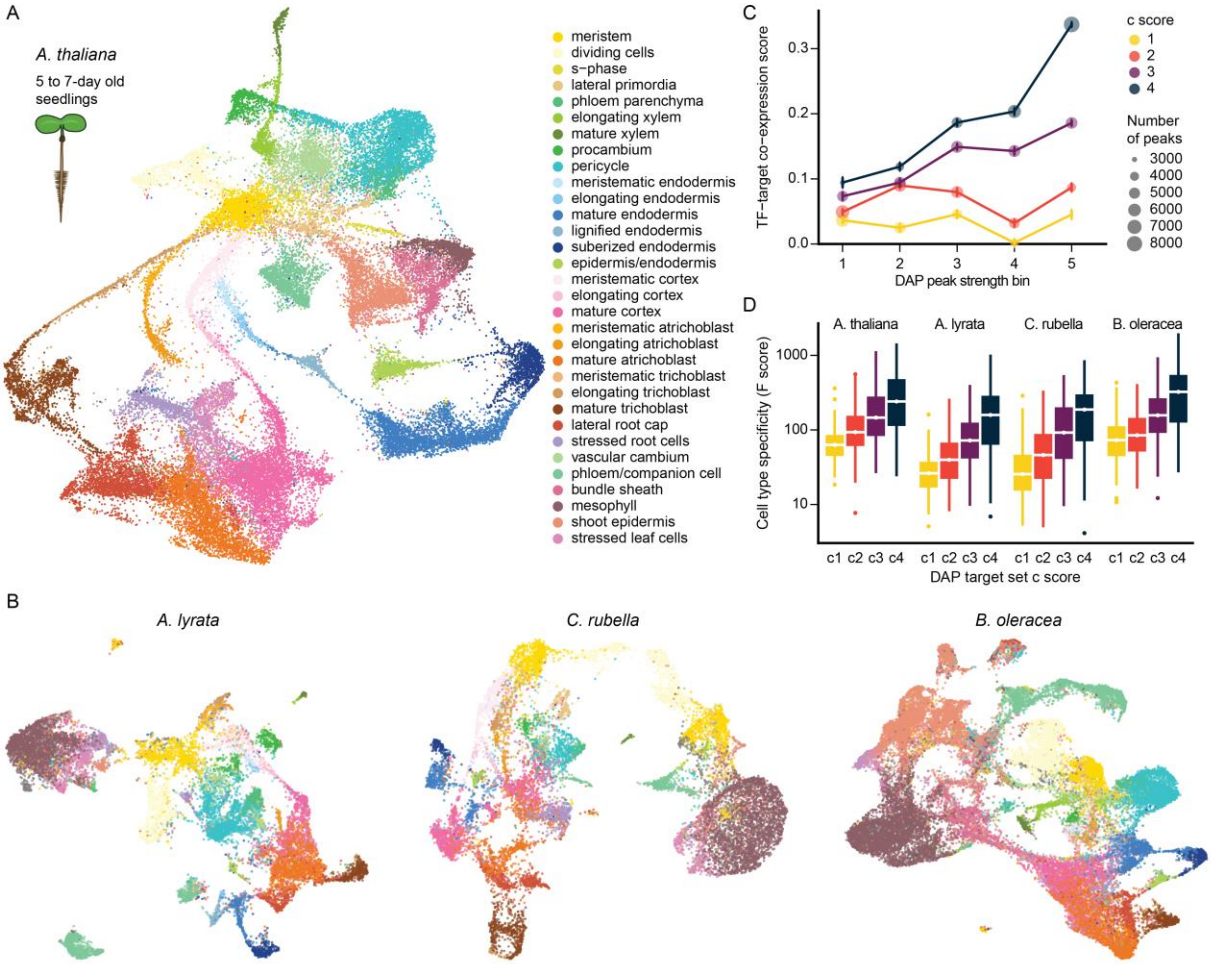


Figure 4. Expression patterns of TFs and target genes across snRNA-seq seedling atlases. See also Figure S5 and Table S6. (A) UMAP representation of the transcriptomes of >45,000 single *A. thaliana* seedling nuclei corresponding to 32 labeled cell types. (B) UMAP representations of corresponding data from three Brassicaceae relatives, with cell type labels transferred from the *A. thaliana* atlas. Colors correspond to labels in panel A (C) TF-target co-expression across all nuclei in the *A. thaliana* seedling atlas as a function of multiDAP TF binding affinity and multiDAP peak conservation score. Y-axis represents the mean fold-change between true correlation of normalized TF expression with target expression vs. correlation with a matched control target gene. Point size represents the number of DAP peaks considered, and error bars represent standard error of the mean. (D) Boxplots showing the distribution of cell type specificity scores as a function of target conservation for 244 TFs in each of the four Brassicaceae seedling snRNA-seq atlases. Cell type specificity was defined as the F-statistic from an ANOVA test of the extent to which per-nucleus target gene expression varied systematically across labeled cell types. All distributions within a species were significantly different from each other (FDR < 0.05) using paired one-sided t-tests. Boxplots show median +/- interquartile range.

Conserved target gene expression illuminates known and new TF biology

We first tested the simplifying assumption that higher TF expression results in increased TF protein levels, leading to stronger activation or repression of target genes. To do so, we

leveraged the *A. thaliana* seedling atlas to calculate a co-expression score for each TF-target gene pair, measuring how well expression of the 244 *A. thaliana* TFs predicted expression of each target gene. As TF binding affinity can determine how consistently a given TFBS is bound, we binned all target genes by both DAP-seq peak strength and c score. We found that target genes with the most conserved binding sites (c4) consistently exhibited the highest co-expression with their respective TFs, regardless of peak strength (**Figure 4C**). Interestingly, peak strength was only a strong predictor of co-expression when binding sites were conserved across three or more species. These findings support a paradigm in which conserved TFBSs are more likely to impact target gene expression, and binding affinity can further predict a TF's regulatory strength.

While we observed enriched co-expression between TFs and conserved target genes, accurately capturing fluctuations in the expression of individual genes across individual nuclei is challenging due to the sparse nature of snRNA-seq data. Additionally, TF function is often regulated by mechanisms beyond transcription, including post-transcriptional modifications[58], presence of hormones or other cofactors[59], and dynamic changes to chromatin state[60]. Thus as a proxy for TF activity in each cell, we averaged expression of all target genes relative to matched background genes[61], independent of TF expression. As a first test, we asked whether these summarized target gene expression scores varied systematically across labeled cell types, as would be expected from active CREs involved in coordinating growth and development. We found that c4 target gene sets indeed were associated with significantly higher cell type specificity in every species (**Figure 4D**).

We then directly quantified the enrichment or depletion of expression of each TF's c4 target genes in each cell type (**Table S7**). Our results yield an expansive map of TF regulatory roles in cell growth, differentiation, and function, often reflecting known regulatory biology (**Figure 5A**). Examples include the role of WRKY TFs in stress response and senescence[32], with target gene expression enriched in senescing leaf and root cells, or enrichment of NAC007/VND4 target gene expression in mature xylem cells[62]. Interestingly, while NAC regulation of secondary cell wall biosynthesis during xylem vessel formation has been clearly documented in *A. thaliana*[63], extending our analysis to examine the ultra-conserved target sets identified by the 10-species multiDAP assay reveals that this regulatory relationship can be traced back across long evolutionary distances (**Figure S5D**). The strongest xylem-specific expression enrichment was associated with NAC007 c9 target genes (p=1.8e-204, Cohen's d=4.70) and among the 214 targets with a c score of 8 or above, 58 were markers for elongating or mature xylem-specific expression (10.2-fold enrichment over background; **Data File S4**). Accordingly, a TF's conserved target genes often retain the same cell type specificity across species. For example, MYB107 target genes are strong and specific markers of late-stage endodermis cells (including lignified and suberized endodermis) in seedling atlases of all four species (**Figure 5B**). Similarly, AP2/ERF TF RAP2.9 target gene expression marks epidermal leaf cells in all four species (**Figure 5C**), consistent with the role of close paralog RAP2.4 in leaf cuticular wax biosynthesis[64]. The target genes of ASR3 (Trihelix) show petal-specific expression in all four species, consistent with the involvement of Trihelix family members in petal initiation[65] (**Figure 5D**).

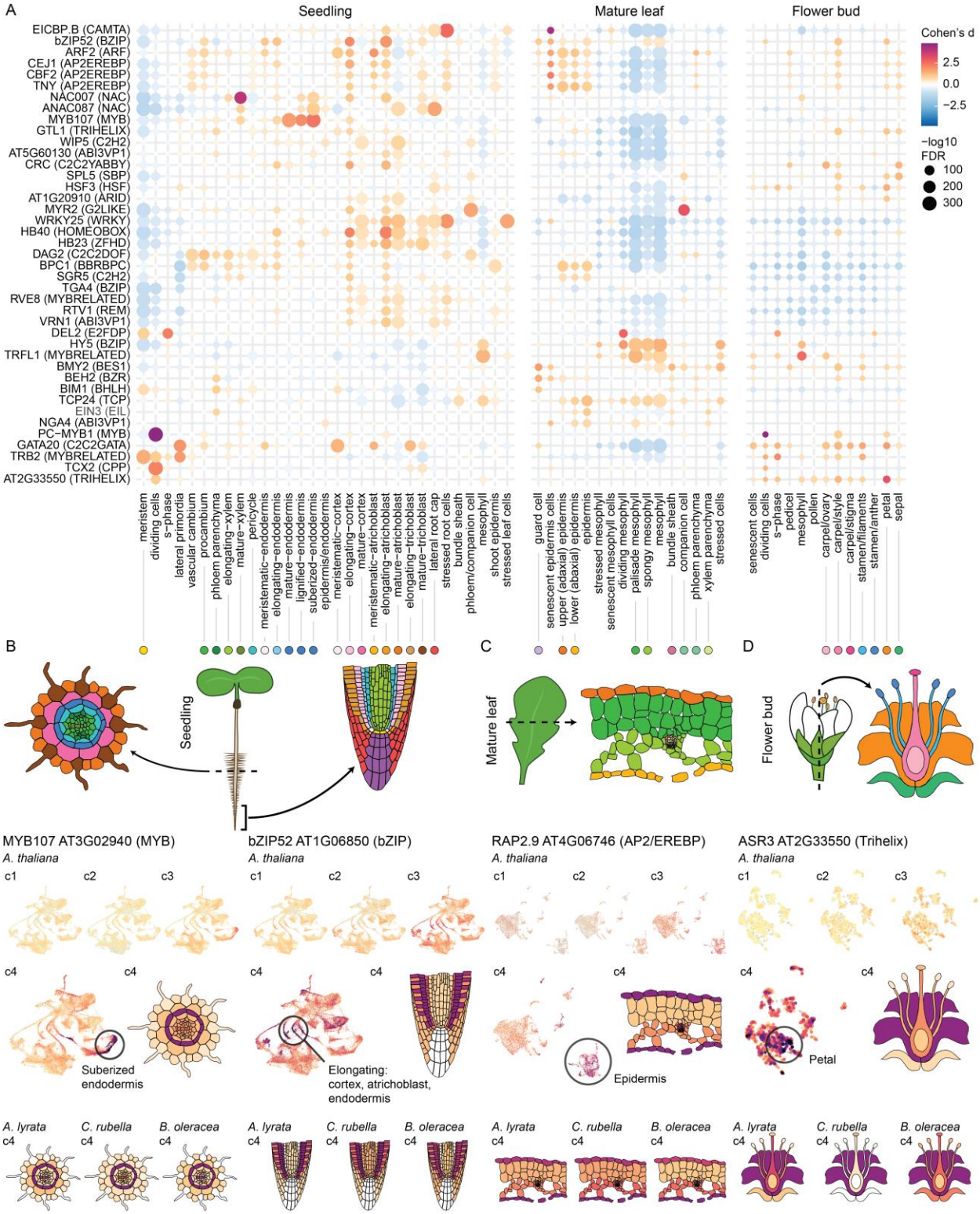


Figure 5. A multi-tissue regulatory map from cell type-specific enrichment of conserved target gene expression. See also Figure S5 and Table S6. (A) Enrichment or depletion of c4 target gene expression for representative TFs (y-axis) across labeled cell types in *A. thaliana* seedling, leaf, and flower bud snRNA-seq atlases (x-axis). Dot size represents the FDR-corrected -log10 p-value from a Wilcoxon rank sum test of summarized target expression scores for each cell type label vs. all other cells, and color represents the effect size of the expression enrichment (warm colors) or

depletion (cool colors). Enrichments with FDR-corrected p-value < 0.01 are shown. (B, C, and D) show detailed plots for one example TF in each of four different anatomical regions. In (B), the spatial arrangement of mature root cell types is shown in a graphic at the top, with cell types colored according to dots annotating cell type labels in the x-axis of panel (A). Target enrichments for TF MYB107 are highlighted in the plots below, starting with a set of *A. thaliana* seedling UMAPs colored by summarized expression scores for MYB107 targets grouped by conservation level (c1/c2/c3/c4). Strong enrichment of c4 target expression in suberized endodermis cells are highlighted with a labeled circle. Average c4 expression scores for each cell type are then projected onto the spatial map of mature root cell types to the right of the c4 UMAP, and projections of c4 scores for the other Brassicaceae are shown below. Following panels use the same structure, but highlight bZIP52 targets in root tip, RAP2.9 targets in mature leaf, and ASR3 targets in flower bud.

Certain TF families mark specific cell types across multiple tissues, for example, G2-like TF MYR2 target gene expression is strongest in the phloem companion cells of both seedling and mature leaf, consistent with previously observed phloem-specific expression of MYR2 and its homolog MYR1[66]. Similarly, mesophyll cells in seedling shoots, leaves, and flower buds are marked by enrichment of HY5 (bZIP) and TRFL1 (Myb-related) target gene expression. Although HY5 is known to be active in mesophyll cells[67], a role for TRFL1 or any close paralogs in mesophyll cells has not yet been characterized. We also observed enrichments reflecting known partnerships between TF families. For example PC-MYB1/MYB3R1 and TCX2 (a close paralog of TSO1) interact to coordinate cell division[68,69] and both show enriched target gene expression in root and floral dividing cells. Finally we found support for suggested but less well-documented associations, such as the role of Group I bZIP TFs in elongation of multiple root cell types[70], reflected here in the enrichment of bZIP52 target genes in elongating cortex, atrichoblast, and endodermis (**Figure 5B**). In total, of 244 TFs assayed, 241 show significant cell-type-specific expression patterns of their c4 target gene sets in at least one labeled cell type, including many previously uncharacterized associations. Integration of cell type-resolved snRNA-seq and multiDAP data provides a comprehensive view of TF activity across plant cell types.

Cell type-specific accessible chromatin overlaps conserved TFBSs

Given the strong cell type-specific expression patterns of many TFs' c4 target genes, we reasoned that conserved TFBSs would be enriched within cell type-specific open chromatin. To explore TFBSs in cell type-specific differentially accessible regions (DARs), we generated a multispecies, multi-tissue snATAC-seq atlas of 74,307 high-quality nuclei from combined root and shoot seedling samples for A. thaliana, and more than 10,000 high-quality nuclei for each of the other species (**Figure 6A, Table S6**). Leaf and the flower bud were represented by 6,000 and 7,284 nuclei on average across the four species (**Figure S6A**). We verified that the pseudo-bulk of our snATAC-seq was consistent with published bulk open chromatin profiles[43] (Pearson r^2 0.58).

