Identification of plant transcriptional activation domains

https://doi.org/10.1038/s41586-024-07707-3

Received: 26 June 2023

Accepted: 12 June 2024

Published online: 17 July 2024

Check for updates

Nicholas Morffy[1], Lisa Van den Broeck[2], Caelan Miller[1], Ryan J. Emenecker[3,4], John A. Bryant Jr.[5], Tyler M. Lee[1], Katelyn Sageman-Furnas[1], Edward G. Wilkinson[1], Sunita Pathak[1], Sanjana R. Kotha[6], Angelica Lam[6], Saloni Mahatma[2], Vikram Pande[2], Aman Waoo[2], R. Clay Wright[5], Alex S. Holehouse[3,4], Max V. Staller[6], Rosangela Sozzani[2] & Lucia C. Strader[1]✉

Gene expression in *Arabidopsis* is regulated by more than 1,900 transcription factors (TFs), which have been identified genome-wide by the presence of well-conserved DNA-binding domains. Activator TFs contain activation domains (ADs) that recruit coactivator complexes; however, for nearly all *Arabidopsis* TFs, we lack knowledge about the presence, location and transcriptional strength of their ADs[1]. To address this gap, here we use a yeast library approach to experimentally identify *Arabidopsis* ADs on a proteome-wide scale, and find that more than half of the *Arabidopsis* TFs contain an AD. We annotate 1,553 ADs, the vast majority of which are, to our knowledge, previously unknown. Using the dataset generated, we develop a neural network to accurately predict ADs and to identify sequence features that are necessary to recruit coactivator complexes. We uncover six distinct combinations of sequence features that result in activation activity, providing a framework to interrogate the subfunctionalization of ADs. Furthermore, we identify ADs in the ancient AUXIN RESPONSE FACTOR family of TFs, revealing that AD positioning is conserved in distinct clades. Our findings provide a deep resource for understanding transcriptional activation, a framework for examining function in intrinsically disordered regions and a predictive model of ADs.

Transcription factors (TFs) are at the foundation of development and response to stimuli, and their abilities to (1) bind to a target DNA sequence and (2) recruit transcriptional coactivators or corepressors are fundamental[1]. Because TF DNA-binding domains (DBDs) are conserved and well-structured, we have a deep molecular understanding of DBD activity, and several large-scale efforts have mapped TF-binding sites[2,3]. Conversely, despite substantial recent work, many questions remain about the molecular interactions that enable TFs to recruit a diverse set of coactivator and corepressor complexes, especially in plants. Activation domains (ADs) are effector domains that recruit coactivator complexes to increase transcription. ADs tend to reside in intrinsically disordered regions (IDRs)—protein regions that lack a well-defined three-dimensional structure and are often poorly conserved as assessed by multiple sequence alignments. These IDR characteristics make ADs inherently challenging to study. Consequently, only a few plant TFs have annotated ADs (Fig. 1a). The inability to easily identify TF ADs hampers researchers' progress towards understanding TF activity, reconstructing gene regulatory networks and engineering synthetic tools.

ADs recruit the transcriptional machinery—that is, coactivators—to the TF-binding site, but the interactions with this machinery are not well-defined despite intense study[4]. ADs are classically defined by their most abundant residues, including acidic, Gln-rich and Pro-rich ADs. Among these classes, acidic ADs are the most well-studied, and contain interspersed aromatic and acidic residues. Hydrophobic motifs make large contributions to activity, and, in nuclear magnetic resonance structures, motif residues contact hydrophobic surfaces of coactivators[5]. Aromatic residues contribute most strongly to activity, whereas leucine and methionine residues have a smaller role. The acidic residues sometimes mediate fast, low-affinity binding[6–8]. A proposed unifying idea for acidic ADs is that the acidic residues keep the hydrophobic residues exposed to solvent, allowing for coactivator interaction[9,10]. However, little sequence similarity exists among strong ADs and the sequence features that govern AD activity remain poorly understood. Moreover, despite the crucial role of ADs, they are difficult to identify, in part because approaches for identifying ADs are labour intensive. This is the case in plant systems, which have traditionally relied on a combination of genetic and cross-species approaches to identify ADs. A previous study was able to identify the transcriptional activity of tens of TFs[11], but lacked the resolution to interrogate AD activity specifically.

In this work, we experimentally identify plant transcriptional ADs on a genomic scale using an approach that we call the plant activation domain identification (PADI) high-throughput assay (Fig. 1b), and develop a set of rules for the future identification of ADs using a deep learning algorithm that we name the transcriptional activation domain activity (TADA) network. This information, together with the constant

[1]Department of Biology, Duke University, Durham, NC, USA. [2]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. [3]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA. [4]Center for Biomolecular Condensates, Washington University in St. Louis, St. Louis, MO, USA. [5]Biological Systems Engineering, Virginia Tech, Blacksburg, VA, USA. [6]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. ✉e-mail: lucia.strader@duke.edu

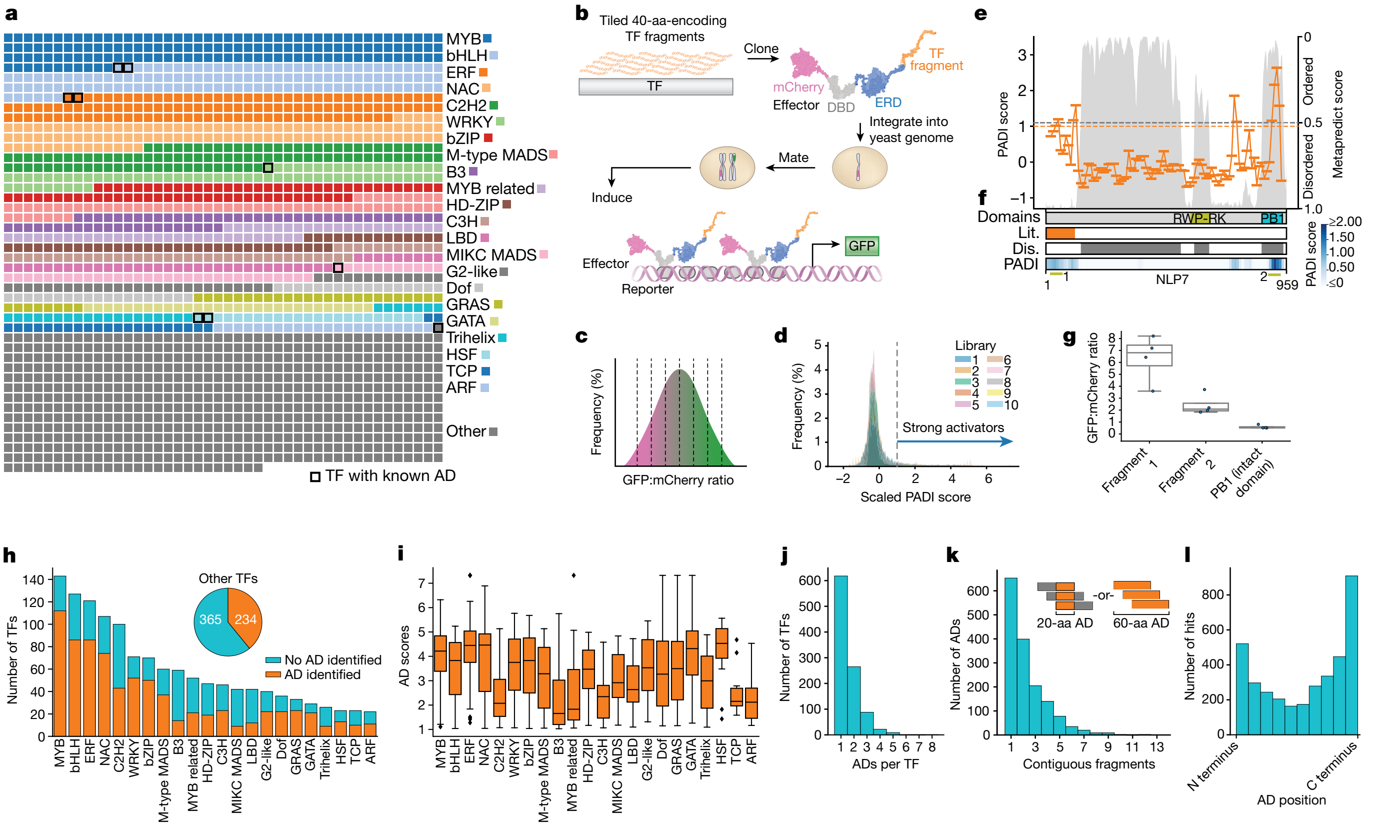


Fig. 1 | High-throughput tiling of *Arabidopsis* TFs uncovers thousands of ADs. a, Waffle plot of the 1,918 *Arabidopsis* TFs analysed. Those with previously identified ADs are marked with a black box. **b**, Schematic of PADI. Ten pooled libraries of synthetic TFs were integrated into the yeast URA3 locus before mating to yeast carrying a 5×UAS reporter. aa, amino acids. ERD, estrogen receptor D. **c**, After induction, cells were flow sorted into bins on the basis of GFP:mCherry ratio to assess transcriptional output. **d**, Scaled AD (PADI) score; fragments with a PADI score ≥ 1 (one standard deviation from the mean) were considered strong activators. **e**, Hits need to be filtered by disorder. PADI (orange) and predicted disorder (white background) scores for NLP7 show strong activity in disordered regions as well as in ordered regions (grey background) that overlap with the known PB1 domain. The orange (PADI = 1) and grey (Metapredict[15] score = 0.5) dashed lines are considered cut-offs for activation and disorder, respectively. **f**, Schematic of NLP7 protein domains. From top to bottom: ordered domains (Uniprot Q84TH9, olive and teal); previously annotated as containing AD activity[13] in the literature (Lit.; orange); predicted disorder (Dis.; grey); and PADI scores (blue). **g**, Forty-amino-acid fragments (underlined in **f**) or intact PB1 domain (teal in **f**) were tested for activation activity using a modified version of the PADI assay ($n = 4$ independent experiments). **h**, Distribution of identified ADs across *Arabidopsis* TF families ($n \geq 22$). **i**, Distribution of highest-scoring hits from each TF in each family ($n \geq 11$). **j**, Distribution of the number of ADs identified per *Arabidopsis* TF. **k**, Distribution of the number of contiguous hits identified per identified AD. Contiguous hits could be indicative of a short AD contained in neighbouring fragments or of an extended AD for which a subset of residues is sufficient to activate transcription; our data cannot distinguish between these. **l**, The distribution of hit locations revealed a bias towards the amino and carboxy termini of proteins. The data in **h**–**l** have been filtered for hits that are present in IDR regions of the parent TF. Unfiltered data can be found in Extended Data Fig. 1g. All box plots show the interquartile range and the median. Whiskers are 1.5 times the interquartile range.

development and improvement of genome-editing techniques, will have a transformative effect on the design of synthetic ADs that will allow for tunable transcriptional activation.

PADI

At present, determining the transcriptional activation activity of proteins solely on the basis of protein sequence remains a challenge, and limits our understanding of transcriptional regulation. To address this knowledge gap, we implemented an experimental approach to identify ADs in all annotated *Arabidopsis* TFs. We used a high-throughput yeast-based system[12] to systematically identify ADs from thousands of protein fragments. Although this system provided us with the ability to identify fragments of *Arabidopsis* TFs that could activate transcription at scale, it has two limitations: (1) it is possible that *Arabidopsis*-specific ADs could be missed in this assay; and (2) it is possible that *Arabidopsis* TF fragments that activate transcription in yeast might not activate transcription in *Arabidopsis*.

We tested the transcriptional AD activity of 1,918 TFs from *Arabidopsis*, representing more than 20 TF families (Fig. 1a). To assess AD activity, we designed synthetic TFs comprising an N-terminal mCherry tag, a mouse DBD and an inducible nuclear localization domain[12]. Each synthetic TF contained a 40-amino-acid fragment derived from an *Arabidopsis* TF and was integrated into the yeast genome at the *URA3* locus to ensure the presence and stability of a single synthetic TF in each transformant for reliable assessment (Fig. 1b).

We screened each TF in 40-amino-acid fragments, with a step size of 10 amino acids, testing 68,441 fragments. Oestradiol-induced libraries were sorted on the basis of their relative GFP:mCherry ratio, and subsequent sequencing was performed to determine the relative abundance of fragments in each bin (Fig. 1c). Fragments with high PADI scores were found predominantly in bins of a high GFP:mCherry ratio, suggesting strong activation. To allow comparison across ten libraries, we applied *z*-score normalization to assign an AD score (PADI score) to each fragment (Fig. 1d and Extended Data Fig. 1). Of the fragments tested, 6,205 showed strong activation, as denoted by scoring at or above a threshold of one standard deviation from the mean (PADI score ≥ 1). Using this definition, approximately 10% of tested fragments exhibited strong transcriptional activation activity.

When examining PADI hits (PADI score ≥ 1) in TFs with previously known AD locations, such as NIN-LIKE PROTEIN 7 (NLP7)[13], we noticed that we often observed extraneous hits in regions of the protein that were not previously associated with the transcriptional activity of these proteins. Consistently, these extraneous hits were observed in protein regions that were predicted to be well-folded (Fig. 1e,f). We hypothesized that fragments from folded regions have sequence attributes that are capable of recruiting the transcriptional machinery when taken out of their well-ordered context; however, these fragments do not exhibit AD activity when in folded regions. Accordingly, previous studies have shown that eukaryotic ADs are commonly found in IDRs[5,14]. To test this idea, we further investigated fragments and domains from NLP7. NLP7 possesses an N-terminal IDR that is both necessary and sufficient to activate transcription in vivo[13]. PADI identified a high-scoring region that overlapped with the N-terminal region of NLP7, as well as fragments from the C-terminal PHOX and BEM1 (PB1) oligomerization domain (Fig. 1e,f). In contrast with the NLP7 N-terminal IDR, the NLP7 PB1 domain has been shown not to activate transcription in vivo[13]. To determine whether the PB1 domain can activate transcription on its own in our PADI system, we tested the full PB1 domain, as well as high-scoring PADI fragments from the N terminus and PB1 domain. As previously reported[13], we observed that the intact PB1 domain did not activate transcription in our experimental system (Fig. 1g). This result suggests that certain fragments when removed from their ordered domain context might exhibit AD-like behaviours that do not reflect their native function. In the PADI dataset, 43% of high-scoring fragments are from globular domains, which have sequence features consistent with ADs, but might lack AD activity in their folded context, and thus we urge caution for researchers following up on specific ADs from these regions. Therefore, in our dataset, we report both PADI data and IDR predictions generated by Metapredict[15] (Supplementary Table 1 and Supplementary Data 1).

We found that 53% of the 1,918 tested *Arabidopsis* TFs contained at least one AD, defined as a PADI hit that originated from an IDR. These ADs were not evenly distributed across TF families. For instance, 79% of the MYB family exhibited at least one AD, whereas only 20% of the B3 family possessed a strong AD (Fig. 1h). The strength of AD output varied across families (Fig. 1i), with a correlation between prevalence of ADs and transcriptional output by family (Extended Data Fig. 2e). Of those TFs with an identified AD, a majority (61%) possessed only a single AD, defined as a single region of one or more contiguous hits (Fig. 1j). Furthermore, these ADs (81%) were found predominantly within three or fewer contiguous fragments (Fig. 1k). From our data, it is challenging to ascertain whether these contiguous regions share a small, overlapping core AD or whether they work in a combinatorial manner to strongly activate transcription. ADs were preferentially located at either the N terminus or the C terminus of *Arabidopsis* TFs (Fig. 1i). Together, these results present a thorough survey of the presence and location of ADs in the *Arabidopsis* genome. By performing this screen in a synthetic yeast system, we identified more than 1,500 ADs in more than 1,000 TFs. This not only provides us with a better understanding of how *Arabidopsis* TFs function, but also allows us to interrogate the relationship between sequence features in PADI hits to systematically understand which combinations of features contribute to AD activity.

TADA network and sequence features

Understanding the features that correlate with strong activation will help researchers gain an understanding of the mechanisms that underlie transcriptional regulation. Unlike structured domains, truncated ADs retain activity[16,17], consistent with their intrinsic disorder[18]. Residue enrichment has historically been used to classify ADs, and the acidic aromatic class of ADs is the best studied and most common. The acid-exposure model suggests that negatively charged acidic residues, such as Asp and Glu, serve to expose otherwise-buried bulky aromatic and hydrophobic residues, including Phe, Trp, Leu and Tyr, which act as contact points with the transcriptional machinery to activate transcription. In line with these findings, we plotted all tested fragments in this feature space to further explore the relationship between residue enrichment and ADs. Our analysis showed that an enrichment in aromatic residues and net negative charge correlated with higher average AD scores (Fig. 2a). However, after further examination focusing on hits that scored above the threshold (Fig. 2a), we observed a significant number of hits with neutral or positive charges, as well as fragments with few aromatic residues. These findings suggest that additional classes of ADs are present in our library.

To gain a more comprehensive understanding of the underlying sequence features associated with transcriptional activity, we validated and refined proposed models for AD activity. Moreover, we reasoned that by examining the features that contribute to strong activation without bias towards existing models, we could potentially uncover novel paradigms of transcriptional activation. Because we observed a strong prevalence of the acidic aromatic class of ADs (Fig. 2a), we first investigated the contribution of individual acidic and aromatic amino acids towards PADI scores (Fig. 2b). When examining negatively charged residues (Asp and Glu), we found a positive trend between Asp frequency and PADI scores, but not between Glu frequency and PADI scores (Fig. 2b). Thus, within the acidic-exposure model, the negative charge provided by a shorter side chain (Asp) might allow for a greater exposure of the functional aromatic residues than the negative charge provided by a longer side chain (Glu). Amongst the examined aromatic residues (Trp, Phe, Tyr and Leu), we found that only Phe enrichment showed a positive trend with PADI score. Despite these mild trends, no individual amino acid showed a significant correlation with PADI score, suggesting that single aromatic residues are unlikely to be good predictors of AD activity.