We used gene expression profiles from our snRNA-seq atlases to assign cell type labels to our snATAC-seq atlases based on gene accessibility profiles (see **Methods**). We looked for regions that were more open in a given cell type than other cell types, identifying 24,627 of these DARs ($p < 0.01$, regression test) localized close to a gene start codon (-2000 to 500 bp) (**Figure S6B**, see also **Data File S5**). 94% of DARs contained a c2+ TFBS, and 67.3% contained a c4 TFBS. The observation that a large percentage of DARs contain conserved

TFBSs supports a role for chromatin accessibility in regulating cell type-specific TF binding that results in cell type-specific gene expression. This type of relationship is exemplified by a c4 TFBS of DOF5.8, a key regulator of phloem development[71]. This c4 site falls within a phloem companion cell DAR in the promoter of a gene expressed specifically in phloem companion cells (**Figure 6B**). We then quantified enrichment of c4 binding sites for each TF in each cell type's differentially accessible chromatin. For many TFs, we found that c4 TFBSs were associated with differentially accessible chromatin and enriched target gene expression in the same cell type (**Figure 6C, Figure S6C**, see also **Figure 5A**). For example, we and others[72] observed WRKY TF binding enrichment in epidermal and cortex cell DARs, which is consistent with the enrichment of c4 WRKY target gene expression in these same cell types as revealed by our snRNA-seq seedling root atlas. MYB107, a master regulator of suberin deposition in endodermal cells, showed enrichment of both cell type-specific accessible chromatin regions and target gene expression in suberized-endodermis cells. Additional examples include C2C2-DOF TFs[73] enriched in vascular tissues, Homeobox TF ANL2 in epidermal cells[74], TCP3 and TCP24[75] and AT3G11280[76] in leaves, and G2-like MYR2 enriched in phloem companion cells[66].

We also applied this analysis to the other Brassicaceae species and found many examples of c4 TFBS enrichment in cell type-specific accessible chromatin regions shared across species. For example, the defense related TF WRKY40 has c4 TFBSs in all four species in cortex-specific accessible chromatin regions located in the promoter of BAK1-interacting receptor-like kinase 1 (BIR1) (**Figure 6D**), a gene known to mediate growth and defense response in roots[32,77]. Other examples include the TF BPC1 binding sites in an atrichoblast-specific accessible chromatin site in the promoter of the gene BIN2 in mature atrichoblast (**Figure 6D**) which has been implicated in inhibiting atrichoblast differentiation[78,79]. Also a c4 binding site for MYB36 falls in an accessible chromatin regions found in mature endodermis in all four Brassicaceae, consistent with the role of MYB36 as a master regulator of the formation of the Casparian strips in the endodermis[80] (**Figure 6D**). In summary, we found that integrating our conserved TFBS atlas with either snRNA-seq or snATAC-seq data presents complementary views of cell type-specific TF activity.

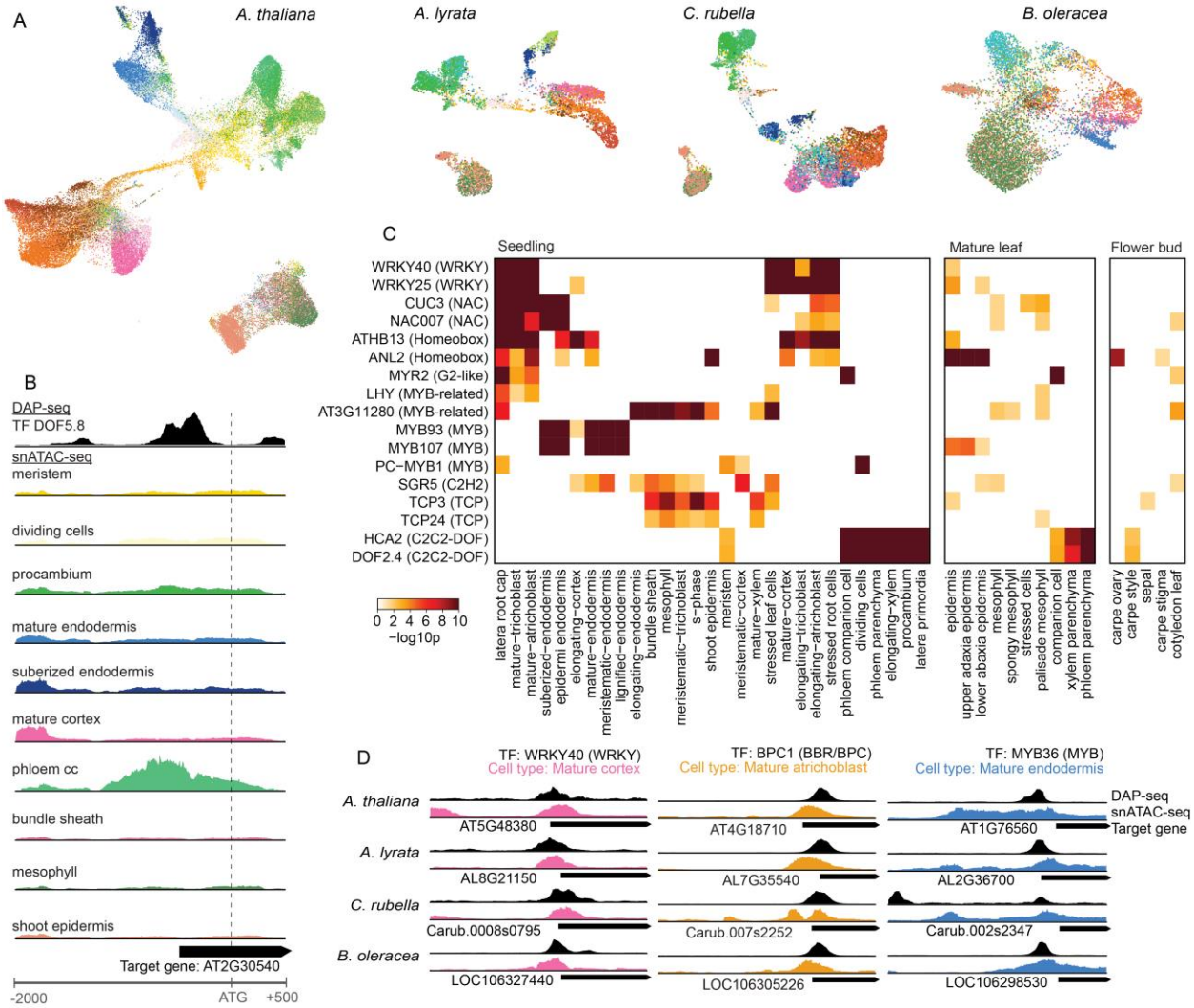


Figure 6. Integration of snATAC-seq atlas with conserved multiDAP sites. See also Figure S6 and Table S6. (A) Whole-seedling snATAC-seq atlases of four Brassicaceae species. Colors match snRNA-seq atlas in Figure 4. (B) Example of an *A. thaliana* c4 binding site for DOF5.8 in differentially accessible open chromatin regions in phloem companion cells. (C) Enrichment of c4 TFBSs in differentially accessible regions by cell type (D) Examples of c4 TFBSs in open chromatin regions conserved across all four species.

Resolving TF activity in response to stress hormones over time and across cell types

Aside from their roles in development and cell type differentiation, TFs orchestrate plants' highly evolved responses to a broad range of environmental cues. To explore these response networks, we profiled TF activity during treatment with two agriculturally important and ancient plant stress hormones: abscisic acid (ABA), a master regulator of responses to abiotic stressors including drought, cold, and soil salinity[81], and salicylic acid (SA), a key activator of plant defense responses to infection[82].

We exposed 5-day-old *A. thaliana* seedlings to ABA, SA, or a mock treatment, collected roots and shoots separately at multiple time points, and performed bulk RNA-seq to identify differentially expressed genes at each timepoint **(Figures 7A** and **7B**). Using our multiDAP atlas to track the response of conserved target genes, we inferred hormone-induced activity for many TF families that have known roles in ABA or SA responses. TF families with ABA-responsive c4

target genes included bZIP[83], AP2/EREBP[84], NAC[85], WRKY[32,86], MYB-related[87], BES[88], and MYB[23], while SA-responsive TFs included members of the bZIP[31], NAC[89], WRKY[90], MYB-related[91], and MYB[92] families (**Figure 7C**). Two distinct clades of the bZIP TF family are known to drive responses for these hormones: ABA triggers a signal transduction cascade leading to the phosphorylation of ABA-responsive element binding proteins (AREB bZIPs)[83], while SA activates TGACG-binding factors (TGA factor bZIPs)[31]. Accordingly, our analysis revealed strong upregulation of AREB3 target genes in ABA-treated seedlings and elevated expression of TGA factor target genes in SA-treated seedlings. We also observed instances of the two hormones having opposite effects. For example, WRKY family target genes were generally down-regulated in response to ABA but up-regulated in response to SA, consistent with previous observations that the WRKY family includes both repressor and activator TFs with roles in ABA and SA response[32,90]. These dynamics may contribute to the antagonism previously observed between ABA and SA response networks[93].

We also found support for previous observations that the strongest ABA-responsive genes are often targeted by more than one TF family[94]. For example, many genes consistently upregulated by ABA treatment are targets of both the AREB bZIP TFs, as well as the dehydration-response element B1a (DREB1a) AP2-EREBP TF, a central regulator of ABA-induced drought and cold responses[23]. Of the 83 genes that are c4 targets of either AREB3 or DREB1a and were upregulated by ABA in both seedling tissues at all tested timepoints, 21 genes (25.3%) were target genes of both AREB3 and DREB1a. This is significantly higher than expected by random chance (94-fold enrichment over shuffled background, $p < 0.001$). These 21 ABA-induced co-targeted genes were enriched for GO terms such as *response to abscisic acid* (10 genes; FDR=2.75e-09) and *response to water deprivation* (11 genes; FDR=1.64e-11). More broadly, we observed a clear trend wherein the promoters of genes showing the strongest ABA-induced expression contained more binding sites for ABA-responsive TFs (**Figure S7**). Combined with previous observations[94], these findings support a model where independent binding of distinct TFs to the same gene promoter may facilitate consistent and sustained gene expression across tissues and time.

We next examined whether the responses observed with bulk RNA-seq could be localized to specific cell types by performing snRNA-seq on hormone-treated and mock-treated samples from the two-hour time point (**Figure 7C**). In general we found that the summarized response of a given TF's target genes was consistent across cell types, a pattern that was repeated for both ABA and SA treatments. Moreover, individual target genes showed similar responses across cell types as measured by average cross-cell type correlation coefficients for target genes compared to correlations obtained after shuffling target gene IDs (paired t-test p-value < 2.2e-16). One exception to this was the repression of WRKY family target genes in response to ABA, which we observed exclusively in the outer layer of the plant (trichoblast, atrichoblast, cortex, lateral root cap), where WRKY family activity was strongest in untreated samples (see **Figures 5A** and **6C**). This result suggests that ABA triggers WRKY target gene repression specifically in the cell types where the TF family is basally active under non-stress conditions in the root. Repressed WRKY target genes were enriched for involvement in biotic defense responses (GO term *response to external biotic stimulus* p=1.10e-16**)**, consistent with evidence that ABA can repress defense responses in *A. thaliana*[93]. The opposite pattern was observed in response to SA, with activation of WRKY target genes in all cell types except those

with basal WRKY activity. SA-induced target genes were also associated with defense responses (GO term *response to external biotic stimulus* p=2.59e-10), suggesting that SA activates a WRKY-mediated defense response root-wide to combat microbial infection. Overall, when combined with multiDAP data, both bulk RNA-seq and snRNA-seq yield complementary views of TF activity in response to hormone-induced regulation of plant stress responses.

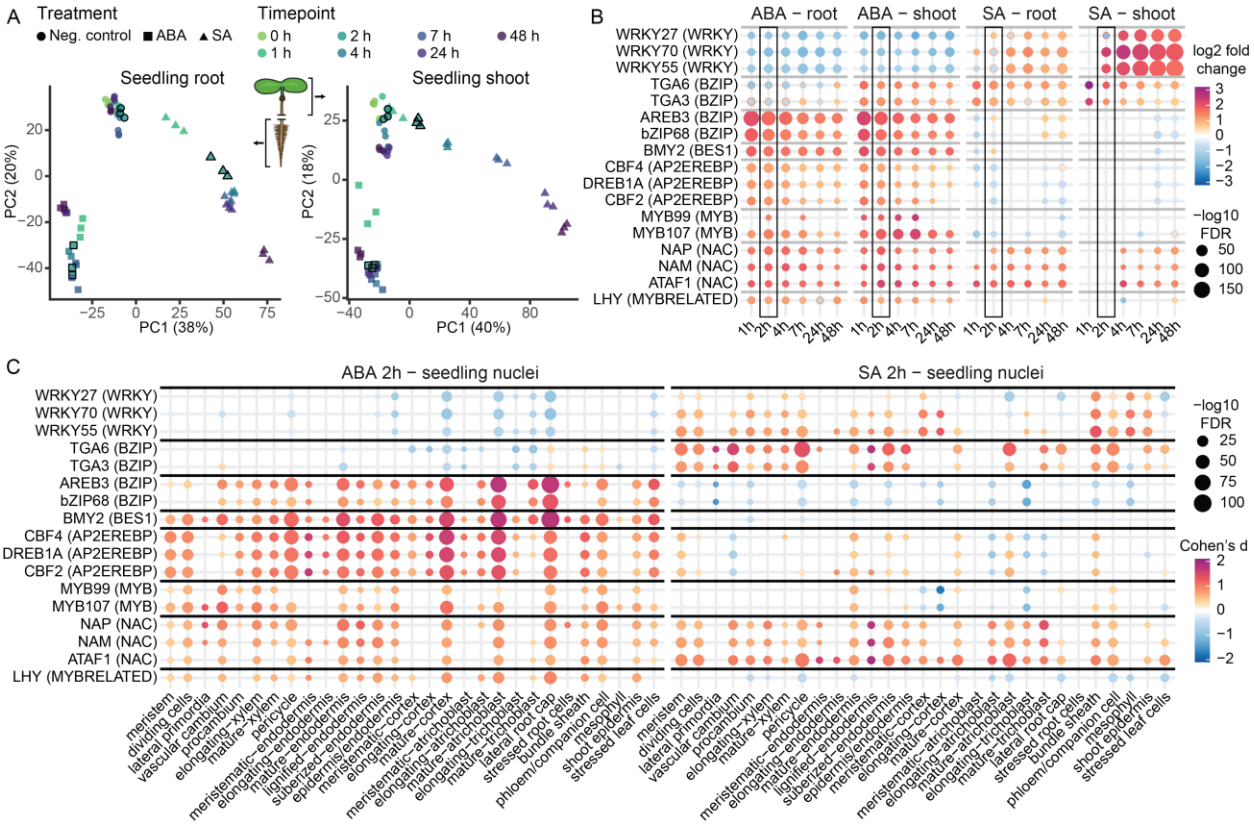


Figure 7. Hormone-induced TF activity across temporal and spatial scales. See also Figure S7. (A) PCA plot of genome-wide gene expression during hormone treatment time courses assayed by bulk RNA-seq of *A. thaliana* seedling root and shoot samples. Treatment regimens are indicated by shape (Neg. control = untreated, ABA = abscisic acid, SA = salicylic acid), and time point is indicated by color. Samples from the 2-hour time point are outlined in black to denote that they were also assayed by snRNA-seq. (B) Enrichment of hormone-induced TF target gene expression shifts assayed by bulk RNA-seq. Dots represent significant enrichment of hormone-upregulated (orange) and/or hormone down-regulated (blue) genes among the c4 targets of each representative highly-responsive TF. Enrichments from the 2-hour time point are outlined in black. Overlapping orange and blue dots (e.g. LHY in ABA-root at 24h) represent simultaneous up- and down-regulation of different subsets of target genes. (C) Enrichment of hormone-induced TF target gene expression shifts in each seedling cell type at the 2-hour time point assayed by snRNA-seq. Differences between summarized c4 target scores in hormone-treated versus untreated nuclei within the same labeled cell type are represented by dot size (FDR-corrected p-value from Wilcoxon rank sum t-test) and dot color (magnitude measured by Cohen's d).