Owing to the limited correlation between individual amino acids and AD activity, we reasoned that higher-order and more complex correlations might be descriptive and/or predictive for AD activity. As such, we recognized the need to investigate which features strongly contribute to the high-scoring fragments in our dataset. To gain insights into the sequence features and structural properties that determine AD function, we developed a neural-network-based approach. Our neural network, which we name the transcriptional activation domain activity (TADA) network, incorporates convolutional, activation and recurrent layers. By doing so, it captures both linear and non-linear relationships between input features and AD predictions. Considering that AD functionality is not determined solely by the primary amino acid sequence and that ADs lack a defined secondary structure, we opted to predict ADs on the basis of side-chain properties and IDR descriptors, rather than relying solely on the raw sequence. To capture this information, we computed 42 sequence properties, referred to as features, using a sliding window of 5 amino acids (Fig. 2c). These computations resulted in a 42 × 36 matrix for each tested fragment. In our first iteration of the neural network[19], we trained TADA on these 42 features computed on a dataset of 75,845 random peptides[20]. Notably, our neural network outperformed an existing classification neural network (ADpred), which was also trained on these random peptides, in both sensitivity and F1 score[19]. This suggests the effectiveness of our approach in predicting ADs. To capture the specific characteristics of identified ADs, we retrained our neural network on the PADI dataset, which consisted of 64,552 non-hits and 6,385 hits. Given the data imbalance, we implemented several cost-sensitive approaches, and stratified splitting of the dataset resulted in 70% for training, 20% for validation and 10% held out as a test set (see Methods). On the test set, our neural network slightly outperformed the previous iteration trained either on random peptides alone or on a combination of random peptides and PADI data across all

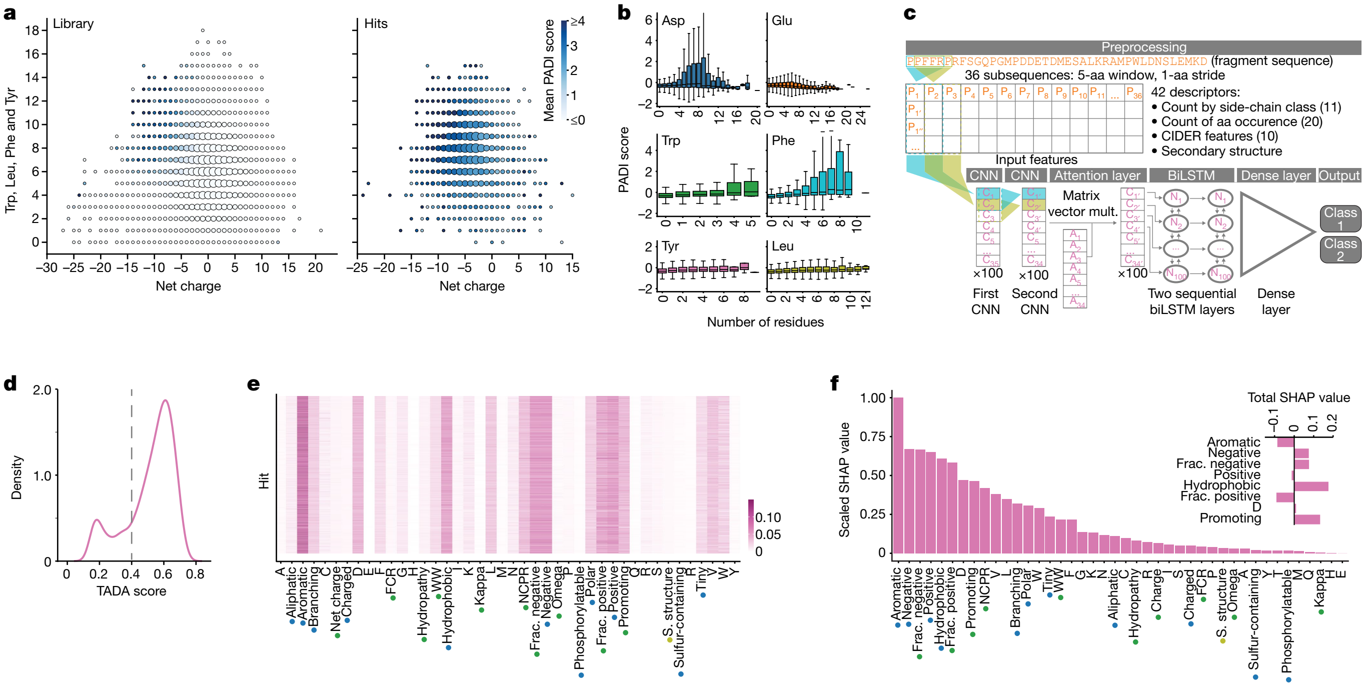


Fig. 2 | Using AD sequence features to create a predictive model. a, Scatter plot showing the distribution of net charge and the number of Trp, Leu, Phe and Tyr residues in the library (left) and hits (right). The size of each point represents the number of fragments at each coordinate and the colour corresponds to the mean PADI score fragments at that coordinate. **b**, Box plots showing the distribution of PADI scores for fragments on the basis of the number of Asp, Glu, Trp, Phe, Tyr and Leu residues per fragment. Boxes represent interquartile range with the median drawn within the box. Whiskers are 1.5 times the interquartile range (*n* = 1–44,633 fragments). **c**, TADA architecture and 42 descriptors, including counts of side-chain class, counts of amino acid occurrence, attributes calculated by LocalCIDER[38] and secondary structure prediction by Metapredict[15]. From these 42 descriptors, TADA uses two convolutional neural network (CNN) layers, an attention layer, two sequential bi-directional long short term memory (biLSTM) layers and a dense layer to classify sequences. Mult., multiplication. **d**, TADA score across PADI hits. Using TADA to predict hits from the PADI dataset suggests that a TADA cut-off score of 0.4 will capture most fragments that activate transcription. **e**, SHAP values averaged across the 26 subsequences for each input feature, as calculated for the test dataset classified as fragments scoring higher than 1. Features derived by counting number of residues by side-chain property (blue), derived from LocalCIDER[38] (green) and the Metapredict[15]-based secondary structure score (olive) are shown. **f**, Normalized SHAP values ranked from most important to least important for fragments scoring higher than 1. Inset, the top eight features plotted as having a positive or negative effect on prediction. Features derived by counting number of residues by side-chain property (blue), derived from LocalCIDER[38] (green) and the Metapredict[15]-based secondary structure score (olive) are shown. FCR, fraction charged residue; frac., fraction; NCPR, net charge per residue; s., secondary; WW, Wimley and White hydropathy.

performance metrics (Extended Data Fig. 3a). Thus, we reasoned that certain AD characteristics are not adequately captured within a random peptide dataset alone. In addition, on a dataset of human TFs, TADA outperformed existing predictors, including PADDLE, ADpred and a composite model (Extended Data Fig. 4). A TADA score of 0.4 captured most true hits and is used for our score cutoff (Fig. 2d). Overall, TADA achieved a high F1 score of 93.40% and an area under the precision recall curve (AUPR) of 97.31%, indicating that our unique encoding approach and neural network design can advance the prediction of ADs (Extended Data Fig. 3a).

To gain insights into the contribution of each of the 42 features to TADA network predictions, we determined the effect of the individual input features. To this end, we used Shapley additive explanations (SHAP) analysis[21] and examined the local and global effects of the features on TADA's predictions. Unlike our previous analyses of sequence features (Fig. 2b), which rely on linear correlations, TADA and SHAP analysis capture non-linear relations and uncover complex correlations with AD identification. To identify local explanations and thus aim to explain individual predictions, we computed the effect of each feature for each AD fragment with SHAP (Fig. 2e). We found that the total counts of aromatic residues (Trp, Phe and Tyr), negative residues (Asp and Glu), positive residues (Lys, Arg and His) and hydrophobic residues (Trp, Phe, Leu, Val, Ile, Cys and Met) emerged as key features across all hits. To assess the overall importance across all fragments, and thus the global explanation, we also computed the total absolute effect of each feature and ranked them according to importance (Fig. 2f). Consistent with local explanations (Fig. 2a,b), we found that aromatic, negative, hydrophobic and positive residues were most important. In addition, the fraction of negatively or positively charged residues and the fraction of residues predicted to be 'disorder-promoting' strongly affected AD predictions (Fig. 2f). Notably, aromatic residues collectively had a greater importance than individual aromatic residues (Fig. 2f). When examining our hits in aggregate, we found that many previously known drivers of AD activity, such as the presence of aromatic, hydrophobic and negative residues, were important to plant ADs; however, we also found differential contributions of negatively charged residues (Glu > Asp) and aromatic residues (Trp > Phe, Tyr).

By using the deep learning interpretability technique SHAP, we identified key properties that are important for AD prediction (Fig. 2f). We then used the most important and predictive features to perform unsupervised classification of ADs. To this end, we normalized the retrieved importance scores and selected eight features above a threshold of one standard deviation from the mean (Fig. 2f (inset) and Extended Data Fig. 4e), which allowed us to retain the strongest signals from the most influential features. To visualize and analyse the resulting AD fragments, we used a combination of principal component analysis (PCA) and *t*-distributed stochastic neighbour embedding (t-SNE). By projecting all hits onto a two-dimensional (2D) space, we observed distinct clusters, ultimately identifying a total of six AD subtypes (Fig. 3a and Supplementary Table 2).

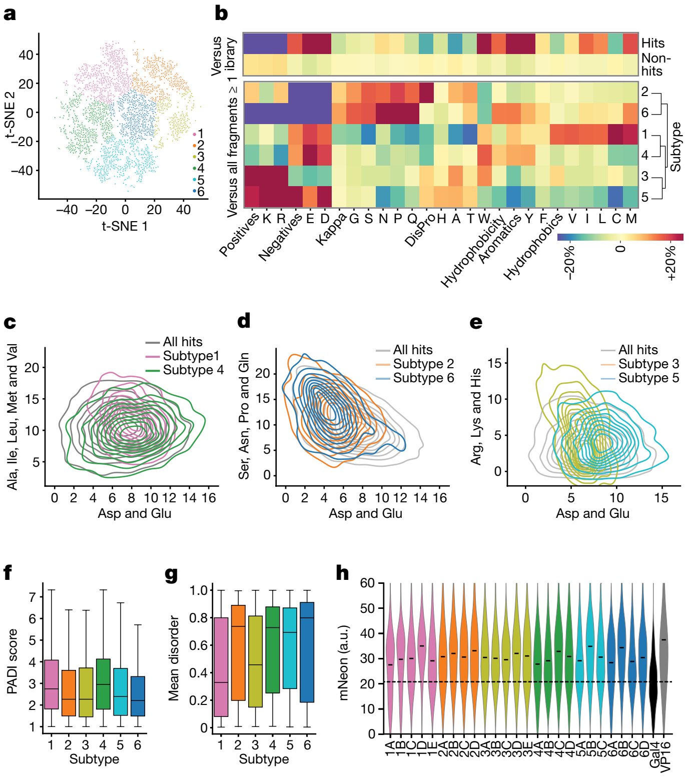


Fig. 3 | AD subtypes show distinct compositional biases. a, The ADs were divided into six subtypes on the basis of *k*-means clustering of the 2D t-SNE output. t-SNE was performed on a ten-component PCA of the eight most important features and their SHAP values. **b**, Top, comparative analysis of the fragment composition of AD versus non-AD fragments in relation to the library as a whole. Bottom, comparative analysis of fragment composition of each AD subtype in relation to all AD fragments (scoring above 1). **c**–**e**, Distribution of subtypes in feature space against all hits (grey) for subtypes 1 and 4 (**c**), subtypes 2 and 6 (**d**) and subtypes 3 and 5 (**e**) on the basis of the enrichment of depicted amino acids. **f**, PADI score by subtype ($n \geq 625$). **g**, Mean disorder by subtype ($n \geq 625$). **h**, All examined yeast-identified hits promote transcription in plant cells. Protoplasts were transfected with a synthetic TF containing an N-terminal mScarlet-I tag, the Gal4 DBD and the identified 40-amino-acid PADI hit, or just the Gal4 DBD (Gal4). The cells were also transfected with a reporter of NLS-mNeonGreen driven by 5× Gal4 UAS sites. The mNEON reporter was assayed in mScarlet-positive cells using flow cytometry. Violin plots depict mNEON signal in arbitrary units (a.u.) with the mean mNEON signal depicted as a black bar. All examined hits were significantly different from the control (Student's *t*-test; $P \leq 0.0001$) ($n \geq 520$ cells from 3 independent transfections). All box plots show the interquartile range and the median. Whiskers are 1.5 times the interquartile range.

To identify whether these subclasses have divergent features that might indicate distinct functionality, we investigated sequence features of these subtypes (Fig. 3b and Extended Data Fig. 2). The six subtypes not only differed from one another in strongly predictive features, but also exhibited other differentiating features. Subtypes 1 and 4 are enriched in negatively charged residues when compared with all identified ADs, and are likely to represent two types of acidic ADs. Subtype 4 is enriched in aromatic residues whereas subtype 1 is enriched in aliphatic residues (Fig. 3c). We hypothesize that these subtypes function through the acidic-exposure model, in which negative charge leads to the exposure of residues for interaction with the transcriptional machinery.

Subtypes 2 and 6 are relatively depleted in negatively charged residues and enriched in Ser, Pro, Asn and Gln residues, which suggests that a loss of negative charge can be compensated for by increases in these residues (Fig. 3d). These combinations of features might promote side-chain exposure, whereby Ser phosphorylation might generate additional negative charge and Pro and/or Asn residues disrupt structure to allow for expanded peptide backbones. Considering the enrichment for Pro and Gln residues, we propose that subtypes 2 and 6, respectively, represent the Pro-rich and Gln-rich AD classes.

Subtypes 3 and 5 are relatively enriched for positively charged residues amongst our hits. Subtype 5 also showed an enrichment in negatively charged residues and a reduction in the total number of crucial aromatic residues (Fig. 3e), raising the possibility that subtype 5 AD relies on a limited number of aromatic residues for interaction with the transcriptional machinery. By contrast, subtype 3 does not show a compensatory enrichment in negative charge (Fig. 3b), suggesting that it has a different mechanism of action to that of subtypes 1, 4 and 5. All subtypes displayed a range of PADI scores (Fig. 3f) and predicted radii of gyration (Fig. 3g).

Because these subtypes represent sequence features of fragments from *Arabidopsis* TFs that can elicit transcriptional output in yeast, we thought that they would be a good test of whether identified hits are active in *Arabidopsis*. For this test, we cloned representative fragments from each subtype (Extended Data Fig. 2i,j) into a reporter system consisting of the Gal4 DBD fused to the tested fragment. These TFs were cloned into plasmids carrying an mNeonGreen reporter driven by 5× Gal4 UAS sites and a minimal promoter to result in a 1:1 ratio of effector and reporter. In *Arabidopsis* mesophyll protoplasts, representatives from each subtype activated transcription (Fig. 3h), confirming that each subtype identified in our yeast-based PADI screen is active in plant cells.

Together, our network and biophysical analyses provide a framework for correlating functional outputs of IDR protein sequences with their features. The six identified AD subtypes differ in the enrichment of crucial features, compared with all hits, and serve as a foundation for further investigation of ADs and their features beyond the acidic–aromatic paradigm.

PADI and TADA validation

To validate PADI-identified ADs and examine their correspondence with previously reported ADs, we investigated the only 9 *Arabidopsis* TFs with ADs mapped to within 100 amino acids whose activities have been shown to be important for function in plants[13,22–28] (Fig. 4a). For each previously published AD (orange), we both identified the same region in our yeast-based assay (PADI, blue) and predicted the AD using TADA (pink). These identified and predicted ADs were consistently found in predicted IDRs. We also identified PADI hits in most *Arabidopsis* TFs that have recently been found to activate transcription in a transient tobacco-based assay[11] (Extended Data Fig. 5a). Our identification of PADI hits that match previously reported *Arabidopsis* domains, along with our testing of 24 additional hits in protoplasts, are consistent with the possibility that PADI accurately identifies *Arabidopsis* ADs; however, it is still possible that some PADI hits will not be active in *Arabidopsis* and that some *Arabidopsis* ADs do not activate transcription in yeast.

To expand our validation, we examined ADs in the 23-member AUXIN RESPONSE FACTOR (ARF) family of TFs. Transcriptional activity for 4 members of this family has been mapped to their long central IDRs[29,30] (Fig. 4b), providing us with the opportunity to validate ADs in a family for which the ADs had not been fully mapped. We identified ADs in the central IDRs of each of these TFs, but also found strongly scoring regions in their well-ordered PB1 domains (Fig. 4b). To test our hypothesis that only the high-scoring fragments from IDRs are functional in the context of an intact TF, we used the well-developed yeast synthetic auxin signalling system[31], focusing on ARF7 (Fig. 4d). Whereas full-length ARF7 showed strong reporter activation, ARF7 lacking either AD1 (ΔAD1) or AD2 (ΔAD2) exhibited a reduced transcriptional output while retaining activity, suggesting that each individual AD is sufficient to confer partial activity. Deletion of both ARF7 ADs from the IDR (ΔAD1ΔAD2) abolished all ARF7 transcriptional activity,

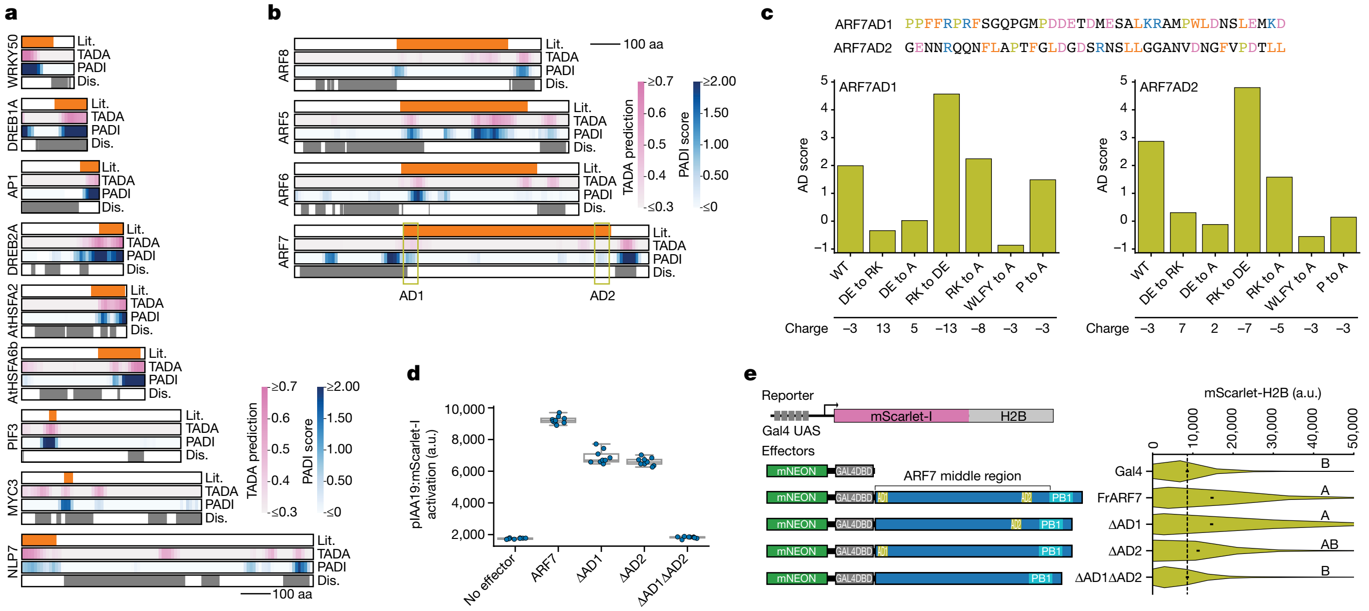


Fig. 4 | Validation of identified ADs. a, Schematic of WRKY50[22], DREB1A[23], AP1[24], DREB2A[25], AtHSFA2[26], HtHSFA6b[26], PIF3[27], MYC3[28] and NLP7[13] protein domains previously annotated as containing AD activity (orange), TADA scores (pink), PADI scores (blue) and the predicted disorder (white). **b**, Schematic of ARF8, ARF5, ARF6 and ARF7[29] protein domains previously annotated as containing AD activity (orange), TADA scores (pink), PADI scores (blue) and the predicted disorder (white). The two identified ADs in the ARF7 middle region are annotated as AD1 and AD2. **c**, ARF7AD1 and ARF7AD2 variants alter transcriptional output. AD sequences were modified as indicated and tested in the PADI assay. **d**, Deletion of ARF7AD1 or ARF7AD2 results in decreased ARF7 output in a reconstructed yeast system. pIAA19:mScarlet-I reporter fluorescence was measured by flow cytometry with the results depicted as median values of three transformants and three replicate experiments (20,000 cells per replicate) with underlaid box plots. Box plots show the interquartile range and the median. Whiskers are 1.5 times the interquartile range. **e**, Deletion of ARF7AD1 or ARF7AD2 results in decreased ARF7 output in plant cells. Protoplasts were transfected with a synthetic TF containing an N-terminal mNEON tag, the Gal4 DBD and the middle region and C terminus of ARF7 with or without the identified ADs (FrARF7, ΔAD1, ΔAD2 and ΔAD1ΔAD2) or just the Gal4 DBD (Gal4), along with an mScarlet-I fused to the histone 2B (H2B) reporter driven by 5× Gal4 UAS. The mScarlet reporter was assayed in mNEON-positive cells using flow cytometry. Violin plots depict mScarlet signal in arbitrary units with black bars marking the average mScarlet signal ($n \geq 2,212$ cells from 4 independent transfections). Letters are statistically significant groupings based on the Tukey HSD test with an alpha-level of 0.01.

which indicates that the high-scoring fragments identified by PADI and TADA from well-folded regions do not contribute to transcriptional activation in the context of the intact ARF7 protein.

The two distinct ADs found in ARF7 provided us with an ideal opportunity to interrogate the roles of ADs from different subtypes, because ARF7AD1 is a subtype 5 and ARF7AD2 is a subtype 4 AD, and they have distinct sequence features (Fig. 4c). We generated and examined ARF7AD1 and ARF7AD2 variants in our PADI system, and found that systematic substitution of residues that were identified by SHAP analysis as impactful for AD activity alters the output for each. Mutating negative residues to either positive residues or Ala led to a loss of activity (Fig. 4c). Conversely, mutating positively charged residues to negatively charged residues resulted in an increase in activity. In addition, mutating aromatic and hydrophobic residues resulted in decreased AD activity. We found that mutating Pro negatively affected the activity of ARF7AD2, but not that of ARF7AD1 (Fig. 4c), suggesting that Pro has distinct molecular roles in these two ADs. Together, these results further validate our classification of distinct AD types and suggest that different AD types achieve similar activities through distinct combinations of features.

To validate identified ARF7 ADs in planta, we created FrankenARF7 (FrARF7), which consists of the Gal4 DBD fused to the ARF7 central IDR (middle region) and PB1 domain. FrARF7 was cloned into a plasmid that also carried an mScarlet-H2B reporter driven by 5× Gal4 UAS sites and a minimal promoter to result in a 1:1 ratio of effector and reporter (Fig. 4e). In *Arabidopsis* mesophyll protoplasts, wild-type FrARF7 exhibited strong transcriptional activation, whereas FrARF7$^{\Delta AD1\Delta AD2}$ did not activate transcription, confirming that the two ADs identified in the ARF7 middle region are necessary for transcriptional activity. FrARF7$^{\Delta AD1}$ showed no loss of activity, whereas FrARF7$^{\Delta AD2}$ showed a mild reduction of transcriptional output (Fig. 4e). Thus, ARF7AD2 is sufficient for transcriptional activity in this system and the contribution of ARF7AD1 can only be unmasked when AD2 is missing. In the future, examining ADs from distinct subtypes will allow the subfunctionalization of ADs in different cellular contexts and tissue types to be investigated.

AD evolution in the ARF family

The well-studied ARF family has three deeply conserved clades: clade-A ARFs are considered transcriptional activators, whereas clade-B and clade-C ARFs are considered transcriptional repressors; activator and repressor functions are encoded in the intrinsically disordered middle region of the ARF TFs[32]. The rich evolutionary history and dichotomy of functions makes the ARFs an ideal TF family through which to interrogate the evolution of transcriptional function and ADs. We identified ADs within the middle regions of all three *Arabidopsis* ARF clades (Fig. 5a). Each clade-A ARF contained one or more AD (Fig. 5a), consistent with their historical definition as ‘activator ARFs’. Clade-B and clade-C ARFs generally lacked ADs, consistent with their presumed roles as transcriptional repressors; however, some members of these ‘repressor ARF’ families had high-scoring AD regions (Fig. 5a).

To examine the evolution of ARF ADs, we performed a new experiment with an additional 11 species that span the flowering plant lineage (Fig. 5b, Extended Data Fig. 6 and Supplementary Table 3). Among the 112 clade-A ‘activator’ ARF middle regions tested, we found that 97% contained at least one AD, defined as a PADI score of at least 1. Notably, approximately 58% of clade-B and 62% of clade-C ARFs also contained an AD (Fig. 5c). Consistent with their proposed roles in vivo, clade-A

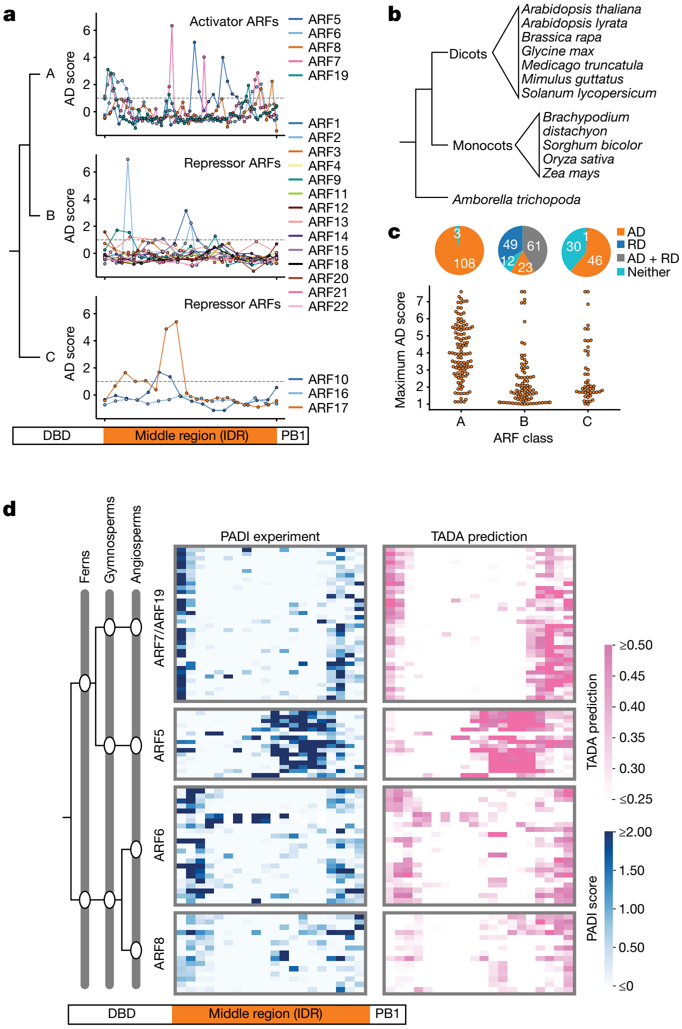


Fig. 5 | The position of ARF ADs has remained constant over evolutionary time. a, *Arabidopsis* class A ARFs are enriched in ADs. **b**, Flowering plant species examined in the ARF evolution library. **c**, A breakdown of the number of ARFs with at least one AD region (orange), putative RD (blue), AD and putative RD (grey) and neither AD nor RD (teal) in each of the three clades and the maximum PADI score found in each of the tested ARFs that scored above the threshold. RDs were identified by searching for the following motifs in the ARF fragments: LxLxL, [R/K]LFG[F/I/V], DLNxxP and LxLxPP (where x denotes any amino acid)[33,39]. **d**, Heat maps showing the average PADI score and TADA prediction scores of ARF middle-region fragments from different clade-A subclades. Each column is 5% of the length of the tested ARF middle region and each row is one examined ARF. When multiple fragments reside within a column, the colour represents the mean PADI score (blue) or TADA prediction (pink) of all fragments within that window.

fragments scored higher on average than did those from either the clade-B or the clade-C ARFs. In plants, the presence of a repression domain (RD), such as those found in clade-B ARFs[33], leads to transcriptional repression even if an AD is also present[34]. We therefore examined the co-occurrence of annotated RDs[33] and fragments that elicited a transcriptional response in the tested ARFs, and found that 76% of clade-B ARFs have annotated RDs, on the basis of the presence of known motifs[33], in contrast with clade-A and clade-C ARFs (Fig. 5c). Thus, although several clade-B ARFs contain ADs, they might not act as transcriptional activators in planta owing to the presence of strong RDs.

To better understand the evolution of ARF ADs, we focused our attention on the clade-A 'activator' ARFs, which exist in four functionally conserved subclades in the flowering plants[35]. Although ADs within the subclades showed minimal sequence similarity (Extended Data Figs. 5–8), they shared common positioning (Fig. 5d). AD fragments from the ARF5 subclade were distributed in the centre of the middle region, whereas ADs in the other three subclades showed a preference for proximity to the DBD (N terminus of the middle region) or the PB1 domain (C terminus of the middle region). This result of conserved positioning without conserved sequence suggests selective pressure on the locations of ADs, even when they reside in extended regions of intrinsic disorder and low complexity, as is found in the ARF middle region. Selection on AD position can also be found within TF families in *Arabidopsis*. The MYB family, for example, shows a preference for C-terminal ADs (Extended Data Fig. 9a). These findings suggest that functionality is encoded not only in AD sequence, but also in AD location.

To investigate whether the TADA network, trained on an *Arabidopsis* dataset, can predict ADs in other plant species, we predicted ARF ADs in the species examined in our ARF evolution dataset. On this unseen dataset, TADA achieved an AUPR of 96.14% and outperformed existing methods in terms of accuracy and F1 score, indicating that TADA generalizes well to other plant species (Extended Data Fig. 9c,d). The predicted ADs in the clade-A ARFs overlapped with our PADI findings (Fig. 5d), which suggests that training the data on *Arabidopsis* ADs is sufficient for the prediction of ADs across the flowering plants. Together, these results provide evidence for the functional and positional conservation of ADs throughout the more than 145 million years of angiosperm evolution.

Discussion

The identification and characterization of ADs has lagged behind that of DBDs across all eukaryotic taxa, which has hindered a comprehensive understanding of TF function. Unlike DBDs, which are easily identified by amino acid sequence, ADs are typically located within IDRs and defined by biochemical features rather than by linear sequence, thus posing a challenge for traditional bioinformatic methods. To overcome this limitation, we conducted the PADI high-throughput assay and developed the TADA prediction network to identify ADs within plant TFs. In our study, we assayed 79,298 sequences from 2,316 plant TFs and discovered 2,069 ADs in 1,275 TFs (Extended Data Fig. 10). Our identification and classification of ADs represents a first step towards comprehending plant TF function on a genome-wide scale. Similar previous studies using random peptides[20], yeast peptides[4] and human peptides[36] have resulted in considerable advances in our knowledge of AD activity, revealing that 59% of yeast[4] and 14% of human[36] TFs contain identified ADs. From this study, 53% of *Arabidopsis* TFs contain at least one region that can activate transcription in a yeast-based assay.

A limitation of our study is the use of yeast to identify fragments of *Arabidopsis* TFs that activation transcription. Although the tested hits were active in plant cells, it is possible that we failed to identify ADs in some TFs that would activate transcription only in *Arabidopsis*. However, a strength to our approach is that we are identifying generalizable features that are likely to directly recruit the transcription machinery. Performing this work directly in plant cells would prevent us from ruling out the possibility that we had identified a domain that recruited another *Arabidopsis* TF, which itself recruited the transcriptional machinery. Thus, this dataset allowed us to create TADA, which outperforms existing AD predictors on human and plant datasets (Extended Data Figs. 3b,c and 9d).

We found that the acidic aromatic class of ADs dominates our dataset, which might be expected from a yeast-based screen. However, around 15% of *Arabidopsis* ADs were neutral or positive and had few aromatic residues. Using our TADA network output, we identified six distinct AD subtypes with differing feature properties. We speculate that each subtype might recruit distinct transcriptional machinery or function only in certain cellular environments, similar to our observations when interrogating the two distinct ADs found in ARF7 (Fig. 4d,e).

Moreover, these identified AD subtypes showed differential enrichment across TF families (Extended Data Fig. 6), indicating potential subfunctionalization. Investigating the contribution of these subtypes to transcriptional activation using biochemical and genetic assays will be crucial to understand their roles.

Whereas the sequence of ADs varies, their position within the examined TF families remains conserved (Fig. 5 and Extended Data Fig. 9a). This suggests functional conservation of domains in rapidly evolving regions of intrinsic disorder. Our findings imply that AD location contributes to TF function, potentially by providing additional means of regulating transcription. Nearby interaction domains could occlude or reinforce recruitment of the transcriptional machinery depending on the context. For example, the PIF3 AD is physically blocked by a protein interaction occurring at an adjacent site, preventing transcription[27]. Similarly, an interaction complex facilitated by the ARF19 PB1 domain and proximal to AD2 regulates Mediator assembly[37]. We expect that our genome-wide annotation of ADs in *Arabidopsis* will lead to the discovery of similar examples.

Certain TFs exhibit bifunctionality, containing both ADs and RDs. Similar to many human TFs[36], several clade-B ARFs possess both ADs and putative RDs (Fig. 5c). Because RDs are strong transcriptional effectors and override ADs[34], we postulate that if the RD and AD are equally accessible, the TF will act as a transcriptional repressor. However, if the RD were to be buried or occluded, the TF could then act as an activator. The ancient and dual roles of ARFs as transcriptional activators and repressors represent an intriguing model for studying the relationship between RD and AD activity.

Our study, using empirical PADI data and the innovative TADA network, offers a powerful approach to identify and classify ADs. This development provides a much-needed tool to fully understand ADs and their role in transcriptional regulation. Moreover, our work goes beyond expanding our understanding of TF function. It introduces a model that investigates the intricate connection between sequence and function within IDRs. Our experimental PADI approach provides a framework to uncover patterns of conservation that are based on function and position rather than on sequence similarity. To further enhance our understanding, we have developed the TADA network, which uses feature space and amino acid properties to capture both linear and non-linear relationships among features. This innovative approach, coupled with downstream analyses, provides a roadmap for investigating sequence features that contribute to IDR function. We believe that this comprehensive methodology will have a far-reaching effect on the characterization of IDRs and will enable considerable advances to be made in this field.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07707-3.