Discussion

Our study presents a comprehensive atlas of conserved TFBSs across flowering plants, providing a foundational resource for understanding gene regulation in development and environmental responses. Using multiDAP, we generated a high-resolution, cross-species map of conserved TFBSs and demonstrated that they are enriched for CREs, as indicated by their

associations with regions under purifying selection, active chromatin features, and specific target gene functions. Our results suggest that TFBSs cluster in preferred promoter "neighborhoods", including some TF families with a high rate of conserved binding sites within coding sequences, and that these neighborhoods are associated with distinct epigenomic environments. The minimal differences in motif and binding site preferences that we observed between TF orthologs, even across highly divergent species like *A. thaliana* and *O. sativa*, suggest that CRE variation is a primary driver of regulatory evolution. By integrating this atlas with bulk RNA-seq, snRNA-seq, and snATAC-seq datasets, we demonstrated that conserved target genes are powerful predictors of TF pathway functions. This allowed us to map known and novel TF functions across 61 distinct cell types shared across *A. thaliana* and other Brassicaceae species.

While our conserved TFBS atlas offers a resource for unraveling gene regulatory networks, further work is required to fully understand the complexities of gene regulation. As the promoter is the fundamental unit of gene regulation, we chose to focus our study on regions near protein-coding genes. However, our plant multiDAP atlas also includes distal binding sites which may contribute to gene regulation and their future integration will be important to fully capture the gene regulatory landscape. Secondly, co-binding of multiple TFs within a family due to overlapping specificities complicates the definitive attribution of binding events and regulatory outputs to specific TFs. Nevertheless, we observed many cases where conserved TFBSs for specific TF family members known to play roles in a developmental or environmental response program are highly predictive of relevant gene expression patterns. Finally, our focus in this study was to explore the TFBSs and target genes conserved between species. However, in our bacterial multiDAP study[95], as well as in a recent DAP-seq study of variation between two maize strains[13], some binding sites found only within a single strain were found to result in functional variations in expression and phenotype. The highly scalable multiDAP approach is ideal for capturing and analyzing the impact of such evolutionary changes in close relatives as well as divergent species.

Future studies combining our atlas with genetic, epigenetic, functional, and comparative genomics approaches will help to further elucidate the complex regulatory interactions underlying gene expression in plants. Expanding this atlas to include more species, developmental stages, and environmental conditions will deepen our understanding of TF networks and their evolution, with broad impacts on the fields of plant biology, synthetic biology, evolutionary studies, and agricultural biotechnology.

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

We also thank the following individuals: Chris Beecroft and Tatiparthi Reddy for assisting with submission of raw data files and curation of associated metadata. Laura Gerard for illustrating plant cartoons. Ben Cole, Jacqueline Humphries, Axel Visel, and Zhuzhu Zhang for editing the text and providing comments. Missy Fix, Nahla Bassil, and Kim Hummer at the USDA ARS NCGR for helping to identify an accession of *F. vesca* and providing germplasm. Joseph

Edwards and the Sundar lab at UC Davis for providing *O. sativa* germplasm. Jesse Schartner at the USDA ARS VCRU for providing *S. tuberosum* germplasm. Melanie Harrison and Tiffany Fields at the USDA ARS PGRCU for providing *S. bicolor* germplasm.

Author contributions

Conceptualization LAB, RCO; Methodology DD, LAB, PW, RCO, YZ; Validation AMC, LAB, SIG, RCO; Formal Analysis AG, AMC, LAB, LY, RCO, SIG; Investigation CC, ES, GH, NG, PW, YZ; Writing, AMC, LAB, PW, RCO, SIG, YZ; Visualization, AMC, LAB, LY, SIG; Supervision CGD, IB, LAB, RCO, SIG, YY

Declaration of interests

The authors declare no competing interests.

Data and Code Availability

Note: All raw data and supplemental data files will be made publicly available pending publication.

Scripts and example data files used for DAP-seq and single nuclei analyses are available in a git repository at https://code.jgi.doe.gov/LBaumgart/plant-multidap-and-single-cell.

NCBI:

- Seurat objects for snRNA-seq and snATAC-seq
- fastq for bulk RNA-seq
- fastq for DAP-seq
- fastq/bam for snRNA-seq
- fastq/bam for snATAC-seq

Zenodo:

- Supplemental data files
- narrowPeak for DAP-seq

Raw fastq sequence data files have been submitted to the National Center for Biotechnology Information under BioProject no. <PRJNA-TBD>, for which curated metadata are managed in GOLD (https://gold.jgi.doe.gov/).

Supplemental data files are available in the Zenodo repository under <DOI-TBD>.

Methods

Plant Germplasm and Reference Genomes

Reference genome sequences and annotations for *A. thaliana*[96], *A. lyrata*[97,98], *C. rubella*[99], *B. oleracea*[100], *F. vesca*[101], *P. trichocarpa*[102], *S. lycopersicum*[103], *S. tuberosum*[104], *S. bicolor* (These sequence data were produced by the US Department of Energy Joint Genome

Institute), and *O. sativa*[105] were downloaded from Phytozome or NCBI RefSeq. Gerplasm was obtained either from ABRC, USDA, TGRC, or inhouse collections. See **Table S2** for a full list of germplasm and reference genome sources.

DAP-seq Experiments

The DAP-seq and multiDAP-seq assays were performed using the same reagents as previously described[5,95,106], but with modifications to optimize sensitivity with all 96-well plate operations taking place on a Hamilton Vantage liquid handler. Genomic DNA (gDNA) was extracted from each of the ten plant species using a CTAB protocol, after which gDNA fragment libraries were constructed with an average shearing size at 150 bp. Different species were labeled by unique Illumina i5 index adapters. Upstream of the DAP-seq assay, fragment libraries were amplified for 10 cycles to remove any DNA modifications.

The coding sequence (CDS) of plant transcription factors were cloned into the pIX-HALO-PaqCI vector, a modified vector optimized for chewback cloning but resulting in identical assembled plasmid sequences as the original pIX-HALO[5,106]. The HaloTag-fused TF CDSs were amplified by the primers pIX-Halo-T7-fwd (5'-GTGAATTGTAATACGACTCACTATAGGG) and pIX-Halo-AfterPolyA-rev (5'-CAAGGGGTTATGCTAGTTATTGCTC) using KAPA HiFi HotStart ReadyMix (Roche KK2602). Resulting PCR products were cleaned up using the Mag-Bind TotalPure NGS Kit (Omega M1378-02), and correct amplicon sizes were verified on a Fragment Analyzer system (Agilent) using the HS NGS Fragment Kit (Agilent DNF-474-1000). A total of 10 µL at a minimum concentration of approximately 100 ng/µL of each PCR amplicon was used to express TF proteins in vitro using the TnT T7 Quick for PCR DNA System (Promega L5540) with an incubation at 30ºC for 2 hours. Each 96 well plate of protein included four negative control wells containing a mock in vitro protein expression without any DNA expression template, to be used for background signal subtraction in the downstream peak-calling step.

Magne HaloTag Beads (Promega G7282) were first washed and resuspended in 50 µL of binding/wash buffer (1x phosphate buffered saline + 0.005% NP40 detergent). To bind TFs to magnetic beads, 20 µL of washed beads were incubated with 100 µL of each protein expression reaction at room temperature for 1 hour on a rotator. The HaloTag-TF-bound beads were gently washed in 150 µL of the same binding/wash buffer 3x, using a magnetic plate adapter to pellet the beads and discard the wash buffer between each wash. The beads were then combined with the gDNA libraries in a volume of 50 µL, brought to a total volume of 150 µL using binding/wash buffer, and incubated on a rotator at room temperature for 1 hour. Input gDNA library amounts were titrated based on genome size to target equal sequence coverage of each species: *A. thaliana* 100 ng, *A. lyrata* 172 ng, *B. oleracea* 407 ng, *C. rubella* 112 ng, *F. vesca* 183 ng, *O. sativa* 312 ng, *P. trichocarpa* 327 ng, *S. lycopersicum* 652 ng, *S. tuberosum* 618 ng, *S. bicolor* 565 ng. In addition, 10 µg of salmon sperm DNA (Invitrogen 15632011) was added alongside gDNA libraries to reduce background signal caused by non-specific binding. After incubation, the beads were again gently washed with the same wash buffer 3x, after which the wash buffer was removed and discarded. We found that the wash conditions are critical to achieving high sensitivity. At each wash step, the bead pellet should be fully resuspended while minimizing shear forces by gentle pipetting of the wash buffer stream directly at the bead pellet. After the final wash, the beads were resuspended and boiled in 22 µL of i7 index primers[95] at

95ºC for 15 min, and 20 µL of eluate from each well was transferred to a new plate and mixed with an equal volume of KAPA HiFi HotStart ReadyMix for 10 cycles of amplification. Equal volumes of each reaction were pooled across rows of the plate, and then purified by gel purification. Pools were sequenced either on an Illumina NovaSeq 6000 S4 flow cell or NovaSeq X plus 10B or 25B flow cell, using 2x150 bp paired-end sequencing, targeting 10 million reads (5 million fragments) per TF for *A. thaliana*, and proportionally scaled to larger genomes as dictated by the relative gDNA library input mass. For the primary four species Brassicaceae dataset, we identified libraries from that first round of sequencing with low sequencing yield and selectively pooled these for a second round of sequencing to increase coverage.

Single nuclei Experiments

Plant growth conditions

A. thaliana, *A. lyrata*, *C. rubella*, and *B. oleracea* seeds were surface sterilized before planting vertically on the agar plates containing 3mM $Ca(NO_3)_2$, 1.5mM $MgSO_4$, 1.25mM $NH_4H_2PO_4$, 1mM KCl, 1x micronutrients (Murashige and Skoog Micronutrient Salts 100x, MSP18-10LT), in 0.8% agar with pH 5.7. The seeds were sterilized by soaking in 75% ethanol for 1 minute and 50% bleach for 5 minutes, 25% bleach with 0.2% Triton for 4 minutes, then rinsed with sterile water 8 times. After sterilization, seeds in the 2 mL Eppendorf tube were wrapped in tin foil and put into a 4°C fridge for cold stratification. Seeds were left in the fridge for 3 days for *A. thaliana* and *C. rubella*, or 5 days for *A. lyrata*. Cold stratification was not needed for efficient germination of *B. oleracea* seeds. We staggered the planting of different species to account for different germination times of the four plant species, allowing us to harvest seedling root and shoot tissues of plants at the same 5-7 day old growth stage of all species on the same day. *A. lyrata* was planted on the agar plates 3 days before *A. thaliana*, *C. rubella* was planted 1 day after *A. thaliana*, and *B. oleracea* seeds without cold stratification were planted 2 days after *A. thaliana*. The conditions of the growth chamber (Percival Scientific) were set to 16 hours light at 22°C and 8 hours dark at 19°C, with the light turning on at 5am and turning off at 9pm. Root, shoot, and whole seedlings were sampled between 10am and 12pm for all experiments.

In order to enable sampling of mature leaf and flower bud samples on the same day of all four species, planting times were staggered and growth conditions were modulated to synchronize growth of all four species. For mature leaves, we sampled the youngest fully expanded rosette leaf from plants grown in soil in a plant grow room or cold room. *B. oleracea* mature leaves were harvested from plants that were two months old grown at 13°C under a light cycle of 12 hours light/12 hours dark. *A. lyrata* was three months old at harvest, grown at 4°C with a light cycle of 12 hours light/12 hours dark. *A. thaliana* and *C. rubella* were grown in conditions of 16 hours light at 28°C and 8 hours dark at 23°C and harvested at 2 months old. Mature leaves were sampled between 10am and 12pm.

For floral buds, we sampled young buds on the main stem from plants grown in soil, where the petal was just emerging (~1 mm visible). All four plant species were grown at 16 hours light at 22°C and 8 hours dark at 19°C when floral buds were collected. *A. thaliana* and *B. oleracea* flowered at about two months old without cold treatment. To induce flowering in *A. lyrata* and *C. rubella*, two month old plants were transferred into a 4°C cold room for two months

and then transferred back to conditions with 16 hours light at 22°C and 8 hours dark at 19°C for one month. In this way, all the four plant species were synchronized to flower on the same day and sampled between 10am-12pm.

Hormone treatment experiments

For hormone treatment experiments, *A. thaliana* seeds were sterilized and cold stratified as described above, then planted on nylon mesh in agar plates with the same growth medium composition as described above. Seeds were sown on plates at a density of 3 rows per plate with ~30 plants per row and plates were placed vertically in a growth chamber under conditions of 16 h light at 22°C, 8 h dark at 19°C for 5 days. On the 5th day, the nylon mesh and plants were transferred using forceps to agar plates containing either 2 μM ABA, 150 μM SA, or the standard growth medium without added hormones as a negative control. Samples were collected at time points of 0, 1, 2, 4, 7, 24, and 48 hours, separating root and shoot tissues with a single razor blade cut at the time of collection. At each time point, five biological replicates per treatment were collected and flash frozen in liquid nitrogen then stored at -80°C. Three of the biological replicates were used for bulk RNA-seq, and the remaining two were used for the 2 hour time point samples selected for follow up single nuclei sequencing.

Nuclei isolation and single nuclei library preparation

Plant tissue samples were either processed immediately after harvest or flash frozen in liquid nitrogen and stored at -80°C for up to one year before nuclei isolation. Freshly harvested (not frozen) samples were used for all snATAC-seq libraries. Nuclei isolation from tissues was performed as follows.

Buffer 1 (lysis buffer) consisted of 0.275 M sorbitol (Sigma-Aldrich S6021), 0.1% Triton X-100 (Sigma-Aldrich 93443), 0.01 M $MgCl_2$ (Ambion AM9530G), 1x protease inhibitor cocktail (Sigma-Aldrich 4693132001), 0.3 U/μl RNase inhibitor (Roche 03335399001), and 1 mM DTT (Teknova D9750).

Buffer 2 (wash and resuspension buffer) consisted of 1x phosphate buffer saline (without Mg and Ca), 1% BSA, 0.4 U/μl RNase inhibitor, and 1 mM DTT.

For isolation of nuclei from mature leaves containing high levels of chloroplasts, an extra modified wash and resuspension buffer (buffer 3) was used to eliminate chloroplasts before the final wash and resuspension step by buffer 2. It consisted of 0.275 M sorbitol, 0.2% Triton X-100, 0.2% NP40 (Thermo Fisher Scientific PI28324), 0.01 M $MgCl_2$, 1x protease inhibitor cocktail, and 1mM DTT.

All nuclei isolation steps were carried out in a cold room at 4°C. For each sample, 20-100 mg fresh or frozen tissue was chopped for 3 minutes using a razor blade on a glass plate with 100-200 μl buffer 1, then transferred to a petri dish with 1.5 mL cold buffer 1 and allowed to rest on ice with gentle shaking for 2 minutes. For multiplexed samples, tissues from different species were weighed separately before being combined and chopped together. The samples were then transferred to a 48-well filter plate (25 μm, 4mL, Agilent, 201003-100), which was pre-wet with 1 mL cold buffer 1 before it was placed on the receiving plate attached to a QIAvac96 (Qiagen). Pressure was maintained below 100 bar during filtration. The filtered nuclei solution was centrifuged at 500x g for 10 minutes at 4°C to pellet the nuclei, after which the supernatant was discarded.

For samples with low or no chloroplast content: The pellet was resuspended by gentle flicking and pipetting, washed with 2 mL cold buffer 2, centrifuged at 500x g for 5 minutes, and the supernatant was discarded.For nuclei isolation from mature leaves with high chloroplast content: The pellet was resuspended by gentle flicking and pipetting and 10 mL cold buffer 3 was added, centrifuged at 500x g for 5 minutes, and the supernatant was discarded. The pellet was washed with 2 mL cold buffer 2, centrifuged at 500x g for 10 minutes, and the supernatant was discarded. For both protocols, pellets from the final centrifugation were resuspended in 100 µl cold buffer 2 for snRNA-seq, or cold 1x nuclei buffer (10X Genomics PN-2000207) for snATAC-seq.