1. Strader, L., Weijers, D. & Wagner, D. Plant transcription factors—being in the right place with the right company. *Curr. Opin. Plant Biol.* **65**, 102136 (2022).
2. O'Malley, R. C. et al. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**, 1280–1292 (2016).
3. Galli, M. et al. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.* **9**, 4526 (2018).
4. Sanborn, A. L. et al. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* **10**, e68068 (2021).
5. Dyson, H. J. & Wright, P. E. Role of Intrinsic protein disorder in the function and interactions of the transcriptional coactivators CREB-binding protein (CBP) and p300. *J. Biol. Chem.* **291**, 6714–6722 (2016).
6. Ferreira, M. E. et al. Mechanism of transcription factor recruitment by acidic activators. *J. Biol. Chem.* **280**, 21779–21784 (2005).
7. Hermann, S., Berndt, K. D. & Wright, A. P. How transcriptional activators bind target proteins. *J. Biol. Chem.* **276**, 40127–40132 (2001).
8. Kim, J. Y. & Chung, H. S. Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science* **368**, 1253–1257 (2020).
9. Staller, M. V. et al. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst.* **13**, 334–345 (2022).
10. Kotha, S. R. & Staller, M. V. Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence. *Genetics* **225**, iyad131 (2023).
11. Hummel, N. F. C. et al. The *trans*-regulatory landscape of gene networks in plants. *Cell Syst.* **14**, 501–511 (2023).
12. Staller, M. V. et al. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* **6**, 444–455 (2018).
13. Konishi, M. & Yanagisawa, S. The role of protein–protein interactions mediated by the PB1 domain of NLP transcription factors in nitrate-inducible gene expression. *BMC Plant Biol.* **19**, 90 (2019).
14. Hahn, S. & Young, E. T. Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**, 705–736 (2011).
15. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120**, 4312–4319 (2021).
16. Hope, I. A., Mahadevan, S. & Struhl, K. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature* **333**, 635–640 (1988).
17. Hope, I. A. & Struhl, K. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell* **46**, 885–894 (1986).
18. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
19. Mahatma, S. *et al*. Prediction and functional characterization of transcriptional activation domains. In *57th Annual Conference on Information Sciences and Systems (CISS)* 1–6 (2023).
20. Erijman, A. et al. A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Mol. Cell* **78**, 890–902 (2020).
21. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems* 4768–4777 (2017).
22. Hussain, R. M. F., Sheikh, A. H., Haider, I., Quareshy, M. & Linthorst, H. J. M. *Arabidopsis* WRKY50 and TGA transcription factors synergistically activate expression of PR1. *Front. Plant Sci.* **9**, 930 (2018).
23. Li, J. et al. Activation domains for controlling plant gene expression using designed transcription factors. *Plant Biotechnol. J.* **11**, 671–680 (2013).
24. Cho, S. et al. Analysis of the C-terminal region of *Arabidopsis thaliana* APETALA1 as a transcription activation domain. *Plant Mol. Biol.* **40**, 419–429 (1999).
25. Sakuma, Y. et al. Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. *Plant Cell* **18**, 1292–1309 (2006).
26. Kotak, S., Port, M., Ganguli, A., Bicker, F. & von Koskull-Doring, P. Characterization of C-terminal domains of *Arabidopsis* heat stress transcription factors (Hsfs) and identification of a new signature combination of plant class A Hsfs with AHA and NES motifs essential for activator function and intracellular localization. *Plant J.* **39**, 98–112 (2004).
27. Yoo, C. Y. et al. Direct photoresponsive inhibition of a p53-like transcription activation domain in PIF3 by *Arabidopsis* phytochrome B. *Nat. Commun.* **12**, 5614 (2021).
28. Fernandez-Calvo, P. et al. The *Arabidopsis* bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *Plant Cell* **23**, 701–715 (2011).
29. Tiwari, S. B., Hagen, G. & Guilfoyle, T. The roles of auxin response factor domains in auxin-responsive transcription. *Plant Cell* **15**, 533–543 (2003).
30. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Activation and repression of transcription by auxin-response factors. *Proc. Natl Acad. Sci. USA* **96**, 5844–5849 (1999).
31. Pierre-Jerome, E., Jang, S. S., Havens, K. A., Nemhauser, J. L. & Klavins, E. Recapitulation of the forward nuclear auxin response pathway in yeast. *Proc. Natl Acad. Sci. USA* **111**, 9407–2412 (2014).
32. Powers, S. K. & Strader, L. C. Regulation of auxin transcriptional responses. *Dev. Dyn.* **249**, 483–495 (2020).
33. Choi, H. S., Seo, M. & Cho, H. T. Two TPL-binding motifs of ARF2 are involved in repression of auxin responses. *Front. Plant Sci.* **9**, 372 (2018).
34. Hiratsu, K., Matsui, K., Koyama, T. & Ohme-Takagi, M. Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *Plant J.* **34**, 733–739 (2003).
35. Mutte, S. K. et al. Origin and evolution of the nuclear auxin response system. *eLife* **7**, e33399 (2018).
36. DelRosso, N. et al. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* **616**, 365–372 (2023).
37. Leydon, A. R. et al. Repression by the *Arabidopsis* TOPLESS corepressor requires association with the core mediator complex. *eLife* **10**, e66739 (2021).
38. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. & Pappu, R. V. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **112**, 16–21 (2017).
39. Kagale, S. & Rozwadowski, K. EAR motif-mediated transcriptional repression in plants: an underlying mechanism for epigenetic regulation of gene expression. *Epigenetics* **6**, 141–146 (2011).

Methods

Library generation

PADI libraries. Protein sequences for primary gene models were downloaded from TAIR (https://www.arabidopsis.org/) using the bulk data retrieval tool. The Araport11 assembly was accessed in June 2020 to download sequences. Proteins were fragmented into 40-amino-acid tiles with a step size of 10 amino acids using a custom Python script. An additional 40-amino-acid tile that corresponds to the final 40 amino acids of each protein was also generated to ensure full coverage of each TF. Each tile was given a unique name corresponding to its AGI locus identified and the starting amino acid position of each tile. Each tile was then reverse-translated into a yeast-codon-optimized DNA sequence. Cloning adapters were added to each sequence at the 5′ and 3′ end as described previously[12] with a minor modification, no barcodes were included in the 3′ adapters. This resulted in 183-bp sequences, 120 of which encode the variable 40-amino-acid sequence. These sequences were distributed into 10 synthesis libraries of around 7,000 fragments each for a total synthesis of 69,347 *Arabidopsis* tiles (Agilent). No TFs were split between synthesis libraries and synthesis libraries were ordered on the basis of AGI locus number to be functionally random in content.

ARF evolution library. ARF sequences for 12 species were identified using BlastP searches on Phytozome (https://phytozome-next.jgi.doe.gov) against the ARF5 protein sequence, with the exception of ARFs from *Zea mays*, which were accessed from MaizeGDB (https://www.maizegdb.org/). All top hits from each species were aligned to AtARF2, AtARF5, AtARF7 and AtARF17 protein sequences to determine the presence of a canonical ARF DBD and PB1 domain. Blast hits that lacked one or both were excluded. The middle region was defined as the first amino acid downstream of the DBD, on the basis of a previous study[40], to the last amino acid upstream of the conserved PB1, on the basis of another study[41]. The extracted middle regions were then passed through our custom Python script to generate tiles as described above, resulting in a synthesis library of 9,069 fragments.

Pilot library. A pilot library consisting of *Arabidopsis thaliana* and *Zea mays* ARF middle region tiles was generated as described above. In total 2,260 fragments were synthesized; these data was used in the training of the TADA network.

Cloning of synthesis libraries into pMVS142

All libraries were cloned into the pMVS142 backbone[12]. The plasmid contains a KanMX gene and the yeast *ACT1* promoter driving the expression of a synthetic TF comprising an N-terminal mCherry tag fused to the mouse Zif268 DBD followed by an oestrogen-binding domain and a multiple cloning site for fragment integration. Synthesis libraries were amplified using primers specific to the shared 5′ and 3′ adapters and Q5 2X Mastermix (NEB) and purified with the Monarch PCR Purification Kit (NEB). The plasmid backbone was digested with NheI-HF and AscI and the synthesis libraries were cloned into the digested backbone using the NEBuilder HiFi assembly at a 2:1 insert to vector ratio eight times to ensure the integration of all fragments. The resulting reactions were pooled and cleaned using the Monarch PCR Purification Kit (NEB) and transformed into 100 µl of Top10 electrocompetent *Escherichia coli*. Transformed cells were grown overnight in 125 ml Luria broth (LB) with ampicillin (Amp) selection. Dilution series up to 1:10,000 of transformed cells were plated on LB + Amp plates to determine colony counts in liquid culture. Synthesis libraries were considered successfully cloned if we reached colony counts higher than 49,000. Plasmid DNA was extracted from transformed *E. coli* using the ZymoPURE II Plasmid Maxiprep Kit (Zymo Research).

Yeast strains used in PADI assays

DHY211 is a *MAT*a yeast strain from A. Chu and J. Horecka and was used to generate yeast pools carrying synthetic TFs at the *URA3* locus. MY435 is the MATα reporter strain that contains a fast-maturing GFP variant driven by six Zif268 binding sites.

Yeast integration, selection and mating

PADI, ARF evolution and pilot libraries. Yeast synthetic TF cloning was performed as described previously[12] with the following modifications: maxipreps of synthetic TF plasmid libraries were triple digested with EcoRI-HF, PacI-HF and SalI-HF before transformation, along with 500-bp homology arms for the 5′ and 3′ ends of the synthetic TF, into DHY211. Transformed cells were recovered overnight in YPD medium and then plated on SC + G418 + 5-FOA plates to identify clones with stable integrations of the synthetic TF pool at the *URA3* locus. Libraries were deemed successfully cloned when 49,000 colonies were reached (77,000 colonies for the ARF evolution library). The resulting clones were then scraped, pooled, washed and grown again in YPD for mating or stored in glycerol at −80 °C for additional mating and experimentation. Positive yeast clones were mated to MATα MY435 twice (or five times in the case of the ARF evolution library) and then pooled to ensure retention of all fragments. The resulting mated cells were grown overnight in bulk in SC + G418 + NAT. Cells were collected and stored concentrated in glycerol at −80 °C until flow sorting experiments. In total, 892 fragments (1.3%) were lost during cloning steps across the 10 PADI libraries, 60 fragments (0.7%) were lost during cloning from the ARF evolution library and four fragments (0.2%) were lost during cloning from the pilot library.

NLP7 fragment and PB1 confirmation. *Arabidopsis* NLP7 coding sequences that encode a 40-amino-acid fragment starting with Leu101 (fragment 1), a 40-amino-acid fragment starting with Glu901 (fragment 2), and the PB1 domain (Thr863 to Val945) were synthesized by Twist Biosciences. Fragments were PCR-amplified using primers with overlap to pMVS219 as described above to generate synthetic TFs for each tested region using NEBuilder HiFi cloning as described above. The resulting clones were confirmed by sequencing with Plasmidsaurus. Yeast strains expressing synthetic TFs with fragment 1, fragment 2 and PB1 domain yeast strains were made as described above using homology-directed integration of synthetic TFs with one modification: MY435 was used for stable integration instead of DHY211. Cells were recovered for 1 h in YPD and plated on SC + G418 + NAT to confirm transformants.

Confirmation of ARF7 ADs in a yeast synthetic auxin signalling system. ARF7 in the pGP8A vector was driven under the *ADH1* promoter[42]. AD deletions were made to this ARF7 plasmid using j5 to design primers[43] and an in vivo assembly strategy to assemble the constructs[44]. Sanger sequencing confirmed the coding sequence of AD deletion constructs.

Plasmids encoding the effectors ARF7, ARF7$^{\Delta AD1}$, ARF7$^{\Delta AD2}$ and ARF7$^{\Delta AD1\Delta AD2}$ were digested with PmeI before Lithium PEG transformation[45]. ARF7 constructs were transformed into YPH500 (MATα ura3-52 lys2-801_amber ade2-101_ochre trp1-Δ63 his1-Δ200 leu2-Δ1) containing an mScarlet reporter driven under the IAA19 promoter housed in a pGP6G vector. Correct integration of transformed colonies was confirmed by diagnostic PCR across the boundary of homologous recombination and confirmed transformants were struck to isolation on YPAD plates.

PADI assay

Mated yeast libraries were grown overnight in SC + G418 + NAT, unmated yeast libraries were grown overnight in SC + G418 and untransformed DHY211 were grown in SC overnight at 30 °C. Yeast grown overnight was subcultured at 1:5 dilution in SC medium without selection and 1 μM β-oestradiol was added to mated libraries and the positive and negative control strains. Cells were grown for an additional four hours at 30 °C and then placed on ice until cell sorted.

Cells were analysed and sorted on a Beckman Coulter Astrios, with the Summit Software package, at the DCI Flow Cytometry Core at Duke University (PADI and ARF evolution libraries) and on a BD Aria-II machine at Washington University in St. Louis (pilot assay). DHY211 and the positive and negative controls were used to define cell gates and gain in each experiment. Experimental yeast libraries were sorted into 12 bins on the basis of the relative GFP:mCherry signal in each cell. A minimum of 60,000 cells per bin were sorted for the PADI assay and 355,000 cells per bin for the ARF evolution experiment into 2 ml fresh SC. Cells were kept on ice until the completion of the experiment and then grown in SC medium at a final volume of 5 ml overnight at 30 °C. The cells were then pelleted and frozen at −80 °C until DNA extractions were conducted.

The GFP- and mCherry-only sorts were conducted as above with some modification. All ten mated PADI libraries were grown overnight in SC + G418 + NAT. The optical density at 600 nm ($OD_{600\,nm}$) of each library was taken and then all ten libraries were pooled with equal numbers of cells, and the pool was diluted 1:5 in SC medium with 1 μM β-oestradiol and grown for four hours before cell sorting. GFP- and mCherry-only sorts were performed with only six bins each that spanned the range of values present in the pooled libraries and one million cells per bin were sorted into fresh SC medium. Overnight growth and storage were performed as described above.

All PADI libraries were assayed once, with the exception of library 3, which was assayed twice. Library 3 replicates are compared in Extended Data Fig. 1d. The control fragments common to all ten libraries were assayed independently in each library and are compared in Extended Data Fig. 1e,f.

Yeast genomic extraction and sequencing

Yeast genomic DNA was extracted with the YeaStar Genomic DNA Kit (Zymo Research). Sequencing libraries were generated through three PCR reactions. PCR1 amplified a 600-bp fragment that contained the tested fragment from the integrated locus. PCR2 amplified the fragment itself and added phasing and Illumina sequencing adapters. PCR3 completed the Illumina sequencing adapters and indexes specific to each sample for each bin and yeast library tested. All yeast libraries and bins were sequenced with 150-bp PE reads, or 150-bp SE for the pilot study, with a minimum of one million reads per sample.

Data analysis

Paired raw Fastq files from each library were aligned to fragment DNA sequences using BWA-mem aligner (v.0.7.15)[46] and SAMtools (v.1.10)[47] to generate BAM files. Fragment counts in each bin were extracted from the resulting BAM files using SAMtools coverage[47]. Count files were then opened in Python using Pandas (v.1.4.1; https://zenodo.org/record/7979740) and NumPy[48] to generate PADI scores. Each library was independently analysed to generate PADI scores by first normalizing each sequenced bin by counts per million reads. Each bin was normalized by the number of yeast collected in each bin. Next, each fragment was normalized across bins by taking the fragment counts in each bin and dividing by the sum of fragment counts across all bins. The raw AD score was generated by taking the dot product of the proportional count of each fragment in each bin and the median GFP:mCherry score for each bin. Raw AD scores were then *z*-score normalized using the preprocessing command in the Scikit-learn package (v.1.2.0)[49] to generate the final PADI score.

Sequence features for each fragment were determined using LocalCIDER (v.0.1.19)[38], including hydrophobicity, kappa and individual amino acid counts. Net charge was calculated by taking the sum of Arg and Lys residues and subtracting the sum of Asp and Glu residues. Disorder predictions for all tested TFs were generated using Metapredict 2 (v.2.2)[15]. Mean disorder values were applied to each fragment by taking the mean Metapredict values assigned to each amino acid in the tested fragment.

NLP7 fragment confirmation

Positive yeast transformants expressing the synthetic TF with fragment 1, fragment 2 or the intact PB1 were grown in SC + G418 + NAT overnight and diluted 1:5 in SC medium. These transformants were then induced with 1,000× β-oestradiol and allowed to incubate at 30 °C for 4 h. Transformed cell populations were scored using the Beckman Coulter Cytoflex S Flow Cytometer and CytExpert software. A general gating strategy was used to identify the population of present yeast. Cells expressing mCherry were identified by comparing untransformed MY435 cells using the Y610 channel (ex: 561 nm, em: 610 ± 20 nm, 2,000 gain). A minimum of 300,000 mCherry-positive yeast cells from four independent induction experiments for each construct (NLP7 AD, PB1 fragment and PB1 domain) were used to collect GFP-reporter levels using the B525 channel (ex: 488 nm, em: 525 ± 40 nm, 2,000 gain). FCS files were generated through Cytoflex and the mean GFP/mCherry score was calculated using the Python packages flowkit (v.1.0.1), seaborn (v.0.13.0) and pandas (v.1.4.1).

Training dataset and encoding

The TADA neural network, which has been previously described[19], underwent training using 75,845 random peptides. To capture the inherent characteristics of plant ADs, TADA was retrained on PADI. PADI consists of a total of 70,937 40-amino-acid fragments, among which 64,552 and 6,385 were identified as non-ADs and ADs, respectively (PADI data). To represent the sequences and capture the side-chain properties of each fragment, 42 features were computed. These features included 11 side-chain properties (Supplementary Table 4) and 9 properties used to describe disordered regions (Supplementary Table 4), and the count of each of the 20 amino acids was computed for a window of size 5 across the entire sequence length with a step size of one amino acid. The intrinsically disordered properties were calculated using LocalCIDER (v.0.1.19)[38] and AlphaFold[50]. This sliding window approach resulted in 36 subsequences. In addition, two intrinsically disordered properties, kappa and omega, were computed for the entire sequence length. To accommodate the 36 subsequences, the computed kappa and omega values were duplicated. The computation of these 42 features, accounting for the 36 × 42 input matrix, was performed. The dataset was then split in a stratified manner into three proportions: 70% for training, 20% for validation and 10% as a test set. After the split, the feature matrices were scaled using a standard scaler, which adjusted the mean and standard deviation of each feature to zero and one, respectively. This was followed by a min–max scaling, which rescaled the features between 0 and 1.

Neural network architecture

The TADA neural network architecture comprises four types of layers: (1) two convolutional neural network layers (CNN); (2) an attention layer; (3) two bi-directional long short term memory (biLSTM) layers; and (4) a dense layer. The purpose of the two CNN layers is to extract potential patterns within the fragments and reduce the dimensionality of the data. These one-dimensional (1D) CNNs perform convolutions using a kernel size of 2 and a stride of 1, allowing for the identification of potential bipeptides that are believed to be characteristic of ADs. To prevent overfitting and enhance generalization, dropout was incorporated into the CNN layers. An attention layer was included to highlight the learned patterns from the CNN layers and to selectively focus on the features that are more crucial for the prediction task. To capture the interdependence of the subsequences within a sequence, the biLSTM layers were added. Finally, the dense layer is connected to the output layer, completing the network architecture.

Experimental settings and evaluation

The hyperparameters found to give the best performance[19] are presented in Supplementary Table 4.

Our input dataset is unbalanced, with the ADs being underrepresented. To achieve accurate prediction for the underrepresented ADs, we used cost-sensitive approaches. Specifically, our cost-sensitive approach is twofold: (i) the misclassification of the minority class penalizes with a focal loss function; and (ii) our class weights are inversely proportional to the class sizes in the dataset.