Nuclei quality and quantity were assessed by flow cytometry and microscopy before continuing to the single nuclei partitioning and barcoding step. A BD Accuri C6 plus flow cytometer was used to analyze the quality and quantity of nuclei stained by propidium iodide (Sigma-Aldrich P4864), added to a final concentration of 50 µg/mL. The stained samples were incubated on ice in darkness for 5-10 min prior to analysis. The analysis was conducted using light-scatter and fluorescence signals produced by a 20 mW laser illumination at 488 nm. Signals corresponding to forward-angle scatter (FSC), 90° side scatter (SSC), and fluorescence were collected. Fluorescence signals (pulse area measurements) were screened using the following filter configurations: (a) FL-2, with a 585/40 nm band-pass filter, and (b) FL-3, with a 670 nm long-pass filter. Threshold levels were set empirically at 80,000 for FSC-H to exclude debris commonly found in plant homogenates. Templates for uni- (FL2-A and number of nuclei) and bi-parametric (FL2-A and FL3-A) frequency distributions were established. Upon identifying the region corresponding to nuclei, data was collected to a total count of 5,000–10,000 nuclei. The flow cytometer was operated at the Slow Flow Rate setting (14 µl sample/minute), with data acquisition for a single sample typically taking 3–5 minutes. Different plant species exhibited varying patterns of somatic endoreduplication within most of their tissues and organs. For example, in young *A. thaliana* roots and cotyledons, this variation is reflected in the form of multiple clusters within the distribution, corresponding to nuclei forming a 2C, 4C, 8C, 16C, etc., of the endoreduplicative series[107]. All the nuclei of all the endoreduplicative series were counted and divided by the actual volume analyzed in the flow cytometer to obtain the concentration of the nuclei. An EVOS M5000 microscope was also used for validation of nuclei quality and quantity by analyzing 7 µl diluted nuclei sample stained by 1 µl of diluted SYBR Green dye (Thermo Fisher Scientific S7585 diluted 1:10,000 in water) using a C-Chip Disposable Hemocytometer. For nuclei prepared from fresh tissue samples, most of the nuclei in each sample showed well-resolved edges without obvious evidence of blebbing. In the case of nuclei prepared from frozen samples, where cell debris around the nuclei can interfere with these microscopy observations, nuclei quality was primarily assessed by comparing to fresh nuclei's distribution and the endoreduplicative series pattern on the flow cytometer.

The nuclei suspension was then diluted to concentration of 700-1000 nuclei/µl using cold buffer 2 for snRNAseq or cold 1x nuclei buffer (10x Genomics PN-2000207) for snATAC-seq. Single nuclei were partitioned and barcoded using a Chromium system (10x Genomics) with the Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index) for snRNAseq or Chromium Single Cell ATAC Reagent Kits v2 for snATAC-seq. Approximately 20,000 nuclei in total were loaded for single species samples, and up to 60,000 nuclei for multiplexed species samples. Library creation was carried out following the manufacturer's protocol. Sequencing was

performed on an Illumina NovaSeq 6000 S4 flow cell or NovaSeq X plus 10B or 25B flow cell, using 2x150 bp paired-end sequencing, targeting 400M reads per species for RNA-seq samples and 250M fragments per species for mixed-species ATAC-seq samples, with higher coverage for *A. thaliana* seedling samples (900M fragments and 1800M fragments for root and shoot respectively).

DAP-seq Analysis

Primary DAP-seq data analysis pipeline

Fastq files were quality filtered and adapters were trimmed using BBTools v38.96 bbduk.sh[108] with parameters k=21 mink=11 ktrim=r tbo tpe qtrim=r trimq=6 maq=10, and then aligned to the corresponding reference genome using bowtie2 v2.4.2[109] and parameters --no-mixed --no-discordant. Bam files from negative control wells were merged using the samtools v1.15.1[110] merge command, to generate a single background file for each 96-well plate. The MACS3 v3.0.0a6[111] callpeak command was used to generate narrowPeak files, using the merged background as the control, with the parameters --call-summits --keep-dup 1 and –gsize with the total fasta genome size. Motifs were called using MEME suite v5.3.0[112] command meme with parameters -dna -revcomp -mod anr -nmotifs 2 -minw 8 -maxw 32, and with a background model (-bfile) generated from the entire reference genome file with fasta-get-markov -m 0 -dna.

The resulting datasets were filtered to exclude those TFs and replicates with poor performance in the DAP-seq assay. As a measure of successful binding site enrichment, we calculated the fraction of reads in peaks (FRIP) scores and only used data from DAP-seq assays that generated FRIP scores of at least 0.05. For experiments including multiplexed species, we only considered those assays where all of the multiplexed species produced FRIP scores of 0.05 or greater. We found that multiplexing larger sets of species with varying genome sizes in a single multiDAP experiment sometimes resulted in large differences in read coverage between different TFs and in some cases differential enrichment of species. To mitigate the effect of low coverage resulting in false negatives in the peak-calling step (and thus lowering apparent c scores), we only used datasets that yielded at least 200,000 read pairs (400,000 reads) for each of the multiplexed species.

Assignment of DAP-seq peaks to target genes and conservation scores (c scores)

DAP-seq peaks files generated by the above analysis pipeline were further processed to identify target genes for each TF and in each species using a set of custom scripts. For each species, the corresponding gene annotation file (gff format) was filtered to retain only features corresponding to a single transcript per gene, for which the primary (or longest) transcript was selected. The script 1_assign_peaks_to_gff_features.sh (a wrapper for bedtools[113] v2.31.0) was used to assign peaks to mRNA, exon, and CDS features individually. The output was further processed using the script 2_annotate_peak_targets.py, which integrates information from the different assigned feature types to assign each peak to up to two gene targets and categorize their location with respect to the target gene as either upstream, downstream, within the 3’ or 5’ UTR, within the CDS, or intronic. Peaks falling in intergenic regions between two divergent

genes were assigned to both flanking gene targets, while peaks falling within the transcript (UTR 5', CDS, intron, or UTR 3') were assigned to only a single gene target.

The output of the above scripts was further processed with the script 3_filter_annotated_peaks_and_calculate_c_scores.ipynb, which was used to filter for peaks falling within the regions spanning -2000 bp to +500bp with respect to the target gene's start codon. Peaks were only considered if they had a strength of at least 5-fold over background (as defined in the 7th signalValue column of the narrowPeak file format), and if they fell in one of the following regions: upstream, 5' UTR, CDS, or intron. Each target gene was then mapped to an orthogroup, as defined by applying OrthoFinder v2.5.5[16] with default parameters to the protein sequences of all species. To improve orthogroup resolution within the more closely related species, separate tables of orthogroups were generated and used to analyze the four Brassicaceae species datasets and the ten species datasets. Finally, a conservation score (c score) was assigned to each TF-orthogroup assignment, representing a count of species in which the given TF also targets at least one gene within the same orthogroup.

See code availability statement for scripts, reference genome files, and sample data files.

DAP-seq signal and peak strength comparisons

For assessment of DAP-seq signal similarity across individual DAP-seq assays, we compared bigwig files using deeptools[114] v3.5.2 multiBigwigSummary bins command with parameter "--binSize 50". In order to reduce computation time, this analysis was restricted to only the first chromosome of the *A. thaliana* genome using the flag "--region Chr1". The resulting output was processed to generate a correlation matrix with the deeptools command and "--corMethod pearson".

For comparison of DAP-seq peak strengths between replicates, we first filtered the narrowPeak files for a single summit per peak region, retaining only the summit with the highest fold-change value. We then used bedtools[113] v2.31.0 intersect command with the parameter "-wo" to match corresponding peaks between replicates.

For comparison of DAP-seq peak strengths between species, we considered only the strongest peak for each target orthogroup in each species.

GO term enrichment

We used clusterprofiler v4.6.2[115] to perform gene ontology (GO) enrichment analysis on the list of target genes for each TF and c score. We first converted the TAIR gene names into the Entrez ids. Then we calculated enrichment values for Biological Processes and filtered for enriched terms with level six to avoid broad GO terms. We then used the "simplify" function to combine terms with a similarity cutoff of 70%. GO term enrichment calculations for specific other examples including ERF11 c4 genes in coding sequence and ABA-induced genes co-targeted by AREB3 and DREB1A were calculated using TAIR's implementation of enrichment analysis with PANTHER Classification System annotations and default parameters, available at https://v2.arabidopsis.org/tools/go_term_enrichment.jsp.

Nucleotide diversity in TFBSs

We used the 1001 genomes polymorphism data[34] to calculate nucleotide diversity (pi) with vcftools v0.1.17[116]. We only considered regions flanking peak summits (+/- 30 bp from the peak summit location) that were located within 2000 bp upstream of the assigned gene's coding sequence start. We did not include regions downstream of the start codon or regions overlapping coding sequences in the 2000 bp upstream of the assigned gene.

TFBS location distribution

We binned the genomic distance of the summit of a TFBS to the start codon in bins of 50 bp. We counted the number of TFBSs, for a given TF, across all target genes that fell inside each of these bins. To compare the distribution across c scores, we normalized the counts using the total number of TFBSs per c score per TF. We calculated the Gini coefficient for each TF and each c score to quantify unequal distributions of TFBSs.

Motif analysis

We used the first positional weight matrix motif identified by the MEME motif caller (MEME-1) from the 60 bp summit regions of the top 100 peaks and the MEME suite command fimo to search for instances of motif occurrences in the genome. We also used the RSAT[117] matrix-clustering algorithm to cluster TFs based on motif similarities with default parameters. We then intersected the motif instances with the DAP peaks to identify motifs inside DAP-seq peaks of the corresponding TF. To explore the motif direction conservation across species, we compared the direction of the *A. thaliana* motifs with those in the other Brassicaceae species. To avoid interference of potential divergent paralogs, we only considered motifs in peaks assigned to single-copy orthologs. We then calculated significance and enrichment based on the probability of finding the same motif direction in each other species for each c score using the binomial test from scipy[118]. We also calculated a motif directionality bias score defined as the absolute value of 1 - (motif count same strand)/(motif count opposite strand).

snRNA-seq and snATAC-seq analysis

snRNA-seq atlas construction

Raw fastq files for each library were filtered using BBTools v38.96 bbduk.sh[108] to remove read pairs with 31-mers in read 2 matching common ribosomal 31-mers. Remaining reads were processed with Cellranger count v7.0.1. Nuclear reference genomes were downloaded from Phytozome where available, and supplemented with organellar genomes from NCBI. QC metrics were compiled for each sample by (1) calculating the total UMIs and genes from each cell barcode (CBC), (2) calculating the proportion of UMIs from each CBC mapping to organellar genomes, (3) running Velocyto[119] to associate each CBC with an overall rate of unspliced vs. spliced transcripts, (4) running the R package diem[120] to estimate the rate of ambient RNA in each CBC. Cell barcodes were then filtered to those with at least 400 nuclear UMIs, 200 nuclear genes, 90% nuclear UMIs, 10% unspliced UMIs, and debris score <=2. (Note that the 400 UMI threshold was relaxed to 300 UMIs for flower bud samples, in order to obtain enough cells for meaningful clustering. All other filters remained unchanged.). In some experiments nuclei from

multiple species were pooled together, as inter-species sequence divergence is sufficient to distinguish transcripts based on genome alignment, and we did not observe any reads multi-mapping to annotated genes from different species.For these mixed-species samples, additional filtering was performed by defining background proportions for each species using all UMIs in the raw expression matrix, and then running a chi-squared test for each CBC to determine whether the species proportions calculated from the CBC's associated UMIs were significantly different ($p<0.01$) from background. CBCs passing this test were assumed to contain nuclei from one or more species. To determine which species were present, a per-species UMI threshold was defined as the 99th percentile of per-species UMIs across all CBCs failing the chi-squared test. Each CBC was then assigned an estimated ambient RNA proportion by comparing total UMIs from 'present' species to total UMIs from all species, and CBCs with >50% estimated ambient RNA were further filtered from downstream analysis.

After filtering each sample, count matrices for all remaining CBCs with organelle genes removed were loaded into Seurat (v5.0.0)[121], and *SCTransform* v2 normalization was performed, followed by an initial clustering using Seurat's *RunPCA* (with 5,000 variable genes and 20 PCs), *FindNeighbors*, and *FindClusters* (resolution 0.8). Next, marker genes were identified for each cluster using *FindAllMarkers* (with *logfc.threshold = .5, min.pct = .4, only.pos = T)* filtered to those with *p_val_adj<0.05*, and clusters with zero marker genes and significantly lower UMI counts than other clusters were removed as low quality clusters. Next, DoubletFinder[122] was run on remaining CBCs and those with doublet score >0.4 were removed. To prepare a final atlas, samples from the same species and tissue (seedling, leaf, or flower bud) were merged into a single seurat object, which was then split by sequencing batch and re-normalized with SCT per batch before performing a final PCA (50 PCs) and clustering (resolution = 1.0). Finally, we assigned clusters and cells to unique cell types and developmental stages via label transfer from three well-annotated large-scale atlases (a root atlas[52], leaf atlas[53], and a 6-day old seedling atlas[54]) as well as cluster-specific enrichments of known marker genes[55]. Label transfer was performed using labeled cells from each published reference atlas to each SCT-normalized batch of the *A. thaliana* seedling and leaf datasets with Seurat's *FindTransferAnchors* function using 20 PCs, and each CBC was putatively annotated with the highest scoring label. Additionally, marker genes from each cluster in each of the three *A. thaliana* tissue datasets were tested for enrichment (hypergeometric enrichment test) of a large suite of previously computed marker genes downloaded from scPlantDB[55]. Final cell type labels were compiled from these two sources, with additional manual annotation based on literature support for data-derived markers when there were discrepancies or no significant matches.

As a means of validating clustering and cell type labeling, we directly integrated our 35,169 *A. thaliana* seedling snRNA-seq transcriptomes with an external *A. thaliana* snRNA-seq dataset[56] derived from 7-day old seedling roots, resulting in a combined dataset of >45,000 nuclei. To combine in-house and external samples, and samples from different tissue subtypes (root, shoot, and whole seedling), we tested a number of different integration techniques (CCA, RPCA, and Harmony) using Seurat's *IntegrateLayers* function, and evaluated each as to the separation of root- and shoot- derived CBCs and mixing of internal and external root-derived CBCs. The RPCA method using a sample tree to fix the integration order proved to be the most appropriate, with root-specific cell type labels (for example, trichoblast-labeled cluster 29)

overwhelmingly composed of nuclei derived from root and whole seedling samples (cluster 29 log2 root:shoot ratio: 5.08, **Figure S5A**), and comprising an approximately even mix of nuclei from the internal and external root datasets (cluster 29 log2 internal:external ratio: -0.65). Conversely, root-derived nuclei (including all nuclei from the external root dataset) were almost entirely absent from clusters receiving leaf-specific labels (for example mesophyll-labeled cluster 13, with log2 root:shoot ratio: -5.24). Integration also revealed mixed internal-external clusters (including clusters 35, 17, and 19) that appeared to be associated with endodermis subtypes not typically captured by protoplast-based assays. These clusters shared expression of top endodermis-specific genes (for example AT3G32980/PRX32; **Figure S5C**) but had distinct expression profiles. Cells in cluster 35 were mostly labeled mature endodermis via label transfer from the protoplast-based reference root atlas (**Figure S5A**), yet unlike other endodermis clusters expressed key genes associated with lignin biosynthesis, including CASP1 and ERK1. Cells in cluster 17 received mixed low-confidence labels from the root atlas label transfer, and strongly expressed GELP96 and CYP86A1, genes with known roles in suberin biosynthesis and polymerization. The vast majority of external root nuclei in this cluster were labeled 'suberized-endodermis' in the original publication. The publication noted that this cell type was only detected in single nuclei rather than single cell datasets, potentially due to difficulty in digesting suberized cell walls. Cells in cluster 19 received low confidence atlas labels as well, yet included external cells from unlabeled 'cluster 14b' in the original publication. This cluster was also noted in the publication as appearing only in snRNA datasets, and showed tentative evidence of a role in cortex development and root epidermal patterning via expression of SCRAMBLED/STRUBBELIG. Both internal and external cells in our cluster 19 were differentiated from other endodermis cells by strong expression of FACT, a HXXXD-type acyl-transferase, and TLL1, a triacylglycerol lipase, both involved in synthesis of root wax components. Clustering of internal *A. thaliana* seedling cells without integration of the external dataset yielded a cluster with very similar expression. We note that a cluster was also identified in a separate published nuclei-based dataset[72] that showed highly similar expression patterns and was described as an 'endodermis cluster with epidermis precursor cells'. Additionally, we found a distinct subset of cells in each of our three other Brassicaceae seedling atlases with a best match to cells in cluster 19, indicating that this cell type is not unique to *A. thaliana.* Thus we referred to this cell type as 'epidermis/endodermis', though further characterization will be needed to fully understand its role.

Final cell type labels from the untreated *A. thaliana* seedling atlas were transferred to SCT-normalized datasets from hormone-treated seedling samples using Seurat's *FindTransferAnchors* function, selecting the top-scoring label per cell. Cell type labels from *A. thaliana* seedling, leaf, and flower atlases were transferred to the respective datasets from the other Brassicaceae species using SAMap[57], again selecting the top-scoring *A. thaliana* label per cell.

TF-target co-expression analysis

SCT-normalized count data from the *A. thaliana* seedling atlas was used to calculate Spearman correlation coefficients between the expression profiles of each assayed TF and every other gene across all nuclei. To compute final TF-target co-expression scores, we compared the squared correlation coefficient for a true TF-target pair to the squared correlation

coefficient for the TF and a matched control gene. To find an appropriate control gene, all genes in the genome were assigned to one of 10 equally sized bins based on their average expression across all nuclei (*expression-bin*), as well as a bin representing the number of species (out of 4) in which an ortholog was present (*ortho-bin*). We then compiled a list of 500 randomly chosen expressed targets per TF, and identified a matched non-target gene for each from the same *ortho-bin* and *expression-bin.* Next each TF's target genes were additionally binned by fold-change of the strongest DAP peak (5 equally sized *peak strength bins* per TF) and c score, and we calculated the average fold-change between the squared correlation coefficient of true targets vs. matched non-target controls within each group. We repeated this procedure 100 times to obtain mean and standard error values for each group.

Cell type specificity and enrichment analyses

In order to calculate target gene summary scores, atlases for untreated seedling, leaf, and flower bud tissues of each species were first merged together. Before proceeding, Seurat's *FindClusterMarkers* was re-run on the merged datasets (with parameters *logfc.threshold = .5, min.pct = .2, only.pos = T*), and filtered to the top 100 genes per cell type to define a final set of cross-tissue marker genes (**Data File S4**).

Target gene expression summary scores were computed for each TF's expressed c1 through c4 target gene sets in each cell of each atlas using Seurat's *AddModuleScore* function with *assay = "SCT"* and default parameters. Summary scores for seedling cells were used to calculate a measure of 'cell-type specificity' for each TF's target sets, by extracting the F statistic from an anova test using R function *aov* with formula *score ~ celltype*. F statistic distributions across all TFs were plotted as boxplots for each c score and each species, and paired one-sided t-tests (using R function *t.test*) were performed for each species comparing F statistics from c4 vs. c3 target sets.

To compute enrichment or depletion of the expression of a TF's c4 target genes in each cell type, separate one-sided Wilcoxon Rank Sum tests were run to test for significantly higher or significantly lower target expression summary scores for cells with a given cell type label compared to all other cells, and magnitude of the difference was calculated using the function *cohen.d* from the R package *effsize*. The same tests were used to compare expression of a TF's c4 target genes in untreated vs. ABA- or SA- treated cells within the same cell type.

snATAC-seq atlas construction

We used the Cell Ranger-ATAC version 2.1.0[123] count command to filter and map reads to each reference, and generate the fragments file containing a unique ATAC-seq fragment created by two separate transposition events. We then used SnapATAC2 v2.50[124] to process the fragment files to downstream analyses from cell filtering to cell type annotation. To keep high-quality cells we required a minimum of 3000 fragments per cell and a 3-fold enrichment around the TSS. We separated the data into 500 bp windows across the genome and created a cell by bin count matrix. We used the scrublet[125] algorithm to identify potential doublets and remove them afterwards. We then combined replicates (**Table S6**) using the MNN-Correct algorithm in SnapATAC2 for batch correction. For visualization purposes we created a UMAP of each atlas.

We used gene accessibility scores as a proxy for gene expression to annotate cells with cell types from our snRNA-seq atlases. We calculated gene accessibility scores using the 'pp.make_gene_matrix()' function in snapATAC2. Briefly, this function assigns genomic bins to nearby genes, counts the transposase insertion events associated with each gene, and creates a matrix of accessibility scores for a particular gene and cell. Using the gene accessibility matrix, we integrated the snATAC-seq and the snRNA-seq datasets selecting the top 1000 most variable genes from our snATAC-seq data. We used scVI[126] integrated in SnapATAC2 to transfer cell type labels from the snRNA-seq atlas to the snATAC-seq atlas. To do this, we trained models using as covariate the batch information, negative binomial distribution ("nb") as the gene likelihood statistic, and a maximum number of epochs of 1000. After assigning cell type labels, we called peaks using MACS3 integrated into snapATAC2 independently for each cell type. We then merged peaks that were in close proximity (250 bp to either edge) to make comparisons across different cell types possible. We also generated a matrix of peaks by cell types, filtered to peaks present in at least 20% of the cells for a given cell type.

To calculate the differentially accessible regions (DARs) we sampled random cells per cell type, excluding the cell type being tested, and combined them to create a background. We then used the regression-based test from snapATAC2 "snap.tl.diff_test()", to identify regions more likely to be open in a given cell type than the background set of cells. We defined DARs as peaks with an adjusted p value smaller than 0.01. For the TF enrichment analysis we intersected c4 sites with all DARs using our script "intersect_dap_with_atac.sh". For a given cell type, we tallied the TF by DAR intersections to create a contingency table. For a given TF and cell type, we counted the number of DARs overlapping and not overlapping a c4 TFBSs. To identify TFs potentially enriched in DARs for a given cell, we created a background by counting the number of DARs of all other cell types overlapping and not overlapping c4s sites. We calculated the significance of enrichment for each TFs and cell type combination using Fisher's exact test from scipy[118].

Bulk RNA-seq Analysis

Raw reads were processed using BBTools v38.96 bbduk.sh[108] for 1) removal of artifact sequence, RNA spike-in reads, PhiX reads, and reads containing any N, 2) quality trimming with method set at Q6, and 3) removal of reads with <50 bases. Filtered reads from each library were aligned to the *A. thaliana* TAIR10 reference genome using HISAT2 v2.2.1[127]. featureCounts[128] was used to generate raw gene counts using Araport 11 gene annotations. Only primary hits assigned to the reverse strand were included (-s 2 -p --primary options). High Pearson correlation of raw gene counts was observed between all pairs of replicates (average correlation 0.992; min. correlation 0.89). Principal component analysis (PCA) and differential expression (DE) analysis were performed using DESeq2 v1.42.0. PCA was performed separately for root and shoot samples on vst-normalized counts from all expressed genes. DE analysis was performed on all samples together, with *design = ~ tissue_treatment_timepoint*. P-values and fold-changes for comparisons between ABA- or SA- treated samples and untreated samples from the same tissue at the same time point were extracted and subject to fold-change shrinkage (*DESeq2::lfcShrink* with *type = "ashr"*). Significant DE genes were considered those with *padj<0.05* and *abs(l2fc.shrunk)>1*. Hypergeometric enrichment tests were performed (using R

function phyper) to identify c4 target sets that were significantly enriched (Benjamini-Hochberg FDR < 0.05) forDE-upregulated and/or DE-downregulated genes at each time point.

Supplemental Information

Supplemental Figures

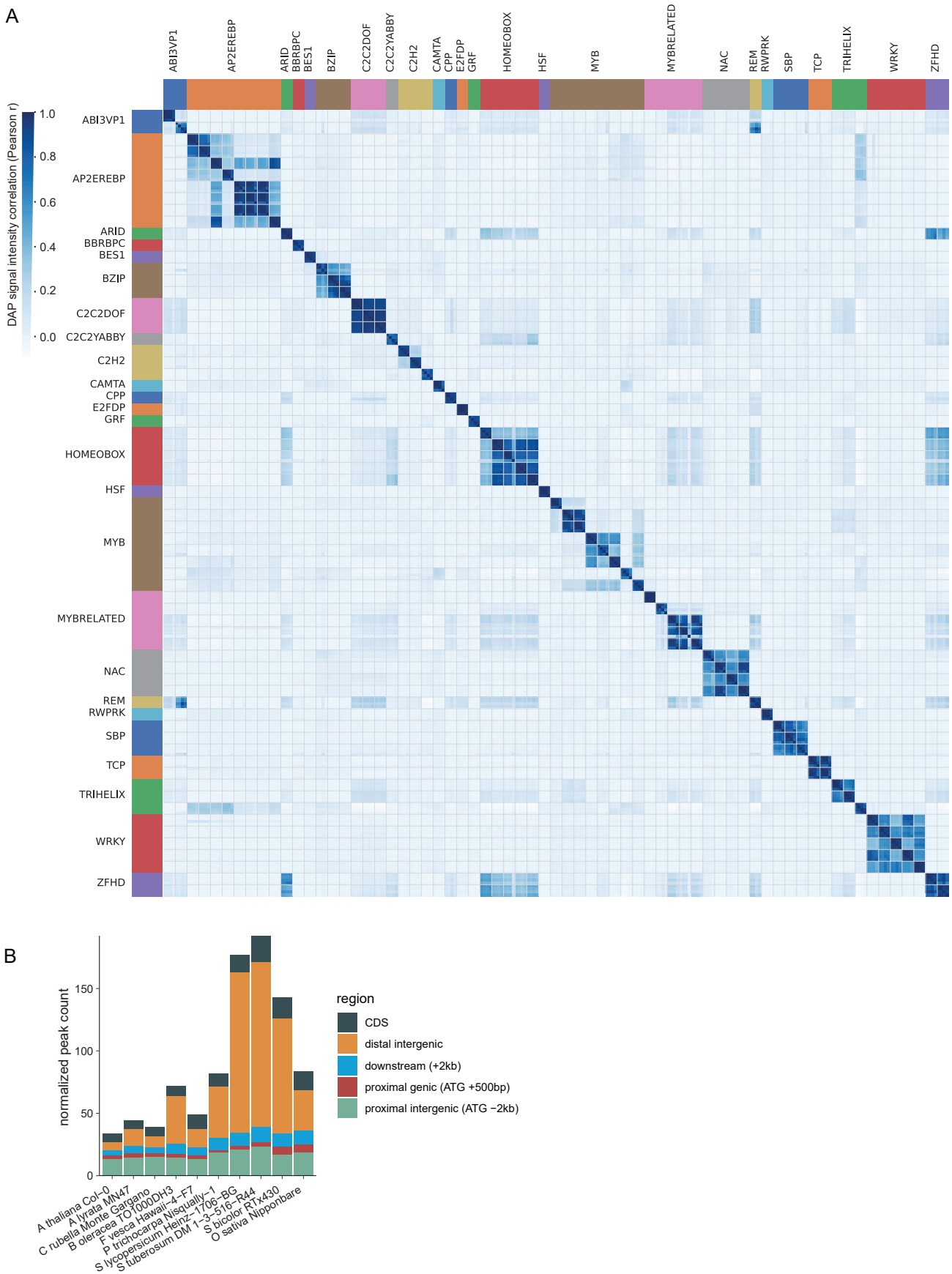

A
DAP signal intensity correlation (Pearson r)
1.0
0.8
0.6
0.4
0.2
0.0
ABI3VP1
AP2EREBP
ARID
BBRBPC
BES1
BZIP
C2C2DOF
C2C2YABBY
C2H2
CAMTA
CPP
E2FDP
GRF
HOMEOBOX
HSF
MYB
MYBRELATED
NAC
REM
RWPRK
SBP
TCP
TRIHELIX
WRKY
ZFHD
B
normalized peak count
0
50
100
150
A thaliana Col−0
A lyrata MN47
C rubella Monte Gargano
B oleracea TO1000DH3
F vesca Hawaii−4−F7
P trichocarpa Nisqually−1
S lycopersicum Heinz−1706−BG
S tuberosum DM 1−3−516−R44
S bicolor RTx430
O sativa Nipponbare
region
CDS
distal intergenic
downstream (+2kb)
proximal genic (ATG +500bp)
proximal intergenic (ATG −2kb)

Figure S1.
A) All-vs-all genome-wide correlations of DAP-seq binding profiles between four replicates for each TF using either 1, 4, or 10 multiplexed species. Gray grid lines delineate individual TFs and colored bars indicate TF families. Mean correlation between single species replicates=0.94, standard deviation=0.08, n=116; mean correlation between single and four species experiments=0.94, standard deviation=0.06, n=211; two-sample t-test p=0.56. Mean correlation of 10 species vs. single species experiments=0.87, standard deviation=0.1, n=163.
B) Count of peaks in genomic regions normalized by the number of total genes per species.
See also Figures 1 and 2.

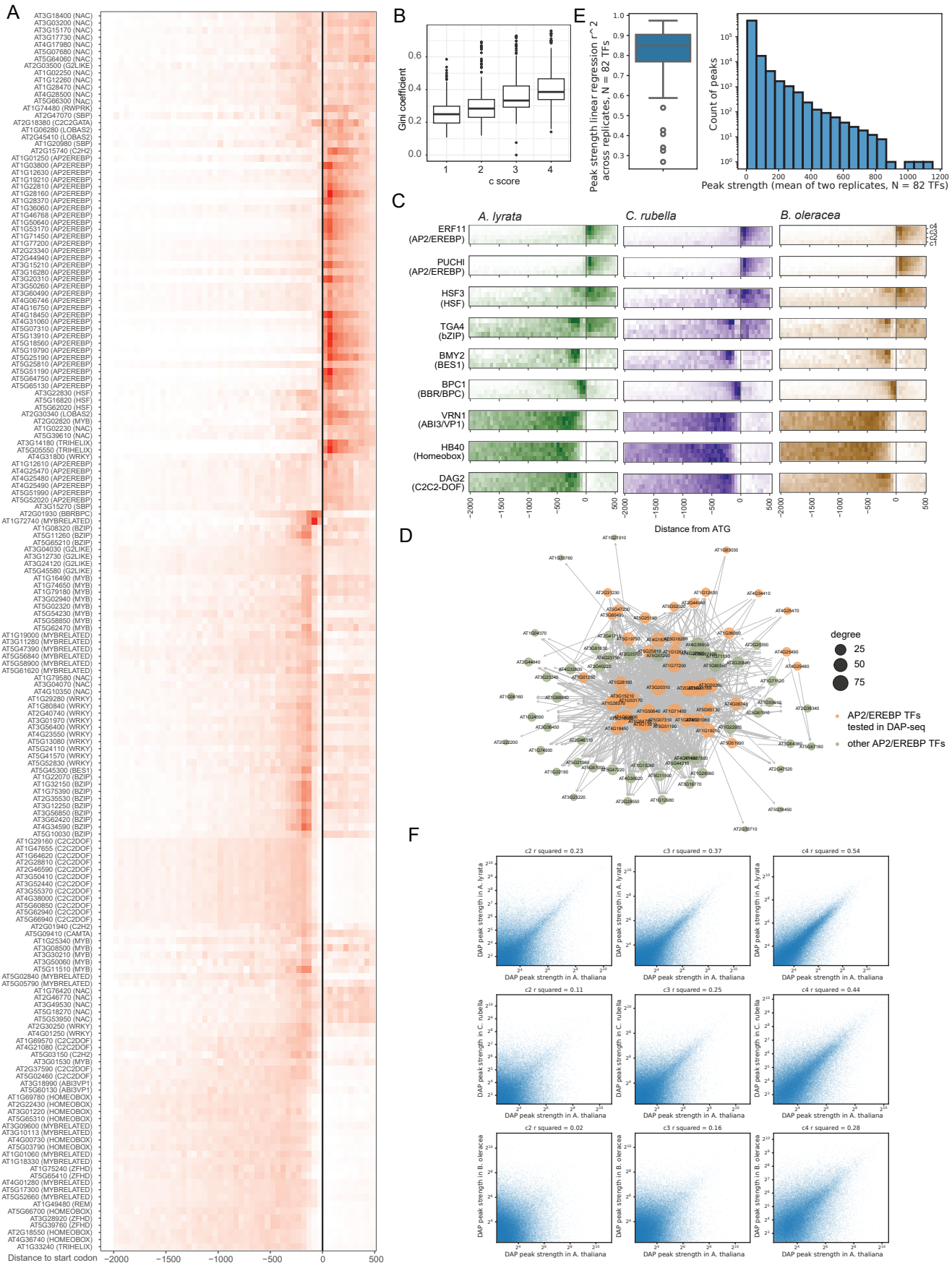

A
Distance to start codon
B
Gini coefficient
c score
E
Peak strength linear regression r^2 across replicates, N = 82 TFs
Count of peaks
Peak strength (mean of two replicates, N = 82 TFs)
C
A. lyrata
C. rubella
B. oleracea
ERF11 (AP2/EREBP)
PUCHI (AP2/EREBP)
HSF3 (HSF)
TGA4 (bZIP)
BMY2 (BES1)
BPC1 (BBR/BPC)
VRN1 (ABI3/VP1)
HB40 (Homeobox)
DAG2 (C2C2-DOF)
Distance from ATG
D
degree
AP2/EREBP TFs tested in DAP-seq
other AP2/EREBP TFs
F
c2 r squared = 0.23
c3 r squared = 0.37
c4 r squared = 0.54
c2 r squared = 0.11
c3 r squared = 0.25
c4 r squared = 0.44
c2 r squared = 0.02
c3 r squared = 0.16
c4 r squared = 0.28
DAP peak strength in A. lyrata
DAP peak strength in C. rubella
DAP peak strength in B. oleracea
DAP peak strength in A. thaliana