To assess the performance of TADA trained using PADI, three neural networks were trained with (i) PADI alone; (ii) the random peptides data as described previously[19]; and (iii) a compiled dataset of PADI and the random peptides. In the combined dataset, the random peptides were extended by adding the first 10 amino acids of the cloning vector to generate 40-amino acid fragments. To account for class imbalance, cost- sensitive approaches were applied in all three neural networks, which were not yet implemented in the previous study[19]. We implemented early stopping criteria during model training based on the F1 score of the validation set. This allowed us to halt the training process when the model performance on the training set no longer improved significantly. Finally, to confirm that TADA is not memorizing sequences and thus biased towards predicting ARF sequences, we fully retrained TADA withholding a total of 2,046 ARF sequences (ARF (494), MPARF (87), PPARF (469), AMARF (996)). To evaluate the performance of the neural networks, we calculated various performance metrics on the test dataset. These metrics include precision, recall, AUPR, area under the receiver operating curve (AUC), accuracy and F1 score. These metrics were computed individually for each class.

Analysis of feature importance and unsupervised clustering

To conduct predictions and SHAP analysis, we retrained the neural network using a 90:10 split between the training and the validation datasets. The best model obtained during training was saved and used for predictions and SHAP analysis. We used SHAP[21] to assess the influence of the 42 computed features on the predictions. Within the SHAP package, we used the GradientExplainer to acquire the SHAP value associated with each feature and each subsequence for all AD positive classes in our dataset. To determine the overall effect of the 42 features, we aggregated the SHAP values for each feature by summing the absolute values of each subsequence and fragment. To select the most important features in an unbiased manner, we normalized the obtained SHAP values using *z*-scores and selected all features with a *z*-score higher than 1. In total, we identified eight features, which were subsequently used for clustering. To determine the directionality of the effect of the top eight features and ascertain whether they had an overall positive or negative influence on the predictions, we summed the SHAP values of each subsequence and fragment. Lastly, to gain an overall understanding of the variability of the important features across the AD fragments, we summed the absolute SHAP values of the 36 subsequences.

To identify subtypes of AD classes, we used an unsupervised approach that involved PCA, t-SNE and *k*-means clustering. First, we extracted the scaled features and SHAP values for the top eight most important features from the ADs. Second, we reduced the dimensions of the 2D feature matrix for each fragment by using kernel PCA, resulting in a 1D matrix that captured the majority of the variance within the subsequences. The 2D SHAP value dimensions for each fragment were reduced by summing the values across the subsequences. Next, we concatenated the features matrix (8 × 6,385) and the SHAP value matrix (8 × 6,385), and performed a kernel PCA with 10 components. The resulting components from the t-SNE were plotted, using PCA initialization, a high learning rate, large perplexity and exaggeration[51]. Finally, we used the output components of the t-SNE for *k*-means clustering, identifying six as the optimal number of clusters on the basis of an elbow plot. For each of the six clusters, we performed another SHAP analysis to assess any global differences in the contribution of each feature to the AD subtype prediction. Using the same approach as that used in the global SHAP analysis, we ranked each feature according to its importance.

Cloning synthetic TFs for protoplast assays and FrARF7

A gene fragment that encoded the Gal4 DBD fused to the middle region and C terminus of ARF7, a nos terminator and 1,000 bp of non-coding DNA including a multiple cloning site was synthesized by Twist in the pENTR gateway-compatible backbone. An additional gene fragment encoding 500 bp of non-coding DNA followed by 5× Gal4 UAS sites, a minimal CaMV 35S promoter and mScarlet-I fused to histone 2B was synthesized and put into the pENTR backbone by Twist. The pENTR backbones containing the synthetic TFs were linearized using SacI, and the reporter was PCR-amplified to include 20 bp overlap with the pENTR synthetic TF multiple cloning site at the 5′ and 3′ ends. The reporter insert was cloned into the linearized pENTR backbone vector using NEBuilder HiFi cloning to generate pENTR synthetic TF+Reporter clones. The entry clones were then cloned into pLCS107, which provided an in-frame mNEON fused to the N terminus of the synthetic TF driven under the UBQ10 promoter and a nos terminator for the mScarlet-I H2B reporter, by gateway cloning. The Gal4 DBD, FrARF7ΔAD1 and FrARF7ΔAD2 variants were generated by in-frame deletions of the ARF7 CDS, ARF7AD1 and ARF7AD2, respectively, by PCR linearization and self-assembly with NEBuilder HiFi cloning. The FrARF7ΔAD1ΔAD2 variant was generated by Hifi cloning to delete ARF7AD2 from the FrARF7ΔAD1 variant. These additional variants were then subcloned to add the reporter and then cloned into the pLCS107 backbone as described above.

A plasmid that encoded the Gal4 DBD, a nos terminator, 1,500 bp of non-coding DNA followed by 5× Gal4 UAS sites, a minimal CaMV 35S promoter, mNeonGreen and a nos terminator was synthesized by Twist in the pENTR gateway-compatible backbone. The pENTR backbone containing the synthetic TFs were linearized using SacI, and the AD fragments were PCR-amplified to include 20 bp overlap with the pENTR synthetic TF multiple cloning site at the 5′ and 3′ ends. The reporter insert was cloned into the linearized pENTR backbone vector using NEBuilder HiFi cloning to generate pENTR synthetic TF+Reporter clones. The entry clones were then cloned into pLCS99, which provided the UBQ10 promoter, by gateway cloning.

Auxin-responsive reporter activation assays in yeast

Activation assays were adapted from a previous report[52] using an Attune NxT Acoustic focusing cytometer with 488-nm excitation, forward-scatter and side-scatter and 637-nm emission for RFP. Events were annotated and plotted using the flowTime R package[53]. Individual colonies of each strain were diluted to 1 cell per µl in synthetic complete medium (Takara). Cultures were incubated overnight for 16 h at 900 rpm in a Talboys microplate shaker. The following morning, three separate measurements were drawn at approximately one hour apart for measurement. Cultures were in exponential growth phase to capture maximum activation. Approximately 10,000 events from biological replicates were recorded 3 times (9 total) for each measurement and the YL1.A channel was used.

Testing synthetic TFs and FrARF7

Arabidopsis mesophyll protoplasts were isolated from 14-day-old Col-0 leaves. A total of 100,000 cells were transformed with 20–30 µg of plasmid DNA carrying Gal4, and synthetic TFs or FrARF7, FrARF7ΔAD1, FrARF7ΔAD2 or FrARF7ΔAD1ΔAD2 with UAS constructs using the tape-sandwich method and incubated for 16 h in the dark. Transformed cell populations were scored using the Beckman Coulter Cytoflex S Flow Cytometer and CytExpert software. A back gating strategy was taken to identify the population of intact protoplasts. For the FrARF7 experiment, cells expressing mNEON reporters were first identified by comparing transformed mNEON-Gal4 cells with untransformed cells using the B525 channel (ex: 488 nm, em: 525 ± 40 nm, 69 gain) and then back gated on FSCvSSC. A minimum of 63 mNEON-positive cells from four independent transformations were used to collect mScarlet-I H2B reporter levels using the Y610 channel (ex: 561 nm, em: 610 ± 20 nm, 1,000 gain). At least 2,212 total mNEON-positive cells from four independent transformations were used to determine the mean levels of mScarlet-I H2B. For the synthetic TF experiment, cells expressing mScarlet-I were first identified by comparing transformed mScarlet-Gal4 cells with untransformed cells using the Y610 channel (ex: 561 nm, em: 610 ± 20 nm, 1,000 gain) and mNeonGreen reporter levels were collected using the B525 channel (ex: 488 nm, em: 525 ± 40 nm, 69 gain). FCS files for mNEON-positive cell populations were generated by Cytoflex and analysed using FlowKit (v.1.0.1)[54], NumPy (v.1.22.3)[48] and Pandas (v.1.4.1; https://zenodo.org/record/7979740) packages in Python (v.3.8.12) with custom scripts. A minimum of 520 cells from three independent transfections were used to determine the mean mNEON values. When used, Tukey HSD statistical tests to determine alpha-groups between populations were conducted in JMP Pro 17 (v.17.0,0) at an alpha-level of .01.

Determining subtypes of the ARF evolution dataset

To determine the subtypes of ADs among the six identified subtypes, we first conducted a SHAP analysis. We then generated a concatenated dataset of the scaled features and SHAP values for the top eight most important features. To incorporate these new points into our existing t-SNE plot, we used a previously described method[51]. Following that, we correlated the 16 data points (8 features and 8 SHAP values) of each AD fragment to the same data of the TADA class 1 ADs. Next, we selected the ten nearest neighbours that have the highest correlation. We then took the median of the t-SNE positions of these ten nearest neighbours to identify the plot position of the AD fragment. To assign the subtype for the ADs, we took the most frequent subtype of the ten nearest neighbours.

Sequence analysis and predicted sequence properties

Sequence analysis calculations for sequence charge decoration (SCD) and sequence hydropathy decoration (SHD) were generated using the Python package SPARROW (https://github.com/idptools/sparrow). Predictions for radius of gyration (Rg), end-to-end distance (Re) and asphericity were generated using the Python package ALBATROSS[55]. In brief, ALBATROSS is a Python package that contains bidirectional recurrent neural networks trained to generate predicted sequence properties for disordered proteins including Rg, Re and asphericity using the primary amino acid sequence as the input.

Sample size

No calculations were performed to predetermine sample size. Biological and technical replicates were performed as described in the Methods for each experiment and conform to standards in the field. Exact *n* numbers for each experiment are provided in each figure legend.

Data exclusions and replication

No data were excluded. Experiments were replicated as described in the Methods. All attempts at replication were successful.

Randomization and blinding

Randomization and blinding were not applicable because the data are quantitative and were not subjectively grouped.

Unique biological materials

Unique biological materials are available from the corresponding author (L.C.S.) upon request.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Library sequencing data have been deposited in the NCBI's Gene Expression Omnibus (GEO) and are accessible through the GEO series accession number GSE234215. Source data are provided with this paper.

Code availability

All scripts for the neural network training and validation and for making predictions are available on GitHub (https://github.com/LisaVdB/TADA).

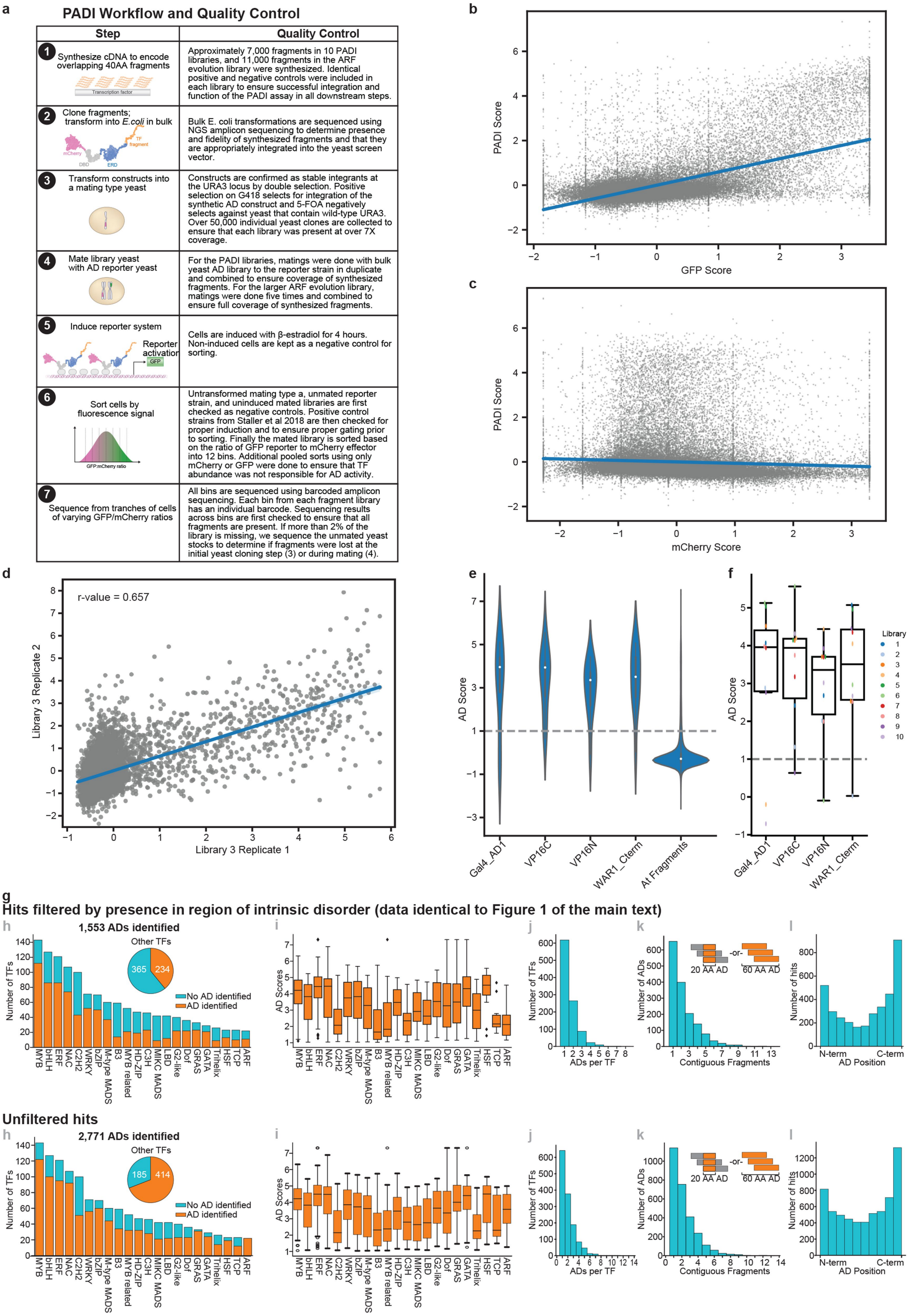
40. Boer, D. R. et al. Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* **156**, 577–589 (2014).
41. Korasick, D. A. et al. Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proc. Natl Acad. Sci. USA* **111**, 5427–5432 (2014).
42. Havens, K. A. et al. A synthetic approach reveals extensive tunability of auxin signaling. *Plant Physiol.* **160**, 135–142 (2012).
43. Hillson, N. J., Rosengarten, R. D. & Keasling, J. D. j5 DNA assembly design automation software. *ACS Synth. Biol.* **1**, 14–21 (2012).
44. Garcia-Nafria, J., Watson, J. F. & Greger, I. H. IVA cloning: a single-tube universal cloning system exploiting bacterial in vivo assembly. *Sci. Rep.* **6**, 27459 (2016).
45. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
47. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
48. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
49. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
51. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
52. Pierre-Jerome, E., Wright, R. C. & Nemhauser, J. L. Characterizing auxin response circuits in *Saccharomyces cerevisiae* by flow cytometry. *Methods Mol. Biol.* **1497**, 271–281 (2017).
53. Wright, R. C., Bolten, N. & Pierre-Jerome, E. flowTime: annotation and analysis of biological dynamical systems using flow cytometry. R version 1.24.0 https://www.bioconductor.org/packages/release/bioc/html/flowTime.html (2023).
54. White, S. et al. FlowKit: a Python toolkit for integrated manual and automated cytometry analysis workflows. *Front. Immunol.* **12**, 768541 (2021).
55. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024).

Acknowledgements This work was supported by the National Science Foundation (PGRP BIO-2112056 to L.C.S., PGRP BIO-2112057 to M.V.S. and PGRP BIO-2112058 to R.S.), the NSF Postdoctoral Research Program (IOS-1907098 to N.M.) and the National Institutes of Health (R35 GM136338 to L.C.S.).

Author contributions N.M., M.V.S., R.S. and L.C.S. designed the study. N.M. and M.V.S. designed the pilot tiling libraries. N.M. designed the PADI tiling and ARF evolution tiling libraries. C.M. and N.M. cloned and integrated libraries into yeast. L.V.d.B., S.M., V.P., A.W. and R.S. designed and implemented the TADA network. R.J.E. and A.S.H. performed biophysical simulations and advised on Metapredict. J.A.B. and R.C.W. assessed ARF7 AD activity in the yeast synthetic auxin signalling system. N.M., T.M.L., K.S.-F., E.G.W., S.P. and L.C.S. tested ADs in protoplasts. S.R.K. and A.L. examined human TFs with TADA. N.M. and L.V.d.B. wrote the manuscript, with important contributions from R.S. and L.C.S., and contributions from all other authors. L.C.S. supervised the project, with contributions from R.S. and M.V.S.