Figure S2.
(A) Distribution of c4 *A. thaliana* TFBS locations for TFs with more than 500 c4 peaks. (B) Distribution of the Gini coefficient across all TFs per c score showing peaks becoming more concentrated across fewer bins as the c score increases. (C) Distribution of TFBS locations in *A. lyrata*, *C. rubella*, and *B. oleracea* for TFs shown in Figure 2E and for each c score. (D) Gene regulatory network among AP2/EREBP family members based on *A. thaliana* DAP-seq results. (E) Correlation between peak strengths in replicate DAP-seq experiments. (F) Correlation between peak strengths across pairs of Brassicaceae species at different c scores. For each pair of species at each c score, peaks were randomly subsampled to 100k to allow direct visual comparisons between subplots.
See also Figure 2.

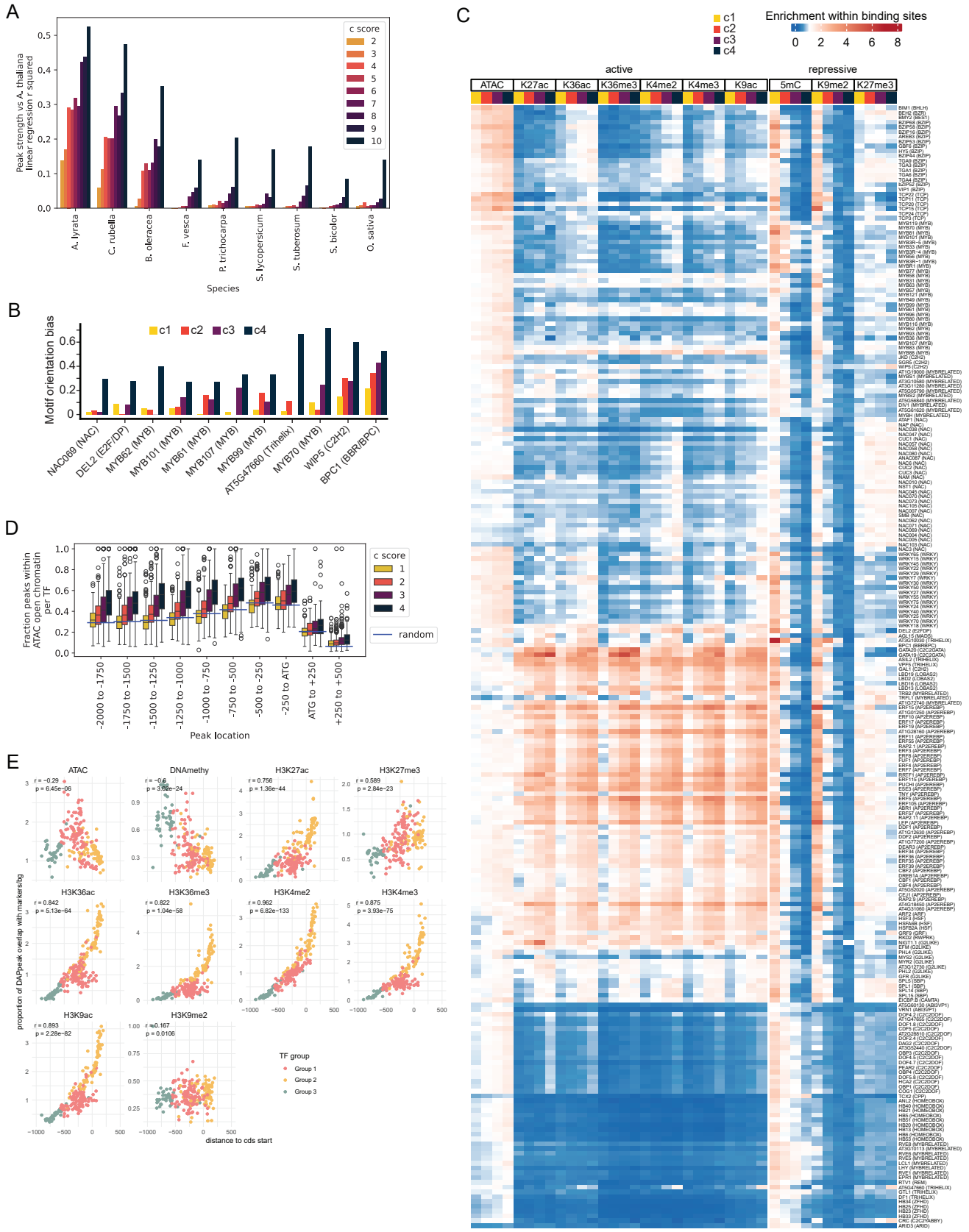

A
Peak strength vs A. thaliana linear regression r squared
c score
2
3
4
5
6
7
8
9
10
A. lyrata
C. rubella
B. oleracea
F. vesca
P. trichocarpa
S. lycopersicum
S. tuberosum
S. bicolor
O. sativa
Species
B
Motif orientation bias
c1
c2
c3
c4
NAC069 (NAC)
DEL2 (E2F/DP)
MYB62 (MYB)
MYB101 (MYB)
MYB61 (MYB)
MYB107 (MYB)
MYB99 (MYB)
AT5G47660 (Trihelix)
MYB70 (MYB)
WIP5 (C2H2)
BPC1 (BBR/BPC)
D
Fraction peaks within ATAC open chromatin per TF
c score
1
2
3
4
random
-2000 to -1750
-1750 to -1500
-1500 to -1250
-1250 to -1000
-1000 to -750
-750 to -500
-500 to -250
-250 to ATG
ATG to +250
+250 to +500
Peak location
E
ATAC
DNAmethy
H3K27ac
H3K27me3
H3K36ac
H3K36me3
H3K4me2
H3K4me3
H3K9ac
H3K9me2
proportion of DAPpeak overlap with markers/bg
distance to cds start
TF group
Group 1
Group 2
Group 3
C
c1
c2
c3
c4
Enrichment within binding sites
0
2
4
6
8
active
repressive
ATAC
K27ac
K36ac
K36me3
K4me2
K4me3
K9ac
5mC
K9me2
K27me3

Figure S3.
A) DAP-seq peak strength correlation between *A. thaliana* and the other nine species for each c score. (B) Motif orientation bias in *A. thaliana* for each c score. TFs with a bias score higher than 0.25 for c4 peaks are shown. (C) Enrichment of epigenetic markers overlapping TFBSs from all TFs tested binned by c score. (D) Fraction of DAP-seq peaks in open chromatin for each TF and each c score, split by peak location into 250 bp bins. (E) Correlation between median distance to TSS and enrichment of epigenetic marks of all TFs. Colors represent groupings of TFs based on the epigenetic mark overlap pattern seen in Figure S3C. Group 1 TFs are ATAC-seq enriched TFs, top cluster; Group 2 TFs are active mark enriched TFs, middle cluster; Group 3 TFs are histone mark depleted TFs, bottom cluster.
See also Figure 2.

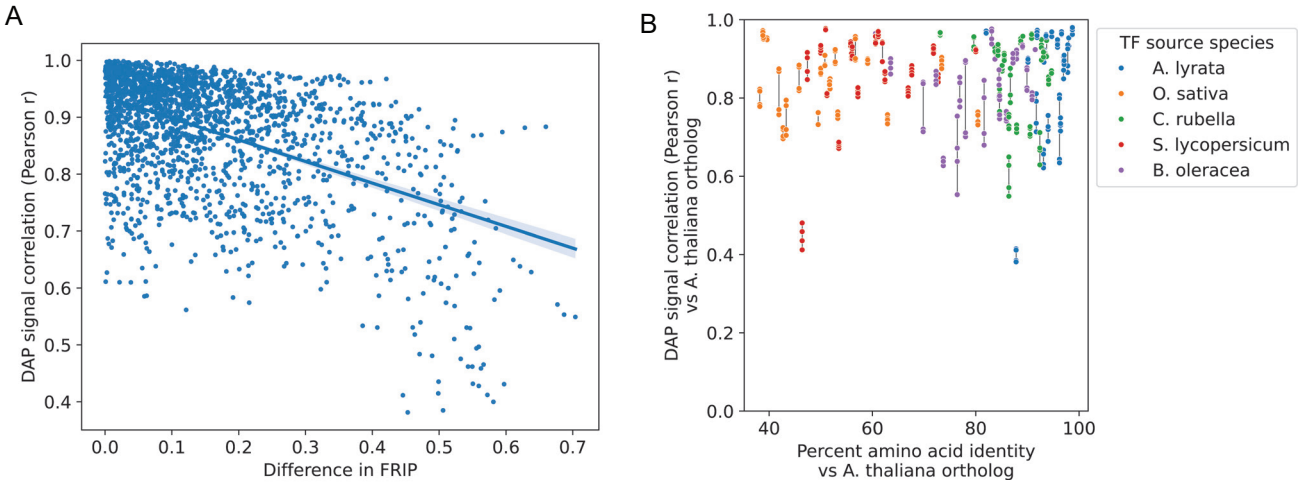

A
DAP signal correlation (Pearson r)
Difference in FRIP
B
DAP signal correlation (Pearson r) vs A. thaliana ortholog
Percent amino acid identity vs A. thaliana ortholog
TF source species
A. lyrata
O. sativa
C. rubella
S. lycopersicum
B. oleracea

Figure S4.
(A) DAP-seq binding signal correlation between orthologs as a function of the difference in fraction of reads in peaks (FRIP) scores. (B) DAP-seq signal correlation between *A. thaliana* TFs and the corresponding orthologs from another species vs. the percent amino acid identity between each pair of orthologs.
See also Figure 3.

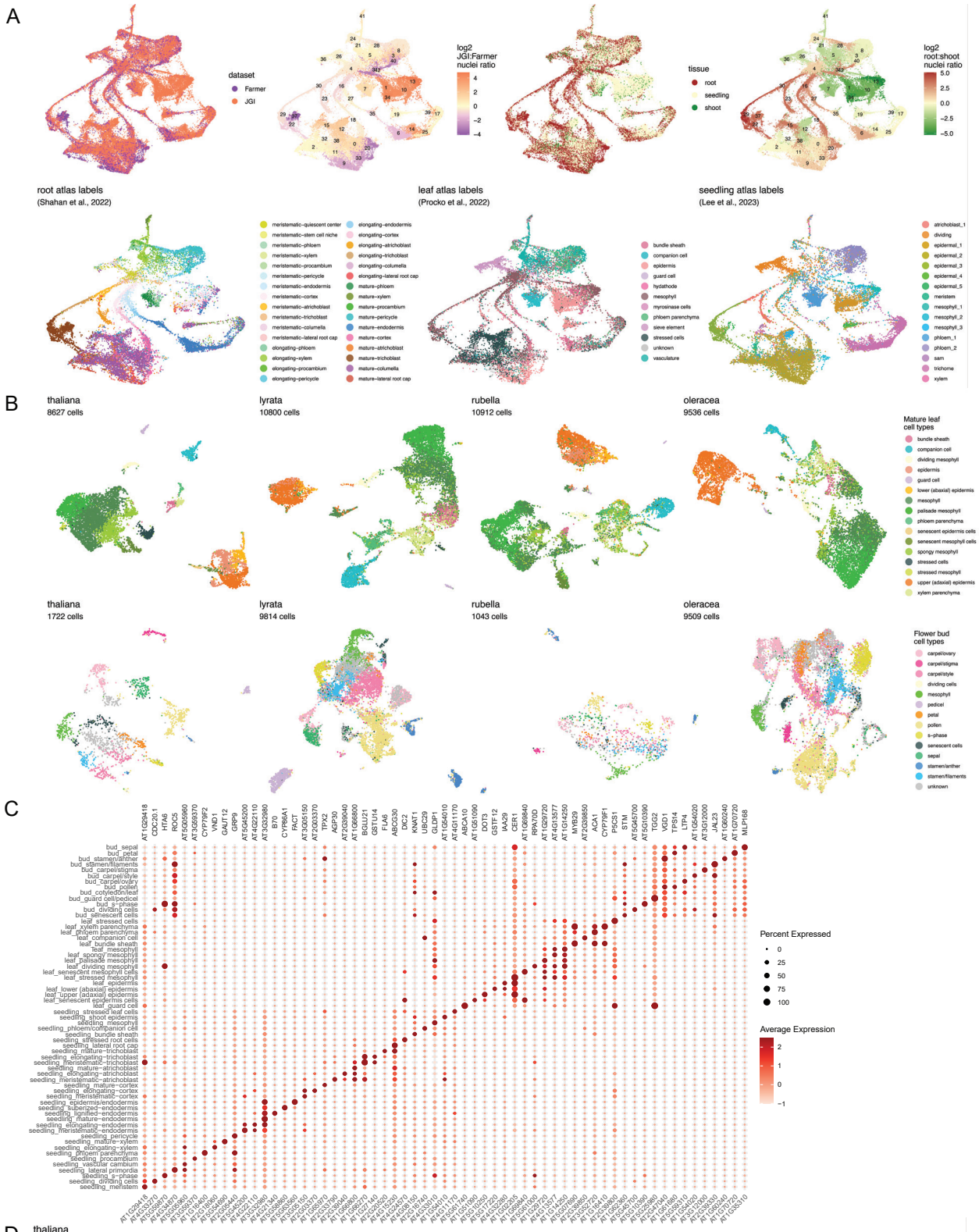
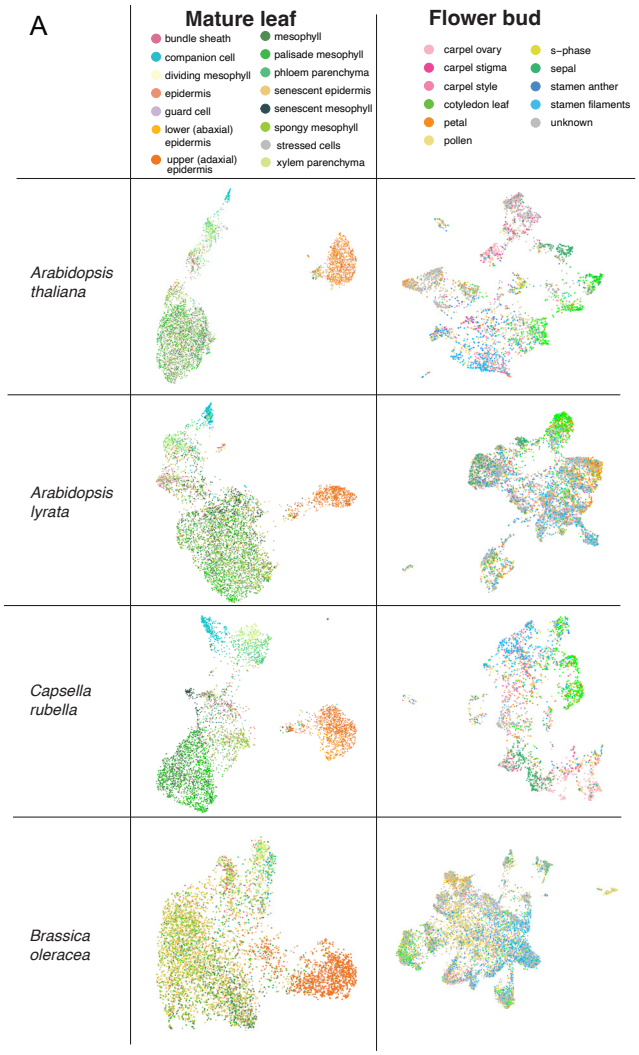
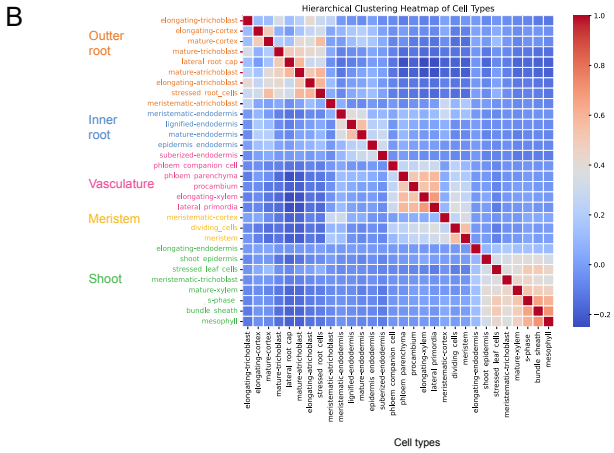

A
dataset
Farmer
JGI
log2 JGI:Farmer nuclei ratio
tissue
root
seedling
shoot
log2 root:shoot nuclei ratio
root atlas labels
(Shahan et al., 2022)
leaf atlas labels
(Procko et al., 2022)
seedling atlas labels
(Lee et al., 2023)
B
thaliana
8627 cells
lyrata
10800 cells
rubella
10912 cells
oleracea
9536 cells
Mature leaf cell types
thaliana
1722 cells
lyrata
9814 cells
rubella
1043 cells
oleracea
9509 cells
Flower bud cell types
C
Percent Expressed
Average Expression
D
thaliana
NAC−2 | NAC007 (AT1G12260)
c1 267 targets
c2 427 targets
c3 493 targets
c4 552 targets
c5 378 targets
c6 259 targets
c7 213 targets
c8 138 targets
c9 47 targets
c10 29 targets
target set summary score
min_q:0 max_q:0.9999

Figure S5.
(A) Integration of internal and external *A. thaliana* seedling snRNA-seq datasets (first row) and results of label transfer from three reference atlases (second row). Only cells receiving a label with confidence score >0.4 are shown in these plots. (B) UMAP representations of mature leaf and flower bud snRNA-seq atlases from four profiled Brassicaceae species. (C) Dotplot of the top marker gene for each cell type, where dot size represents the percent of all nuclei with a specific cell type label showing non-zero expression of the gene, and color represents average normalized and scaled expression across these nuclei. (D) NAC007/VND4 target expression summary scores using target c scores from extended 10-species multiDAP dataset. Note that two targets are excluded as they were not expressed in any cells in the snRNA-seq dataset. See also Figure 4.


A
Mature leaf
bundle sheath
companion cell
dividing mesophyll
epidermis
guard cell
lower (abaxial) epidermis
upper (adaxial) epidermis
mesophyll
palisade mesophyll
phloem parenchyma
senescent epidermis
senescent mesophyll
spongy mesophyll
stressed cells
xylem parenchyma
Flower bud
carpel ovary
carpel stigma
carpel style
cotyledon leaf
petal
pollen
s-phase
sepal
stamen anther
stamen filaments
unknown
Arabidopsis thaliana
Arabidopsis lyrata
Capsella rubella
Brassica oleracea

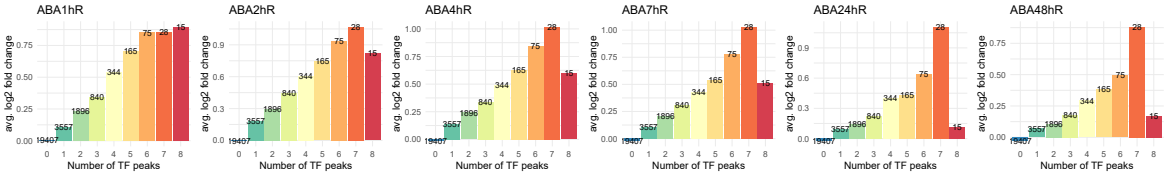

B
Hierarchical Clustering Heatmap of Cell Types
Outter root
Inner root
Vasculature
Meristem
Shoot
Cell types

C
−log10(p)
Count
Value

Figure S6.
(A) UMAP representation of snATAC-seq data for mature leaf and flower bud in *A. thaliana*, *A. lyrata*, *C. rubella* and *B. oleracea*. (B) Pearson correlation of DARs between each cell type in the *A. thaliana* seedling atlas. (C) Heatmap of TFBSs enriched in *A. thaliana* DARs for every TF and cell type with significant enrichment (Fisher's exact test $p<0.01$).
See also Figure 6.

Root samples

TFs: AREB3,BMY2,NAP,DDF1,MYB49,bZIP44,HSFA6B,LHY,AT3G10580,BEH2,SPL14,NAC070



Shoot samples

TFs: MYB93,AREB3,BMY2,NAC047,bZIP44,DDF1,LHY,SPL5,AT1G36060

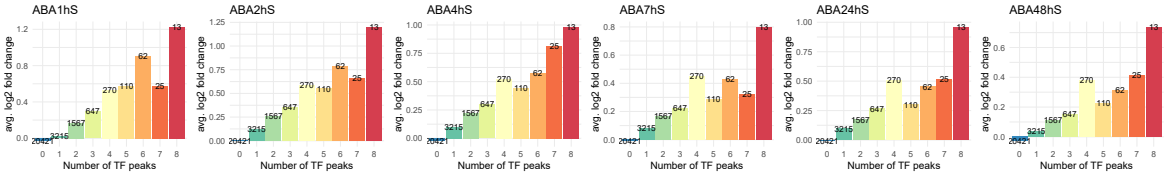


Figure S7.
Average ABA-induced log2 fold-change of the target genes of ABA-responsive TFs, as a function of the total number of ABA-responsive TFBSs in each gene's promoter, for A) root and B) shoot samples. ABA-responsive TFs were chosen separately for root and shoot samples by filtering to TFs whose c4 target sets were significantly upregulated (FDR < 0.01, log2 fold-change > 1) at two or more timepoints, and then selecting one TF per subfamily that was significant at the most timepoints with the highest average log2 fold-change. This resulted in 12 TFs for root samples and 9 TFs for shoot samples; TF names are listed in the figure for each plot. The number of genes in each category is indicated at the top of each bar.
See also Figure 7.

Supplemental Tables

Table S1. DAP-seq and multiDAP libraries. See also Figures 1, 2, and 3.

Table S2. Plant accessions and germplasm sources. See also Figure 1.

Table S3. Significantly enriched biological process GO terms for target genes of c2+ TFBSs. See also Figure 2.

Table S4. Significantly enriched biological process GO terms for target genes of c10 TFBSs. See also Figure 2.

Table S5. Percentage of assigned peaks in coding sequence for each TF. See also Figure 2.

Table S6. snRNA-seq and snATAC-seq samples and quality control metrics. See also Figures 4, 5, 6, and 7.

Table S7. All TFs with significant cell type-specific target gene expression enrichments. See also Figure 5.

Supplemental Data Files

Data File S1. Clustering of all TFs based on their motifs.

Data File S2. Filtered and annotated peaks assigned to genes with c scores from the 4 species Brassicaceae dataset.

Data File S3. Filtered and annotated peaks assigned to genes with c scores from the 10 species dataset.

Data File S4. Top 100 snRNA-seq marker genes for each cell type in each Brassicaceae species.

Data File S5. Significant DARs from the *A. thaliana* snATAC-seq seedling atlas.

References

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

2. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

3. Ruan, S., Swamidass, S.J., and Stormo, G.D. (2017). BEESEM: estimation of binding energy models using HT-SELEX data. Bioinformatics *33*, 2288–2295.

4. Li, M., Yao, T., Lin, W., Hinckley, W.E., Galli, M., Muchero, W., Gallavotti, A., Chen, J.-G.,

and Huang, S.-S.C. (2023). Double DAP-seq uncovered synergistic DNA binding of interacting bZIP transcription factors. Nat. Commun. *14*, 2600.

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

6. Lai, X., Blanc-Mathieu, R., GrandVuillemin, L., Huang, Y., Stigliani, A., Lucas, J., Thévenon, E., Loue-Manifel, J., Turchi, L., Daher, H., et al. (2021). The LEAFY floral regulator displays pioneer transcription factor properties. Mol. Plant. https://doi.org/10.1016/j.molp.2021.03.004.

7. Lai, X., Stigliani, A., Lucas, J., Hugouvieux, V., Parcy, F., and Zubieta, C. (2020). Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. Nucleic Acids Res. *48*, 9637–9648.

8. Huberman, L.B., Wu, V.W., Kowbel, D.J., Lee, J., Daum, C., Grigoriev, I.V., O'Malley, R.C., and Glass, N.L. (2021). DNA affinity purification sequencing and transcriptional profiling reveal new aspects of nitrogen regulation in a filamentous fungus. Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2009501118.

9. Tu, X., Ren, S., Shen, W., Li, J., Li, Y., Li, C., Li, Y., Zong, Z., Xie, W., Grierson, D., et al. (2022). Limited conservation in cross-species comparison of GLK transcription factor binding suggested wide-spread cistrome divergence. Nat. Commun. *13*, 7632.

10. Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036–1040.

11. Boj, S.F., Servitja, J.M., Martin, D., Rios, M., Talianidis, I., Guigo, R., and Ferrer, J. (2009). Functional targets of the monogenic diabetes transcription factors HNF-1alpha and HNF-4alpha are highly conserved between mice and humans. Diabetes *58*, 1245–1253.

12. Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A.J., Funnell, A.P.W., Goncalves, A., et al. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. Elife *3*, e02626.

13. Galli, M., Chen, Z., Ghandour, T., Chaudhry, A., Gregory, J., Li, M., Zhang, X., Dong, Y., Song, G., Walley, J.W., et al. (2024). Transcription factor binding site divergence across maize inbred lines drives transcriptional and phenotypic variation. bioRxivorg. https://doi.org/10.1101/2024.05.31.596834.

14. Sun, Y., Oh, D.-H., Duan, L., Ramachandran, P., Ramirez, A., Bartlett, A., Tran, K.-N., Wang, G., Dassanayake, M., and Dinneny, J.R. (2022). Divergence in the ABA gene regulatory network underlies differential growth control. Nat. Plants *8*, 549–560.

15. Chaw, S.-M., Chang, C.-C., Chen, H.-L., and Li, W.-H. (2004). Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. *58*, 424–441.

16. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for

comparative genomics. Genome Biol. *20*, 238.

17. 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.

18. Gou, M., Hou, G., Yang, H., Zhang, X., Cai, Y., Kai, G., and Liu, C.-J. (2017). The MYB107 Transcription Factor Positively Regulates Suberin Biosynthesis. Plant Physiol. *173*, 1045–1058.

19. Shukla, V., Han, J.-P., Cléard, F., Lefebvre-Legendre, L., Gully, K., Flis, P., Berhin, A., Andersen, T.G., Salt, D.E., Nawrath, C., et al. (2021). Suberin plasticity to developmental and exogenous cues is regulated by a set of MYB transcription factors. Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2101730118.

20. Woolfson, K.N., Esfandiari, M., and Bernards, M.A. (2022). Suberin biosynthesis, assembly, and regulation. Plants *11*, 555.

21. Serra, O., and Geldner, N. (2022). The making of suberin. New Phytol. *235*, 848–866.

22. Fraser, C.M., and Chapple, C. (2011). The phenylpropanoid pathway in Arabidopsis. Arabidopsis Book *9*, e0152.

23. Hrmova, M., and Hussain, S.S. (2021). Plant Transcription Factors Involved in Drought and Associated Stresses. Int. J. Mol. Sci. *22*. https://doi.org/10.3390/ijms22115662.

24. Jiang, C.-K., and Rao, G.-Y. (2020). Insights into the Diversification and Evolution of R2R3-MYB Transcription Factors in Plants. Plant Physiol. *183*, 637–655.

25. Kim, Y., Park, S., Gilmour, S.J., and Thomashow, M.F. (2013). Roles of CAMTA transcription factors and salicylic acid in configuring the low-temperature transcriptome and freezing tolerance of Arabidopsis. Plant J. *75*, 364–376.

26. Shibata, M., Breuer, C., Kawamura, A., Clark, N.M., Rymen, B., Braidwood, L., Morohashi, K., Busch, W., Benfey, P.N., Sozzani, R., et al. (2018). GTL1 and DF1 regulate root hair growth through transcriptional repression of *ROOT HAIR DEFECTIVE 6-LIKE 4* in *Arabidopsis*. Development *145*. https://doi.org/10.1242/dev.159707.

27. Bowman, J.L., and Smyth, D.R. (1999). CRABS CLAW, a gene that regulates carpel and nectary development in Arabidopsis, encodes a novel protein with zinc finger and helix-loop-helix domains. Development *126*, 2387–2396.

28. Chen, P., Takatsuka, H., Takahashi, N., Kurata, R., Fukao, Y., Kobayashi, K., Ito, M., and Umeda, M. (2017). Arabidopsis R1R2R3-Myb proteins are essential for inhibiting cell division in response to DNA damage. Nat. Commun. *8*, 635.

29. Ramirez-Parra, E., López-Matas, M.A., Fründt, C., and Gutierrez, C. (2004). Role of an atypical E2F transcription factor in the control of Arabidopsis cell growth and differentiation. Plant Cell *16*, 2350–2363.

30. Huang, Y.-C., Niu, C.-Y., Yang, C.-R., and Jinn, T.-L. (2016). The Heat Stress Factor HSFA6b Connects ABA Signaling and ABA-Mediated Heat Responses. Plant Physiol. *172*,

1182–1199.

31. Johnson, C., Boden, E., and Arias, J. (2003). Salicylic acid and NPR1 induce the recruitment of trans-activating TGA factors to a defense gene promoter in Arabidopsis. Plant Cell *15*, 1846–1858.

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

33. Ramachandran, P., Augstein, F., Mazumdar, S., Van Nguyen, T., Minina, E.A., Melnyk, C.W., and Carlsbecker, A. (2021). Abscisic acid signaling activates distinct VND transcription factors to promote xylem differentiation in Arabidopsis. Curr. Biol. *31*, 3153–3161.e5.

34. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, and 1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481–491.

35. Voichek, Y., Hristova, G., Mollá-Morales, A., Weigel, D., and Nordborg, M. (2024). Widespread position-dependent transcriptional regulatory sequences in plants. bioRxiv, 2023.09.15.557872. https://doi.org/10.1101/2023.09.15.557872.

36. Ma, Z., Hu, L., and Jiang, W. (2024). Understanding AP2/ERF Transcription Factor Responses and Tolerance to Various Abiotic Stresses in Plants: A Comprehensive Review. Int. J. Mol. Sci. *25*. https://doi.org/10.3390/ijms25020893.