Competing interests L.C.S. is on the science advisory board of Prose Foods. R.S. is founder of Raleigh Biosciences. The remaining authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07707-3.
Correspondence and requests for materials should be addressed to Lucia C. Strader.
Peer review information *Nature* thanks Jennifer Brophy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

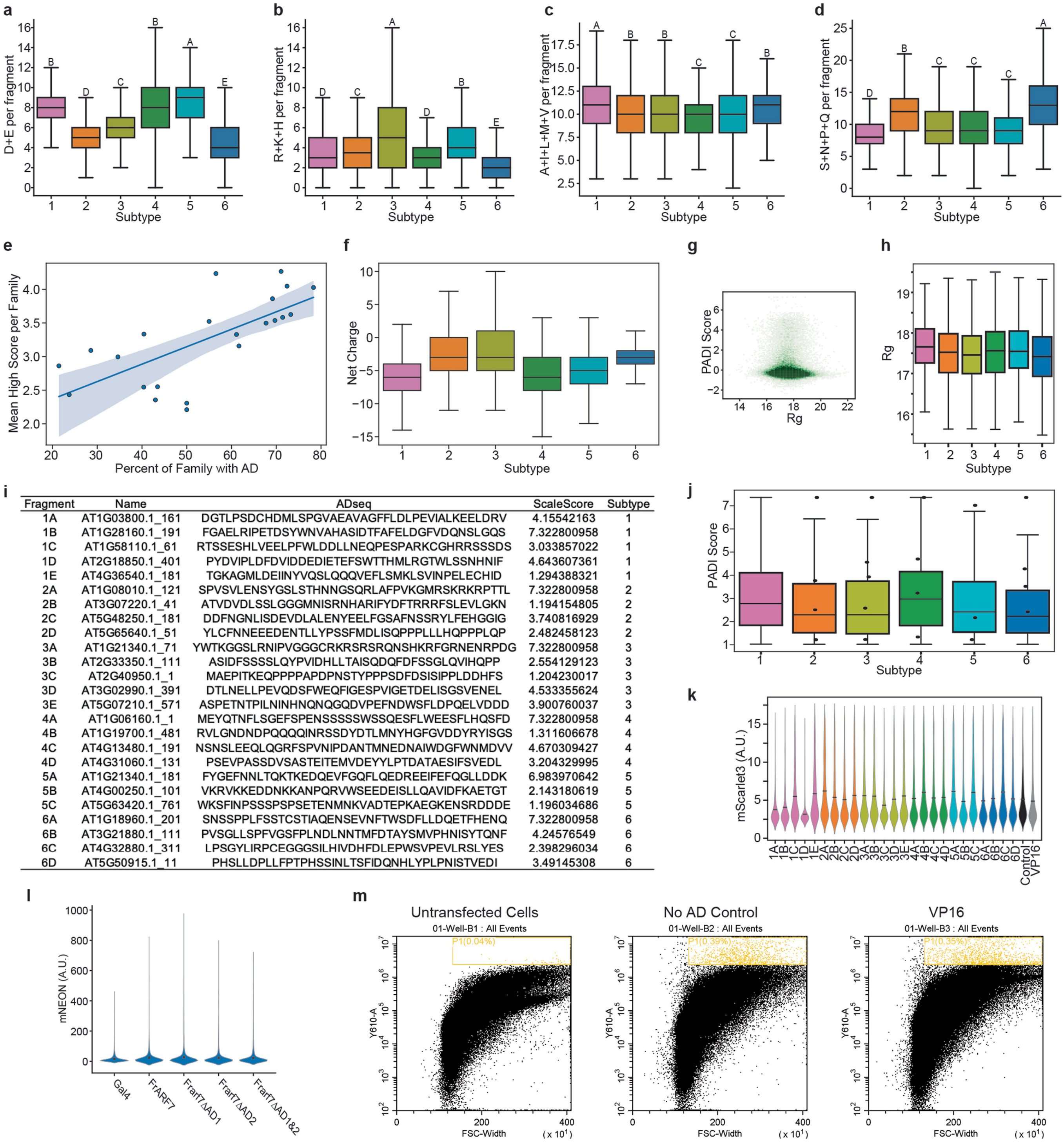


PADI Workflow and Quality Control

Step	Quality Control
1 Synthesize cDNA to encode overlapping 40AA fragments	Approximately 7,000 fragments in 10 PADI libraries, and 11,000 fragments in the ARF evolution library were synthesized. Identical positive and negative controls were included in each library to ensure successful integration and function of the PADI assay in all downstream steps.
2 Clone fragments; transform into *E.coli* in bulk	Bulk E. coli transformations are sequenced using NGS amplicon sequencing to determine presence and fidelity of synthesized fragments and that they are appropriately integrated into the yeast screen vector.
3 Transform constructs into a mating type yeast	Constructs are confirmed as stable integrants at the URA3 locus by double selection. Positive selection on G418 selects for integration of the synthetic AD construct and 5-FOA negatively selects against yeast that contain wild-type URA3. Over 50,000 individual yeast clones are collected to ensure that each library was present at over 7X coverage.
4 Mate library yeast with AD reporter yeast	For the PADI libraries, matings were done with bulk yeast AD library to the reporter strain in duplicate and combined to ensure coverage of synthesized fragments. For the larger ARF evolution library, matings were done five times and combined to ensure full coverage of synthesized fragments.
5 Induce reporter system	Cells are induced with β-estradiol for 4 hours. Non-induced cells are kept as a negative control for sorting.
6 Sort cells by fluorescence signal	Untransformed mating type a, unmated reporter strain, and uninduced mated libraries are first checked as negative controls. Positive control strains from Staller et al 2018 are then checked for proper induction and to ensure proper gating prior to sorting. Finally the mated library is sorted based on the ratio of GFP reporter to mCherry effector into 12 bins. Additional pooled sorts using only mCherry or GFP were done to ensure that TF abundance was not responsible for AD activity.
7 Sequence from tranches of cells of varying GFP/mCherry ratios	All bins are sequenced using barcoded amplicon sequencing. Each bin from each fragment library has an individual barcode. Sequencing results across bins are first checked to ensure that all fragments are present. If more than 2% of the library is missing, we sequence the unmated yeast stocks to determine if fragments were lost at the initial yeast cloning step (3) or during mating (4).

Extended Data Fig. 1 | See next page for caption.

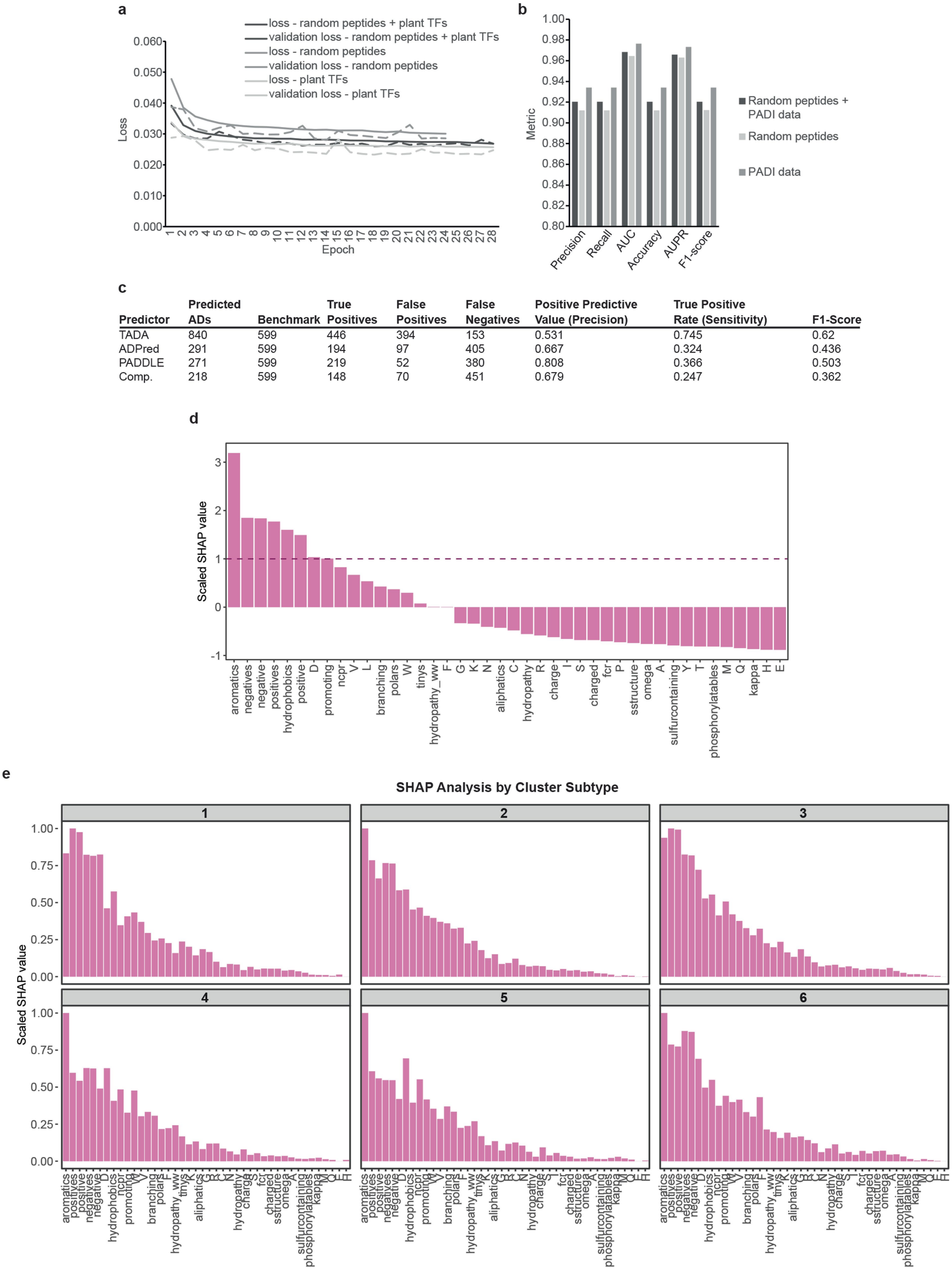
Extended Data Fig. 1 | PADI workflow and quality control. a, Extended depiction of the PADI assay. 1) DNA encoding 40-amino-acid fragments are synthesized and 2) cloned into a synthetic TF backbone in bulk. 3) Confirmed synthetic TF libraries are cloned into the URA3 locus of DHY211 yeast cells and positive clones are selected by G418 and 5-FOA resistance. 4) Positively cloned yeast TF libraries are mated to the MY435 reporter strain[12]. Positively mated clones are selected by G418 (library) and CloNAT (reporter) resistance. 5) Pooled mated libraries and controls are grown overnight and subcultured 1:5 with 1 µM beta-estradiol to induce synthetic TF localization to the nucleus. 6) After 4 hrs beta-estradiol treatment, mated yeast libraries are sorted into bins based on relative levels of GFP (reporter) to mCherry (synthetic TF) to determine AD activity. 7) Populations from each bin were grown overnight and sequenced to determine the distribution of tested fragments across bins. **b**,**c**, These plots show the correlation between PADI scores from all *Arabidopsis* TF libraries plotted against a pooled library where cells were sorted on median GFP (**b**) or mCherry (**c**) values. Each fragment was given a GFP or mCherry score based on the weighted mean of its appearance across all GFP or mCherry bins and then normalized using Z-score normalization consistent with how the PADI score was generated. The blue line represents the linear correlation of the data. There is a positive correlation between PADI score and GFP score, but not between PADI and mCherry scores. These results show that the PADI score is a robust measure of transcriptional activity regardless of the abundance of any TF. **d**, Scatter plot showing the correlation between two sorts of PADI library 3. Replicate 1 is included in all analysis. The blue line represents the linear regression of the two datasets. The linear regression model has an r-value of 0.657. **e**, Violin plots showing the PADI scores of four positive AD controls (n = 10 independent library experiments). The controls are found in all 10 PADI libraries and were consistently positive across libraries. The violin plot of *Arabidopsis* fragments (n = 69,347 fragments from 10 libraries) is also provided as a comparison. Box plots within the violin plot show the interquartile range and the median with whiskers that are 1.5 times the interquartile range. **f**, Box plots showing the PADI scores of tested control fragments across the 10 PADI libraries. Each point is the PADI score of the tested fragment and the colour of each point corresponds to the 10 PADI libraries (n = 10 independent experiments). All box plots show the interquartile range and the median. Whiskers are 1.5 times the interquartile range**. g**, Comparison of panels h–l from main text Fig. 1. The data presented from Fig. 1h–l (top) (n = 3,576) are presented above the same analysis conducted on all positive fragments regardless of mean disorder (bottom) (n = 6,207). The trends hold between the filtered data (top) and unfiltered data (bottom). **h**, Distribution of identified ADs across *Arabidopsis* TF families. **i**, Distribution of highest-scoring hits from each TF in each family. **j**, Distribution of the number of ADs identified per *Arabidopsis* TF. **k**, Distribution of number of contiguous hits identified per identified AD. Contiguous hits could be indicative of a short AD contained in neighbouring fragments or of an extended AD for which a subset of residues is sufficient to activate transcription; our data cannot distinguish between these. **l**, The distribution of hit locations revealed a bias towards the amino and carboxy termini of proteins. All box plots represent the median and interquartile range. The whiskers are 1.5 times the interquartile range.



Fragment	Name	ADseq	ScaleScore	Subtype
1A	AT1G03800.1_161	DGTLPSDCHDMLSPGVAEAVAGFFLDLPEVIALKEELDRV	4.15542163	1
1B	AT1G28160.1_191	FGAELRIPETDSYWNVAHASIDTFAFELDGFVDQNSLGQS	7.322800958	1
1C	AT1G58110.1_61	RTSSESHLVEELPFWLDDLLNEQPESPARKCGHRRSSSDS	3.033857022	1
1D	AT2G18850.1_401	PYDVIPLDFDVIDDEDIETEFSWTTHMLRGTWLSSNHNIF	4.643607361	1
1E	AT4G36540.1_181	TGKAGMLDEIINYVQSLQQQVEFLSMKLSVINPELECHID	1.294388321	1
2A	AT1G08010.1_121	SPVSVLENSYGSLSTHNNGSQRLAFPVKGMRSKRKRPTTL	7.322800958	2
2B	AT3G07220.1_41	ATVDVDLSSLGGGMNISRNHARIFYDFTRRRFSLEVLGKN	1.194154805	2
2C	AT5G48250.1_181	DDFNGNLISDEVDLALENYEELFGSAFNSSRYLFEHGGIG	3.740816929	2
2D	AT5G65640.1_51	YLCFNNEEEDENTLLYPSSFMDLISQPPPLLLHQPPPLQP	2.482458123	2
3A	AT1G21340.1_71	YWTKGGSLRNIPVGGGCRKRSRSRQNSHKRFGRNENRPDG	7.322800958	3
3B	AT2G33350.1_111	ASIDFSSSSLQYPVIDHLLTAISQDQFDFSSGLQVIHQPP	2.554129123	3
3C	AT2G40950.1_1	MAEPITKEQPPPPAPDPNSTYPPPSDFDSISIPPLDDHFS	1.204230017	3
3D	AT3G02990.1_391	DTLNELLPEVQDSFWEQFIGESPVIGETDELISGSVENEL	4.533355624	3
3E	AT5G07210.1_571	ASPETNTPILNINHNQNQGQDVPEFNDWSFLDPQELVDDD	3.900760037	3
4A	AT1G06160.1_1	MEYQTNFLSGEFSPENSSSSSWSSQESFLWEESFLHQSFD	7.322800958	4
4B	AT1G19700.1_481	RVLGNDNDPQQQQINRSSDYDTLMNYHGFGVDDYRYISGS	1.311606678	4
4C	AT4G13480.1_191	NSNSLEEQLQGRFSPVNIPDANTMNEDNAIWDGFWNMDVV	4.670309427	4
4D	AT4G31060.1_131	PSEVPASSDVSASTEITEMVDEYYLPTDATAESIFSVEDL	3.204329995	4
5A	AT1G21340.1_181	FYGEFNNLTQKTKEDQEVFGQFLQEDREEIFEFQGLLDDK	6.983970642	5
5B	AT4G00250.1_101	VKRVKKEDDNKKANPQRVWSEEDEISLLQAVIDFKAETGT	2.143180619	5
5C	AT5G63420.1_761	WKSFINPSSSPSPSETENMNKVADTEPKAEGKENSRDDDE	1.196034686	5
6A	AT1G18960.1_201	SNSSPPLFSSTCSTIAQENSEVNFTWSDFLLDQETFHENQ	7.322800958	6
6B	AT3G21880.1_111	PVSGLLSPFVGSFPLNDLNNTMFDTAYSMVPHNISYTQNF	4.24576549	6
6C	AT4G32880.1_311	LPSGYLIRPCEGGGSILHIVDHFDLEPWSVPEVLRSLYES	2.398296034	6
6D	AT5G50915.1_11	PHSLLDPLLFPTPHSSINLTSFIDQNHLYPLPNISTVEDI	3.49145308	6

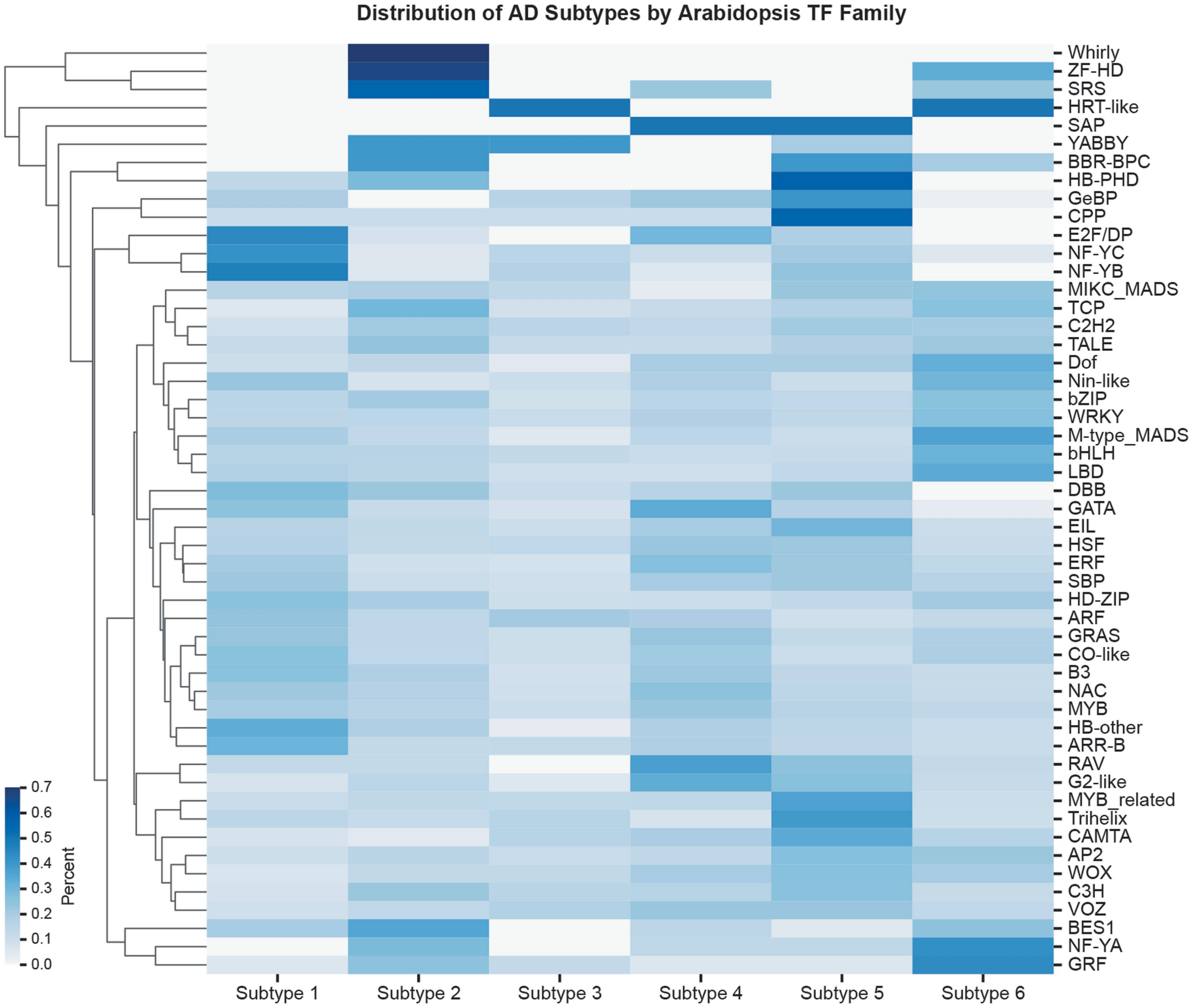
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | PADI hit characterization. a–**d**, Box plots showing the number of D + E (**a**) R + K + H (**b**) A + I + L + M + V (**c**) and S + N + P + Q (**d**) of each subtype (n ≥ 625). Letters correspond to the statistical levels of each subtype based on the Tukey–Kramer HSD metric with an alpha-level of 0.05. **e**, Scatter plot showing the correlation between the percentage of TFs with at least one AD (defined as a PADI score of greater than or equal to 1 and from an IDR) and the mean of the highest-scoring AD from each TF in a family. The line represents the linear regression and the shaded area represent the 95% confidence interval. **f**, Box plots showing the net charge of hits from each of the six AD subtypes (n ≥ 625). **g**, Heat map showing the distribution of Rg values against PADI score for all tested fragments (n = 6,207). We used simulations to examine the radius of gyration (Rg), which is a measure of the volume that an IDR ensemble occupies. Rg is particularly relevant to the AD molecular mechanism, as exposure of interacting side chains is necessary for interaction with the transcriptional machinery. We found that the Rg of our identified ADs occupied a narrow range of radii, as compared to the tested library, raising the possibility that ADs must adopt sufficiently expanded conformations for activity. **h**, Box plots showing the Rg values of each subtype; Rg was similar across subtypes (n ≥ 625). **i**, Table describing the PADI fragments tested in the synthetic TFs in Fig. 3h. The fragment key, its *Arabidopsis* identifier, amino acid sequence, PADI score, and subtype are shown. **j**, Box plots showing the distribution of PADI scores for each of the six subtypes. The stars represent the PADI score of the fragments tested for activity in Fig. 3h and shown in Extended Data Fig. 2i. The tested fragments span the range of PADI scores found in the six subtypes (n ≥ 625). Stars depict the PADI scores of selected hits for testing in protoplasts. **k**, Protein accumulation of Synthetic TFs from Fig. 3h. Violin plots show the mScarlet-TF values of cells. The black lines mark the mean mScarlet-TF value of each sample (n ≥ 529 cells from 3 independent transfections). **l**, Protein accumulation of FrankenARF TFs from Fig. 4e. Violin plots show the mNEON-TF values of cells. The black lines mark the mean mNEON-TF value of each sample (n ≥ 2,212 cells from 4 independent transfections). All cells collected for reporter expression were gated on the presence of TF signal when compared to blank cells. Only positive cells were used to collect output data presented in Figs. 3h and 4e. **m**, Gating strategy for examination of AD activity in protoplasts. Cells were gated based on size and mScarlet (for presence of TF) signal as depicted. Untransfected cells did not display signal above the threshold for mScarlet (left) whereas control cells transfected with the TF lacking an AD (middle) and cells transfected with the TF carrying VP16 (right) were selected for assessment of mNeonGreen (transcriptional output). All box plots represent the median and interquartile range. The whiskers are 1.5 times the interquartile range.