37. Bemer, M., van Dijk, A.D.J., Immink, R.G.H., and Angenent, G.C. (2017). Cross-Family Transcription Factor Interactions: An Additional Layer of Gene Regulation. Trends Plant Sci. *22*, 66–80.

38. de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell *60*, 676–684.

39. Theune, M.L., Bloss, U., Brand, L.H., Ladwig, F., and Wanke, D. (2019). Phylogenetic Analyses and GAGA-Motif Binding Studies of BBR/BPC Proteins Lend to Clues in GAGA-Motif Recognition and a Regulatory Role in Brassinosteroid Signaling. Front. Plant Sci. *10*, 466.

40. Marand, A.P., Chen, Z., Gallavotti, A., and Schmitz, R.J. (2021). A cis-regulatory atlas in maize at single-cell resolution. Cell *184*, 3041–3055.e21.

41. Cheng, K., Xu, Y., Yang, C., Ouellette, L., Niu, L., Zhou, X., Chu, L., Zhuang, F., Liu, J., Wu, H., et al. (2020). Histone tales: lysine methylation, a protagonist in Arabidopsis development. J. Exp. Bot. *71*, 793–807.

42. Kawakatsu, T., Huang, S.-S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., Castanon, R., Nery, J.R., Barragan, C., He, Y., et al. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell *166*, 492–505.

43. Lu, Z., Hofmeister, B.T., Vollmers, C., DuBois, R.M., and Schmitz, R.J. (2017). Combining

ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Res. *45*, e41.

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

45. Mahrez, W., Arellano, M.S.T., Moreno-Romero, J., Nakamura, M., Shu, H., Nanni, P., Köhler, C., Gruissem, W., and Hennig, L. (2016). H3K36ac Is an Evolutionary Conserved Plant Histone Modification That Marks Active Genes. Plant Physiol. *170*, 1566–1577.

46. Li, Y., Mukherjee, I., Thum, K.E., Tanurdzic, M., Katari, M.S., Obertello, M., Edwards, M.B., McCombie, W.R., Martienssen, R.A., and Coruzzi, G.M. (2015). The histone methyltransferase SDG8 mediates the epigenetic modification of light and carbon responsive genes in plants. Genome Biol. *16*, 79.

47. Inagaki, S., Takahashi, M., Hosaka, A., Ito, T., Toyoda, A., Fujiyama, A., Tarutani, Y., and Kakutani, T. (2017). Gene-body chromatin modification dynamics mediate epigenome differentiation in Arabidopsis. EMBO J. *36*, 970–980.

48. Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P., and Lam, E. (2013). Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. Plant J. *73*, 77–90.

49. Cui, X., Lu, F., Qiu, Q., Zhou, B., Gu, L., Zhang, S., Kang, Y., Cui, X., Ma, X., Yao, Q., et al. (2016). REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. Nat. Genet. *48*, 694–699.

50. Jackson, J.P., Johnson, L., Jasencakova, Z., Zhang, X., PerezBurgos, L., Singh, P.B., Cheng, X., Schubert, I., Jenuwein, T., and Jacobsen, S.E. (2004). Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in Arabidopsis thaliana. Chromosoma *112*, 308–315.

51. Mathieu, O., Probst, A.V., and Paszkowski, J. (2005). Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. EMBO J. *24*, 2783–2791.

52. Shahan, R., Hsu, C.-W., Nolan, T.M., Cole, B.J., Taylor, I.W., Greenstreet, L., Zhang, S., Afanassiev, A., Vlot, A.H.C., Schiebinger, G., et al. (2022). A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. Dev. Cell *57*, 543–560.e9.

53. Procko, C., Lee, T., Borsuk, A., Bargmann, B.O.R., Dabi, T., Nery, J.R., Estelle, M., Baird, L., O'Connor, C., Brodersen, C., et al. (2022). Leaf cell-specific and single-cell transcriptional profiling reveals a role for the palisade layer in UV light protection. Plant Cell *34*, 3261–3279.

54. Lee, T.A., Nobori, T., Illouz-Eliaz, N., Xu, J., Jow, B., Nery, J.R., and Ecker, J.R. (2023). A single-nucleus atlas of seed-to-seed development in Arabidopsis. bioRxiv, 2023.03.23.533992. https://doi.org/10.1101/2023.03.23.533992.

55. He, Z., Luo, Y., Zhou, X., Zhu, T., Lan, Y., and Chen, D. (2024). scPlantDB: a

comprehensive database for exploring cell types and markers of plant cell atlases. Nucleic Acids Res. *52*, D1629–D1638.

56. Farmer, A., Thibivilliers, S., Ryu, K.H., Schiefelbein, J., and Libault, M. (2021). Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. Mol. Plant *14*, 372–383.

57. Tarashansky, A.J., Musser, J.M., Khariton, M., Li, P., Arendt, D., Quake, S.R., and Wang, B. (2021). Mapping single-cell atlases throughout Metazoa unravels cell type evolution. Elife *10*. https://doi.org/10.7554/eLife.66747.

58. Hunter, T. (2007). The age of crosstalk: phosphorylation, ubiquitination, and beyond. Mol. Cell *28*, 730–738.

59. Santner, A., and Estelle, M. (2009). Recent advances and emerging trends in plant hormone signalling. Nature *459*, 1071–1078.

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

61. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

62. Yamaguchi, M., Mitsuda, N., Ohtani, M., Ohme-Takagi, M., Kato, K., and Demura, T. (2011). VASCULAR-RELATED NAC-DOMAIN7 directly regulates the expression of a broad range of genes for xylem vessel formation. Plant J. *66*, 579–590.

63. Gushino, S., Tsai, A.Y.-L., Otani, M., Demura, T., and Sawa, S. (2024). VND genes redundantly regulate cell wall thickening during parasitic nematode infection. Plant Cell Physiol., cae048.

64. Yang, S.U., Kim, H., Kim, R.J., Kim, J., and Suh, M.C. (2020). AP2/DREB Transcription Factor RAP2.4 Activates Cuticular Wax Biosynthesis in Arabidopsis Leaves Under Drought. Front. Plant Sci. *11*, 895.

65. Quon, T., Lampugnani, E.R., and Smyth, D.R. (2017). PETAL LOSS and ROXY1 Interact to Limit Growth Within and between Sepals But to Promote Petal Initiation in Arabidopsis thaliana. Front. Plant Sci. *8*, 152.

66. Zhao, C., Hanada, A., Yamaguchi, S., Kamiya, Y., and Beers, E.P. (2011). The Arabidopsis Myb genes MYR1 and MYR2 are redundant negative regulators of flowering time under decreased light intensity. Plant J. *66*, 502–515.

67. Wang, S., Zhou, Z., Rahiman, R., Lee, G.S.Y., Yeo, Y.K., Yang, X., and Lau, O.S. (2021). Light regulates stomatal development by modulating paracrine signaling from inner tissues. Nat. Commun. *12*. https://doi.org/10.1038/s41467-021-23728-2.

68. Wang, W., Sijacic, P., Xu, P., Lian, H., and Liu, Z. (2018). Arabidopsis TSO1 and MYB3R1 form a regulatory module to coordinate cell proliferation with differentiation in shoot and root. Proc. Natl. Acad. Sci. U. S. A. *115*, E3045–E3054.

69. Clark, N.M., Buckner, E., Fisher, A.P., Nelson, E.C., Nguyen, T.T., Simmons, A.R., de Luis

Balaguer, M.A., Butler-Smith, T., Sheldon, P.J., Bergmann, D.C., et al. (2019). Stem-cell-ubiquitous genes spatiotemporally coordinate division through regulation of stem-cell-specific gene networks. Nat. Commun. *10*, 5574.

70. Coomey, J.H., MacKinnon, K.J.-M., McCahill, I.W., Khahani, B., Handakumbura, P.P., Trabucco, G.M., Mazzola, J., Leblanc, N.A., Kheam, R., Hernandez-Romero, M., et al. (2024). Mechanically induced localization of SECONDARY WALL INTERACTING bZIP is associated with thigmomorphogenic and secondary cell wall gene expression. bioRxiv, 2021.02.03.429573. https://doi.org/10.1101/2021.02.03.429573.

71. Konishi, M., and Yanagisawa, S. (2015). Transcriptional repression caused by Dof5.8 is involved in proper vein network formation in Arabidopsis thaliana leaves. J. Plant Res. *128*, 643–652.

72. Dorrity, M.W., Alexandre, C.M., Hamm, M.O., Vigil, A.-L., Fields, S., Queitsch, C., and Cuperus, J.T. (2021). The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. Nat. Commun. *12*, 3334.

73. Guo, Y., Qin, G., Gu, H., and Qu, L.-J. (2009). Dof5.6/HCA2, a Dof transcription factor gene, regulates interfascicular cambium formation and vascular tissue development in Arabidopsis. Plant Cell *21*, 3518–3534.

74. Kubo, H., Kishi, M., and Goto, K. (2008). Expression analysis of ANTHOCYANINLESS2 gene in Arabidopsis. Plant Sci. *175*, 853–857.

75. Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. Nature *425*, 257–263.

76. Cassan-Wang, H., Goué, N., Saidi, M.N., Legay, S., Sivadon, P., Goffner, D., and Grima-Pettenati, J. (2013). Identification of novel transcription factors regulating secondary cell wall formation in Arabidopsis. Front. Plant Sci. *4*, 189.

77. Mirabella, R., Rauwerda, H., Allmann, S., Scala, A., Spyropoulou, E.A., de Vries, M., Boersma, M.R., Breit, T.M., Haring, M.A., and Schuurink, R.C. (2015). WRKY40 and WRKY6 act downstream of the green leaf volatile E-2-hexenal in Arabidopsis. Plant J. *83*, 1082–1096.

78. Galway, M.E., Masucci, J.D., Lloyd, A.M., Walbot, V., Davis, R.W., and Schiefelbein, J.W. (1994). The TTG gene is required to specify epidermal cell fate and cell patterning in the Arabidopsis root. Dev. Biol. *166*, 740–754.

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

80. Kamiya, T., Borghi, M., Wang, P., Danku, J.M.C., Kalmbach, L., Hosmani, P.S., Naseer, S., Fujiwara, T., Geldner, N., and Salt, D.E. (2015). The MYB36 transcription factor orchestrates Casparian strip formation. Proc. Natl. Acad. Sci. U. S. A. *112*, 10533–10538.

81. Chen, K., Li, G.-J., Bressan, R.A., Song, C.-P., Zhu, J.-K., and Zhao, Y. (2020). Abscisic acid dynamics, signaling, and functions in plants. J. Integr. Plant Biol. *62*, 25–54.

82. Ding, P., and Ding, Y. (2020). Stories of Salicylic Acid: A Plant Defense Hormone. Trends Plant Sci. *25*, 549–565.

83. Fujita, Y., Fujita, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2011). ABA-mediated transcriptional regulation in response to osmotic stress in plants. J. Plant Res. *124*, 509–525.

84. Xie, Z., Nolan, T.M., Jiang, H., and Yin, Y. (2019). AP2/ERF Transcription Factor Regulatory Networks in Hormone and Abiotic Stress Responses in Arabidopsis. Front. Plant Sci. *10*, 228.

85. Nakashima, K., Takasaki, H., Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). NAC transcription factors in plant abiotic stress responses. Biochim. Biophys. Acta *1819*, 97–103.

86. Wang, T.-J., Huang, S., Zhang, A., Guo, P., Liu, Y., Xu, C., Cong, W., Liu, B., and Xu, Z.-Y. (2021). JMJ17-WRKY40 and HY5-ABI5 modules regulate the expression of ABA-responsive genes in Arabidopsis. New Phytol. *230*, 567–584.

87. Belbin, F.E., and Dodd, A.N. (2018). ABA signalling is regulated by the circadian clock component LHY. New Phytol. *220*, 661–663.

88. Zhao, X., Dou, L., Gong, Z., Wang, X., and Mao, T. (2019). BES1 hinders ABSCISIC ACID INSENSITIVE5 and promotes seed germination in Arabidopsis. New Phytol. *221*, 908–918.

89. Wu, Y., Deng, Z., Lai, J., Zhang, Y., Yang, C., Yin, B., Zhao, Q., Zhang, L., Li, Y., Yang, C., et al. (2009). Dual function of Arabidopsis ATAF1 in abiotic and biotic stress responses. Cell Res. *19*, 1279–1290.

90. van Verk, M.C., Bol, J.F., and Linthorst, H.J.M. (2011). WRKY transcription factors involved in activation of SA biosynthesis genes. BMC Plant Biol. *11*, 89.

91. Fraser, O., Cargill, S.J., Spoel, S.H., and van Ooijen, G. (2023). Crosstalk between salicylic acid signaling and the circadian clock promotes an effective immune response. bioRxiv, 2023.11.21.568095. https://doi.org/10.1101/2023.11.21.568095.

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

93. de Torres Zabala, M., Bennett, M.H., Truman, W.H., and Grant, M.R. (2009). Antagonism between salicylic and abscisic acid reflects early host-pathogen conflict and moulds plant defence responses. Plant J. *59*, 375–386.

94. Song, L., Huang, S.-S.C., Wise, A., Castanon, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science *354*. https://doi.org/10.1126/science.aag1550.

95. Baumgart, L.A., Lee, J.E., Salamov, A., Dilworth, D.J., Na, H., Mingay, M., Blow, M.J., Zhang, Y., Yoshinaga, Y., Daum, C.G., et al. (2021). Persistence and plasticity in bacterial

gene regulation. Nat. Methods *18*, 1499–1505.

96. Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. *89*, 789–804.

97. Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. *43*, 476–481.

98. Rawat, V., Abdelsamad, A., Pietzenuk, B., Seymour, D.K., Koenig, D., Weigel, D., Pecinka, A., and Schneeberger, K. (2015). Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. PLoS One *10*, e0137391.

99. Slotte, T., Hazzouri, K.M., Ågren, J.A., Koenig, D., Maumus, F., Guo, Y.-L., Steige, K., Platts, A.E., Escobar, J.S., Newman, L.K., et al. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat. Genet. *45*, 831–835.

100. Parkin, I.A.P., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D., Nixon, J., Krishnakumar, V., Bidwell, S.L., et al. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome Biol. *15*, R77.

101. Li, Y., Pi, M., Gao, Q., Liu, Z., and Kang, C. (2019). Updated annotation of the wild strawberry Fragaria vesca V4 genome. Hortic Res *6*, 61.

102. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science *313*, 1596–1604.

103. Hosmani, P.S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L.V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., et al. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv, 767764. https://doi.org/10.1101/767764.

104. Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., and Buell, C.R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience *9*. https://doi.org/10.1093/gigascience/giaa100.

105. Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. *35*, D883–D887.

106. Bartlett, A., O'Malley, R.C., Huang, S.-S.C., Galli, M., Nery, J.R., Gallavotti, A., and Ecker, J.R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat. Protoc. *12*, 1659–1672.

107. Galbraith, D.W. (2009). Simultaneous flow cytometric quantification of plant nuclear DNA contents over the full range of described angiosperm 2C values. Cytometry A *75*, 692–698.

108. BBMap (2022). SourceForge. https://sourceforge.net/projects/bbmap/.

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

110. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

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

112. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. Nucleic Acids Res. *43*, W39–W49.

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

114. 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.

115. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) *2*, 100141.

116. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

117. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. *45*, e119.

118. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

119. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

120. Alvarez, M., Rahmani, E., Jew, B., Garske, K.M., Miao, Z., Benhammou, J.N., Ye, C.J., Pisegna, J.R., Pietiläinen, K.H., Halperin, E., et al. (2020). Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. Sci. Rep. *10*, 11019.

121. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. *42*, 293–304.

122. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. *8*, 329–337.e4.

123. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

124. Zhang, K., Zemke, N.R., Armand, E.J., and Ren, B. (2024). A fast, scalable and versatile tool for analysis of single-cell omics data. Nat. Methods *21*, 217–227.

125. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational identification of cell Doublets in Single-cell transcriptomic data. Cell Syst. *8*, 281–291.e9.

126. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

127. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

128. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.