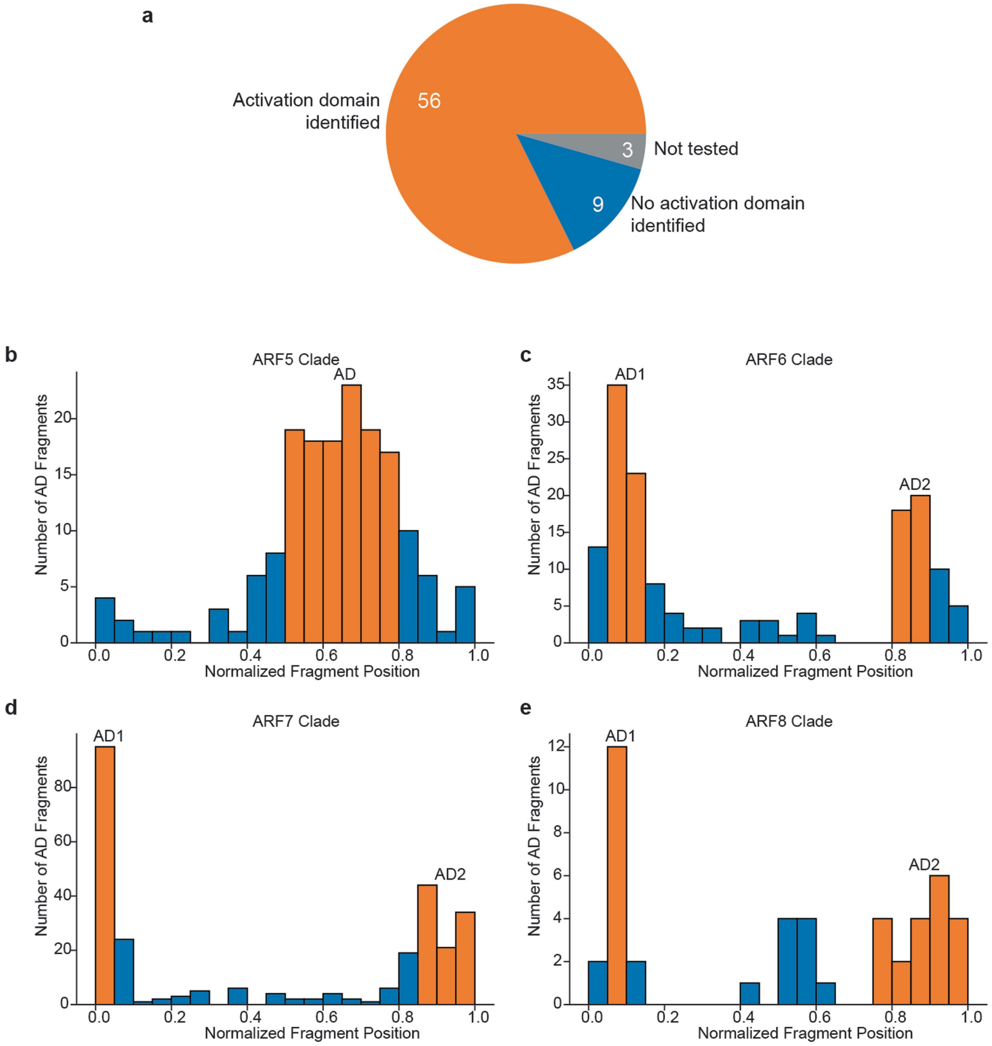


Predictor	Predicted ADs	Benchmark	True Positives	False Positives	False Negatives	Positive Predictive Value (Precision)	True Positive Rate (Sensitivity)	F1-Score
TADA	840	599	446	394	153	0.531	0.745	0.62
ADPred	291	599	194	97	405	0.667	0.324	0.436
PADDLE	271	599	219	52	380	0.808	0.366	0.503
Comp.	218	599	148	70	451	0.679	0.247	0.362

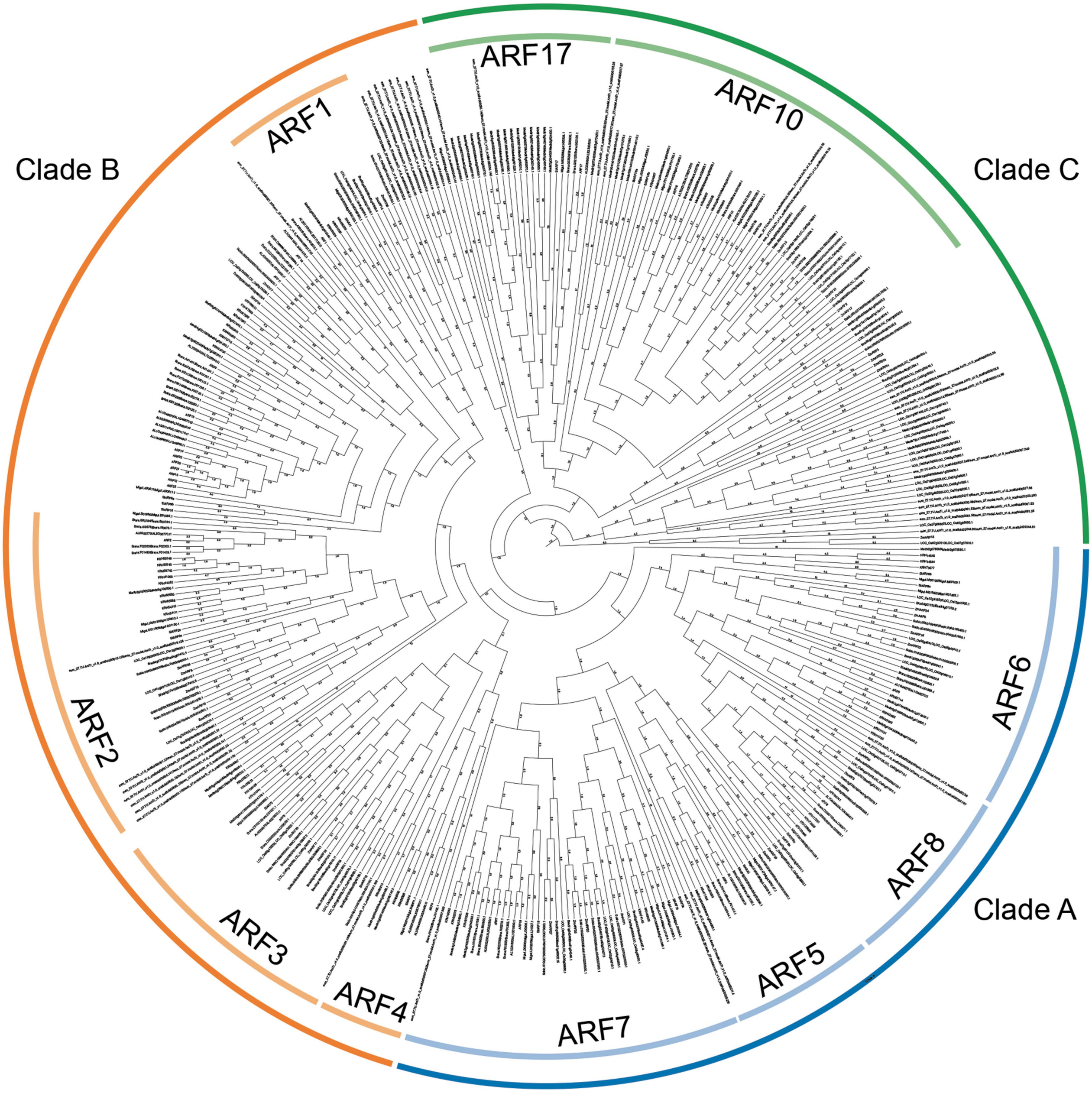
Extended Data Fig. 3 | Classification performance of TADA and effect of features on TADA's prediction performance. a, The loss of TADA during training and validation. **b**, TADA's performance in terms of precision, recall, area under the receiver operating curve (AUC), accuracy, AUPR and F1 score. TADA was trained three distinct times using random peptides[20], PADI (referred to as "plant TFs"), and random peptides and PADI combined. **c**, TADA outperforms all published AD predictors. We compared the performance TADA with three published AD predictors (ADpred, PADDLE and a composition model[4,10,20]. We used a hand-curated list of 599 ADs from 451 human TFs. For each TF, we predicted ADs and considered predictions that overlapped a known annotation by > 10 amino acids to be true positive, using each predictor. TADA made the most predictions, had the highest Sensitivity, and highest F1 score. **d**, Z-score normalized SHAP values leading to the selection of 8 features with a z-score above 1. **e**, Normalized SHAP values ranked from overall most important to least important for fragments scoring above 1 for each of the 6 identified AD subclasses.



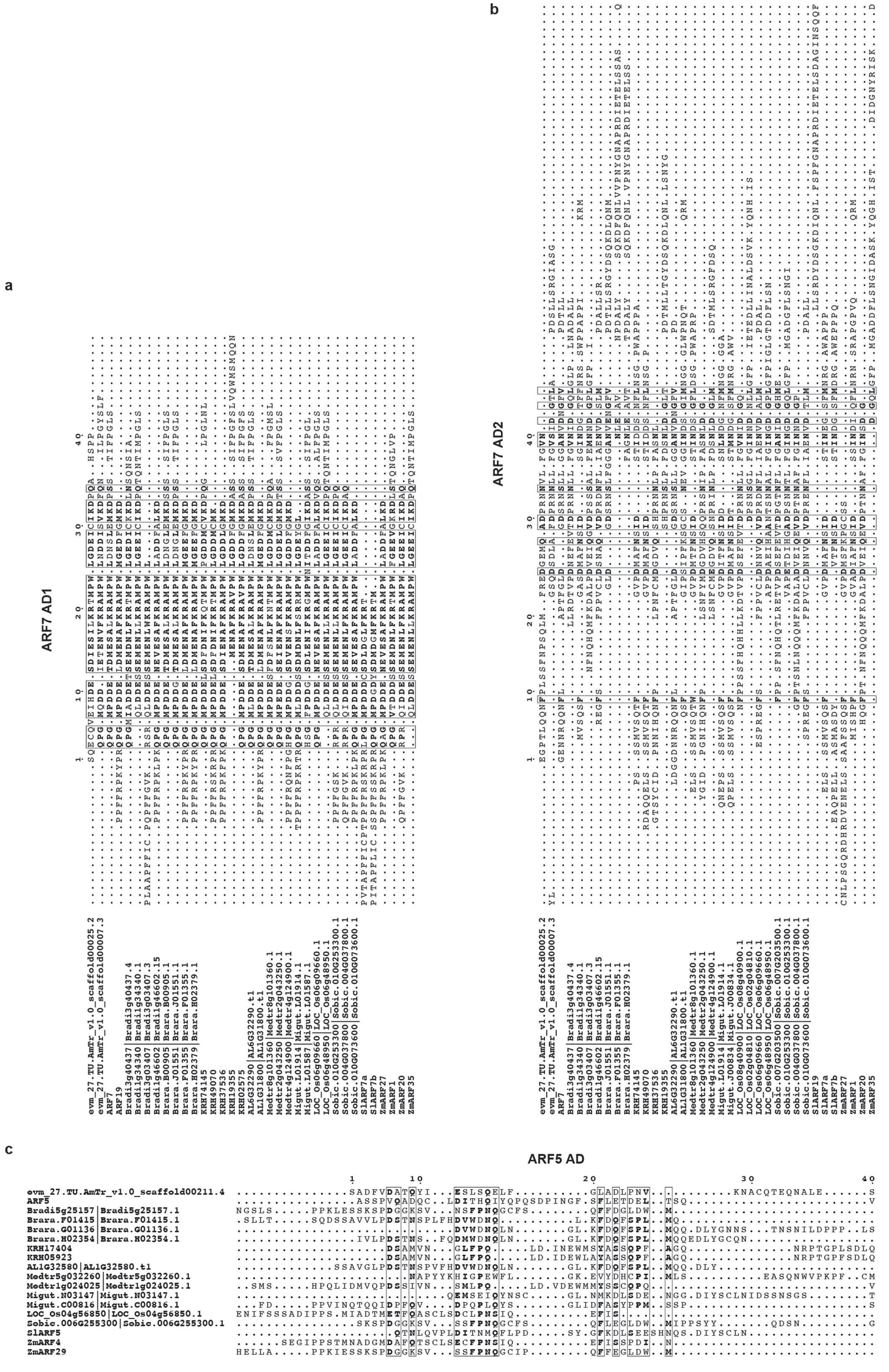
Extended Data Fig. 4 | AD subtypes by TF family. Heat map showing the percentage of hits (defined as a PADI score ≥ 1) from each subtype found in each family in *Arabidopsis*.



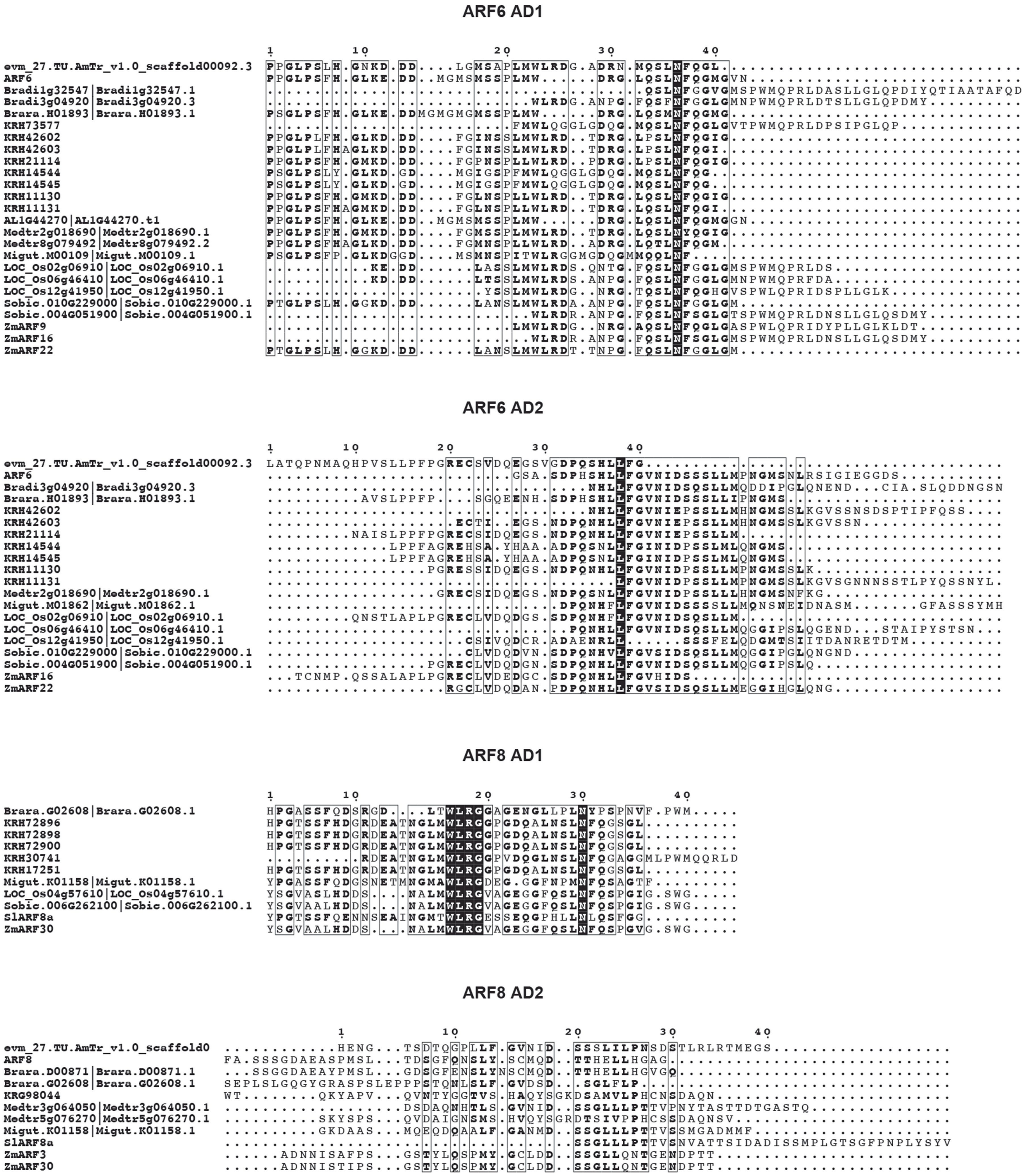
Extended Data Fig. 5 | Comparison of PADI hits to previous activators and distribution of hits across the middle regions of clade-A ARF subclades. **a**, Hummel et al.[11] identified ADs in sixty-eight *Arabidopsis* TFs that could elicit a transcriptional response when transiently expressed in intact tobacco leaves. We identified fragments that could activate transcription in yeast from fifty-six (82%) of the sixty-eight TFs factors identified by Hummel et al. We did not identify fragments that could elicit yeast-based transcription from nine TFs in which Hummel et al. demonstrated transcriptional activity. An additional three TFs were untested in the PADI dataset. It is possible that for the 9 TFs for which Hummel et al. found activation activity and in which we did not identify a hit in our PADI screen that either 1) they contain ADs that are active in plant cells but not in yeast or 2) the nearly intact TFs used by Hummel et al. recruited other coactivators in their system (for example native TFs that contain an AD). **b**–**e**, Orange regions were used to define AD regions for alignment in Extended Data Figs. 7 and 8. **b**, ARF5 clade. **c**, ARF6 clade. **d**, ARF7 clade. **e**, ARF8 clade.



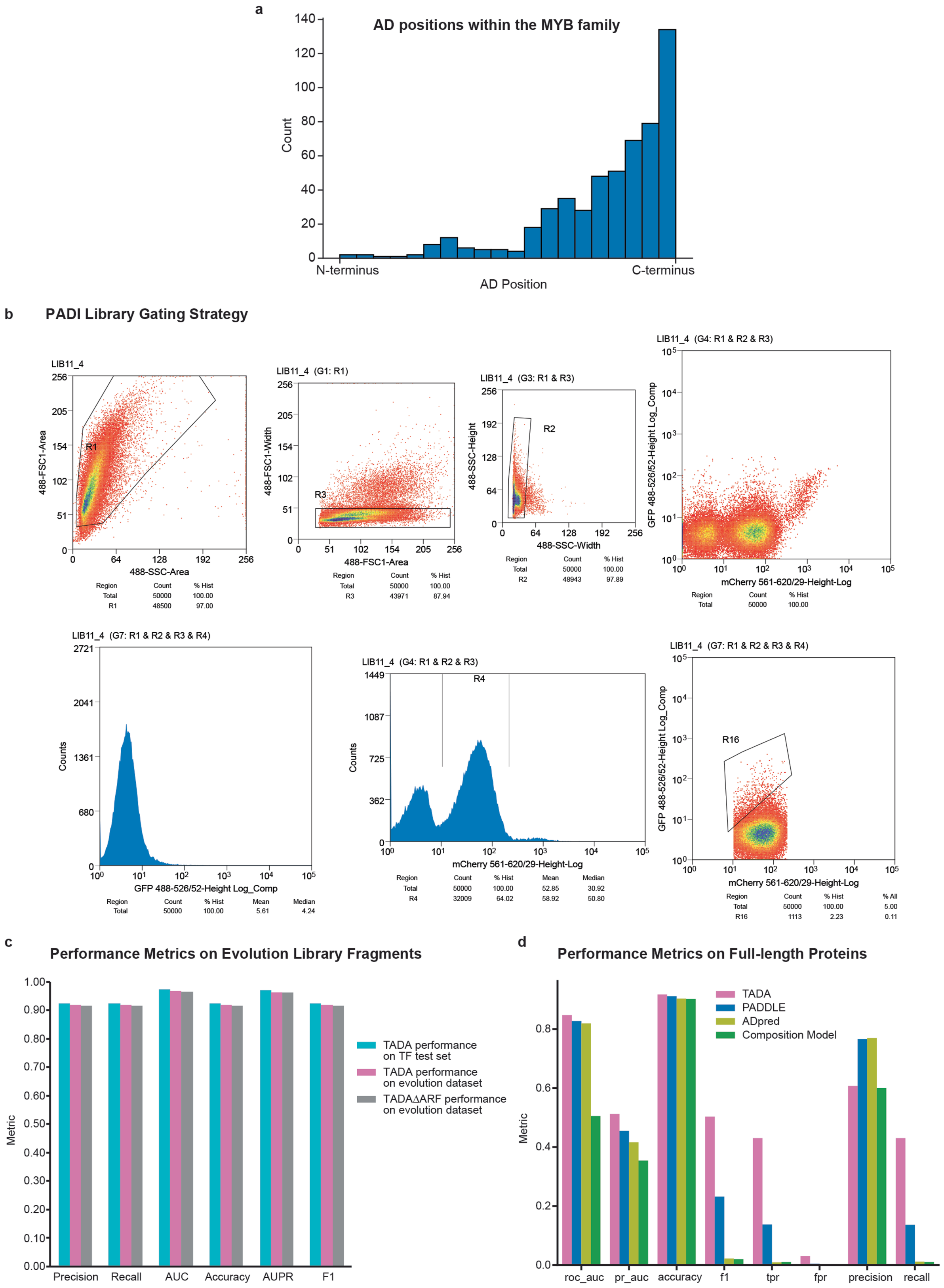
Extended Data Fig. 6 | Phylogeny of examined ARFs. The maximum-likelihood tree was generated using MAFFT alignments of the conserved ARF DBD. Major ARF clades (bright blue, orange and green) and subclades (light blue, orange and green) are annotated. These annotations were used for categorizing sequences in Fig. 4.



Extended Data Fig. 7 | ARF7 and ARF5 subclade AD alignments. a–c, The highest-scoring fragment from each tested ARF within the defined ARF7 and ARF5 AD regions (**a**, ARF7AD1; **b**, ARF7AD2; **c**, ARF5 AD) (orange bars in Extended Data Fig. 5b,d) were used to generate alignments with MAFFT. Alignments were visualized with the ESPript 3.0 webserver. Boxes indicate regions in which 50% of amino acid residues share sequence similarity based on biochemical properties. Bolded residues are the amino acids with shared properties within the region. Black boxes represent sequence conservation.

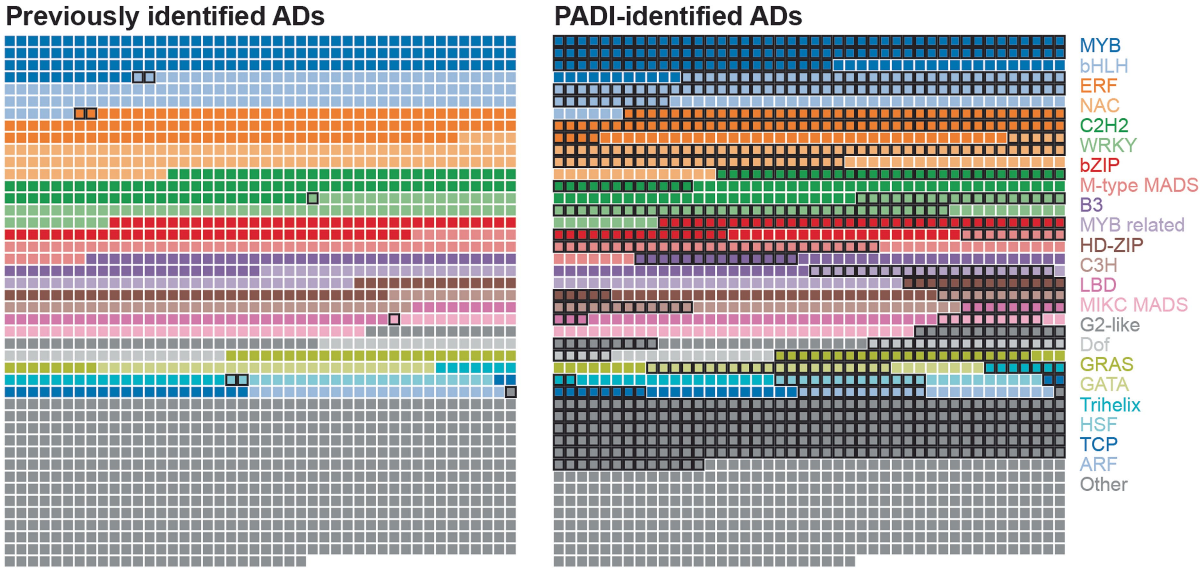

ARF6 AD1
1
10
20
30
40
evm_27.TU.AmTr_v1.0_scaffold00092.3
ARF6
Bradi1g32547|Bradi1g32547.1
Bradi3g04920|Bradi3g04920.3
Brara.H01893|Brara.H01893.1
KRH73577
KRH42602
KRH42603
KRH21114
KRH14544
KRH14545
KRH11130
KRH11131
AL1G44270|AL1G44270.t1
Medtr2g018690|Medtr2g018690.1
Medtr8g079492|Medtr8g079492.2
Migut.M00109|Migut.M00109.1
LOC_Os02g06910|LOC_Os02g06910.1
LOC_Os06g46410|LOC_Os06g46410.1
LOC_Os12g41950|LOC_Os12g41950.1
Sobic.010G229000|Sobic.010G229000.1
Sobic.004G051900|Sobic.004G051900.1
ZmARF9
ZmARF16
ZmARF22
ARF6 AD2
1
10
20
30
40
evm_27.TU.AmTr_v1.0_scaffold00092.3
ARF6
Bradi3g04920|Bradi3g04920.3
Brara.H01893|Brara.H01893.1
KRH42602
KRH42603
KRH21114
KRH14544
KRH14545
KRH11130
KRH11131
Medtr2g018690|Medtr2g018690.1
Migut.M01862|Migut.M01862.1
LOC_Os02g06910|LOC_Os02g06910.1
LOC_Os06g46410|LOC_Os06g46410.1
LOC_Os12g41950|LOC_Os12g41950.1
Sobic.010G229000|Sobic.010G229000.1
Sobic.004G051900|Sobic.004G051900.1
ZmARF16
ZmARF22
ARF8 AD1
1
10
20
30
40
Brara.G02608|Brara.G02608.1
KRH72896
KRH72898
KRH72900
KRH30741
KRH17251
Migut.K01158|Migut.K01158.1
LOC_Os04g57610|LOC_Os04g57610.1
Sobic.006G262100|Sobic.006G262100.1
SlARF8a
ZmARF30
ARF8 AD2
1
10
20
30
40
evm_27.TU.AmTr_v1.0_scaffold0
ARF8
Brara.D00871|Brara.D00871.1
Brara.G02608|Brara.G02608.1
KRG98044
Medtr3g064050|Medtr3g064050.1
Medtr5g076270|Medtr5g076270.1
Migut.K01158|Migut.K01158.1
SlARF8a
ZmARF3
ZmARF30

Extended Data Fig. 8 | ARF6 and ARF8 subclade AD alignments. The highest-scoring fragment from each tested ARF within the defined AD regions (orange bars in Extended Data Fig. 5c,e) were used to generate alignments with MAFFT. Alignments were visualized with the ESPript 3.0 webserver. Boxes indicate regions where 50% of amino acid residues share sequence similarity based on biochemical properties. Bolded residues are the amino acids with shared properties within the region. Black boxes represent sequence conservation.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | MYB family ADs and prediction performance of TADA on the ARF evolution dataset. a, Histogram of all AD hits (defined as a PADI score of greater than or equal to 1 and from an IDR) from the MYB family. Each bar represents the number of ADs found in each 5% interval of the protein length. These results show that MYB ADs are enriched in the final 15% of tested TFs. **b**, Representative gating strategy for all PADI libraries. Yeast cells were gated based on size to exclude doublets (R1 and R3). Single cells were then gated to exclude those with mCherry signal below background (R4) when compared to mCherry negative cells. The mCherry-positive cells were then binned and sorted into twelve populations based on the GFP:mCherry ratio. **c**, Prediction performance of TADA, and the TADAΔARF variation. TADA performance on the PADI data test set and the ARF evolution dataset in terms of precision, recall, area under the receiver operating curve (AUC), accuracy, AUPR and F1 score. We further validated the generalization of TADA by retraining TADA on the original training dataset but withholding the ARF sequences (2,046 of the 70,937 sequences), which we called TADAΔARF. This approach prevents TADA from memorizing/overfitting ARF sequences. **d**, Prediction performance of TADA, PADDLE, ADPred, and the composition model in terms of area under the receiver operating curve (roc_auc), area under the precision recall curve (pr_auc), accuracy, F1 score, true positive rate (tpr), false positive rate (fpr), precision, and recall when tested on the ARF evolution dataset. Because each of these predictors subdivides sequences differently and used different fragment lengths for training, we compared their performance on full-length protein sequence from the evolution dataset.



Extended Data Fig. 10 | *Arabidopsis* TFs with identified ADs. Waffle plots of the 1,918 *Arabidopsis* TFs analysed. Those with previously identified ADs are marked with a black box in the left waffle plot. The right waffle plot depicts those with activating fragments identified by PADI.

nature portfolio

Corresponding author(s): Lucia C. Strader

Last updated by author(s): May 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Cell sorting populations were collected using Summit Software for BC Astrios.
Additional protoplast and yeast flow cytometry data was collected with CytExpert for Cytoflex S.

Data analysis

Paired raw Fastq files from each library were aligned to fragment DNA sequences using BWA-mem aligner and SAMtools to generate BAM files and determine fragment counts.
Data analysis was conducted using Numpy, Pandas, scikitlearn, and custom Python scripts to determine AD scores.
Analysis on flow cytometry data was done using flowTime R package and flowkit, seaborn and pandas python packages as described in the methods.
All graphs were generated with Seaborn and Matplotlib python packages.
Statistical analysis on protoplast flow cytometry was conducted using JMP17.
Sequence features were determined using LocalCider and MetaPredict2 in Python.
The TADA Neural Network is available on https://github.com/LisaVdB/TADA and described in the methods as well as the software submission sheet.
SPARROW is available at https://github.com/idptools/sparrow
ALBATROSS is available at https://github.com/idptools/goose/

Software Version Purpose
BWA 0.7.15 NGS sequence Alignment
SAMtools 1.10 BAM file generation and read count extraction
Numpy 1.22.3 Data analysis in Python 3.8.12
Pandas 1.4.1 Data analysis in Python 3.8.12

Seaborn 0.13.0 Data visualization and graphing in Python 3.8.12
Matplotlib 3.7.1 Data visualization and graphing in Python 3.8.12
Scikitlearn 1.2.0 Data normalization
LocalCider 0.1.19 Amino Acid Feature Extraction
Meatpredict 2.2 Disorder Prediction
JMP17 17.0.0 Statistical analysis
FlowKit 1.0.1 Extract flowcytometry data from .fcs files for analysis
Python 3.8.12

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Library sequencing data has been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO series access number GEO: GSE234215. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We calculated that a minimum of 60,000 cells per bin per pooled library provided sufficient coverage of the tested library for downstream analysis. No calculations were performed to predetermine sample size. Biological and technical replicates were performed as described in the methods for each experiment and conform to standards in the field. Exact n for each experiment is listed in each figure legend panel.
Data exclusions	No data was excluded from this study.
Replication	All experimentation was conducted as described in the methods with replications listed. Internal controls in each flow sort were checked to determine the reproducibility of the assay across sorts. All attempts at replication were successful.
Randomization	Complex fragment libraries were generated from transcription factors spanning the Arabidopsis thaliana genome and tested in genomic order, making tested pools functionally random. Thus randomization is not applicable since data are quantitative and were not subjectively grouped.

Blinding	Not applicable since data are quantitative and were not subjectively grouped.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
□	☒	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	□	Public health
☒	□	National security
☒	□	Crops and/or livestock
☒	□	Ecosystems
☒	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	□	Demonstrate how to render a vaccine ineffective
☒	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	□	Increase transmissibility of a pathogen
☒	□	Alter the host range of a pathogen
☒	□	Enable evasion of diagnostic/detection modalities
☒	□	Enable the weaponization of a biological agent or toxin
☒	□	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

The Col-0 ecotype was used for protoplast experiments. Col-) is available commercially and from the ABRC stock center.

Novel plant genotypes

N/A

Authentication

N/A

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

For yeast-based data, Sorted cells were grown in SC media overnight and then used to extract genomic DNA. For protoplast-based data, cells were extracted from soil-grown plants immediately prior to transfection.

Instrument

Cell sorting was conducted on Beckman Coulter Astrios at The DCI Flow Cytometry Core at Duke University (PADI and ARF Evolution Libraries) and on a BD Aria-II machine at Washington University in St. Louis (Pilot Assay). Additional flow cytometry on protoplasts and yeast was conducted using the Beckman Coulter Cytoflex S Flow Cytometer and Attune NxT Acoustic focusing cytometer as described in the methods.

Software

Analysis on flow cytometry data was done using flowTime R package and flowkit, seaborn and pandas python packages as described in the methods.

Cell population abundance

The entire population of cells were flow sorted into 12 bins for all cell sorting assays. All yeast and protoplast flow cytometry experiments were conducted as described in the methods.

Gating strategy

For cell sorting, initial gating on yeast cells was generated using FSC and SSC. Yeast cells with and without mCherry expression constructs were used to set mCherry positive populations. Activity was scored as a ratio of GFP to mCherry signal in positive cells, all cells were included. For ptotoplast-based experiments, gating was on SSC and either mScarlet or mNeonGreen, depending on the assay, as described.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.