Title

Systematic identification of transcriptional activation domains from non-transcription factor proteins in plants and yeast

Permalink

https://escholarship.org/uc/item/2r74g38f

Journal

Cell Systems, 15(7)

ISSN

2405-4712

Authors

Hummel, Niklas FC
Markel, Kasey
Stefani, Jordan
et al.

Publication Date

2024-07-01

DOI

10.1016/j.cels.2024.05.007

Cell Systems

Report

Systematic identification of transcriptional activation domains from non-transcription factor proteins in plants and yeast

Highlights

- Screening of >17,000 peptides with activator activity from non-transcription factors
- Key residues are biased in their distribution in strongly activating peptides
- Active peptides occur throughout all organelles in a plant and a fungus
- Benchmarking of 51 new activating peptides in a plant gene expression system

Authors

Niklas F.C. Hummel, Kasey Markel, Jordan Stefani, Max V. Staller, Patrick M. Shih

Correspondence

mstaller@berkeley.edu (M.V.S.),
pmshih@berkeley.edu (P.M.S.)

In brief

Hummel et al. utilize a yeast gene expression platform to characterize the potential of >17,000 peptides derived from non-transcription factor proteins to activate transcription. They find a positional bias of key residues in activators, curate a list of putative coactivators, and cross-validate active peptides in a plant gene expression platform.

Hummel et al., 2024, Cell Systems *15*, 662–672
July 17, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2024.05.007

CellPress

CellPress
OPEN ACCESS

Cell Systems

Report

Systematic identification of transcriptional activation domains from non-transcription factor proteins in plants and yeast

Niklas F.C. Hummel,[1,2,3,4] Kasey Markel,[1,2,3] Jordan Stefani,[5] Max V. Staller,[5,6,7,*] and Patrick M. Shih[1,2,3,8,9,*]
[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA
[2]Feedstocks Division, Joint BioEnergy Institute, Emeryville, CA 94608, USA
[3]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4]Department of Biology, Technische Universität Darmstadt, 64287 Darmstadt, Germany
[5]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA
[6]Center for Computational Biology, University of California, Berkeley, CA 94720, USA
[7]Chan Zuckerberg Biohub-San Francisco, San Francisco, CA 9415, USA
[8]Innovative Genomics Institute, University of California, Berkeley, CA 94720, USA
[9]Lead contact
*Correspondence: mstaller@berkeley.edu (M.V.S.), pmshih@berkeley.edu (P.M.S.)
https://doi.org/10.1016/j.cels.2024.05.007

SUMMARY

Transcription factors can promote gene expression through activation domains. Whole-genome screens have systematically mapped activation domains in transcription factors but not in non-transcription factor proteins (e.g., chromatin regulators and coactivators). To fill this knowledge gap, we employed the activation domain predictor PADDLE to analyze the proteomes of *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. We screened 18,000 predicted activation domains from >800 non-transcription factor genes in both species, confirming that 89% of candidate proteins contain active fragments. Our work enables the annotation of hundreds of nuclear proteins as putative coactivators, many of which have never been ascribed any function in plants. Analysis of peptide sequence compositions reveals how the distribution of key amino acids dictates activity. Finally, we validated short, "universal" activation domains with comparable performance to state-of-the-art activation domains used for genome engineering. Our approach enables the genome-wide discovery and annotation of activation domains that can function across diverse eukaryotes.

INTRODUCTION

Transcription factors (TFs) regulate gene expression by binding specific DNA regions with their DNA-binding domains (DBDs) and interacting with protein complexes through their transcriptional effector domains.[1] Transcriptional effector domains that promote transcription are further classified as activation domains (ADs). In this work, we will refer to all short protein sequences that function in AD assays as peptides with AD activity (pADs), regardless of their native function in their respective parent protein. New high-throughput methodologies have helped characterize the regulatory activity of transcriptional effector domains *en masse* in yeast, human, and fly models,[1–8] and these approaches are beginning to be implemented to study plants.[9] Still, most studies have largely biased their focus on TFs, leaving other nuclear proteins and their potential role in transcription understudied. Moreover, non-nuclear proteins have been demonstrated to be involved in transcriptional regulation. For example, Notch1, a plasma membrane localized protein in multicellular animals, contains a C-terminal AD that is cleaved, processed, and localized to the nucleus to induce transcription.[10] Similarly, the cell-adhesion protein beta-catenin is localized to the nucleus when multimerized, where it acts as a transcriptional coactivator in fly and vertebrates, and the closest plant homologs have been linked to root development.[11–13] Thus, there is evidence of proteins with ADs outside of standard TFs. To more thoroughly study all putative proteins that may be involved in transcriptional activation, genome-wide screens of all proteins—not just TFs or nuclear proteins—are needed to identify previously unannotated molecular factors that may play a role in transcriptional regulation.

The availability of large AD activity datasets has enabled the development of deep convolutional neural networks that can predict the activity of eukaryotic ADs from protein sequences.[5,14] These models have helped elucidate how specific amino acid (AA) sequence features of acidic ADs enable their transcriptional activation activity.[15] Notably, because current AD predictive models have been trained on large datasets from select organisms (i.e., yeast and human), the predictive strength of these models in other eukaryotes has not been well defined. The published neural network models have never been directly tested in a large-scale experiment.

Mechanistic studies have shown how acidic residues promote the exposure of hydrophobic residues that, in turn, are essential for AD activity.[16] Hence, the distribution of acidic and hydrophobic residues is key because hydrophobic clusters can lead to the intramolecular collapse of the AD, diminishing its activity.[6] The recently proposed acidic exposure model links these observations to structural disorder in ADs, where acidic residues stabilize an energetically unfavorable solvent exposure of hydrophobic residues, which, in turn, interact with coactivators to promote transcription in a transiently structured fashion.[6] Thus, sequence composition, structural disorder, and small sequence motifs in ADs have been linked to defining AD activity, but we still lack a comprehensive understanding of how positional sequence features affect AD function.

Eukaryotic transcription is facilitated by TFs, coactivators, and chromatin regulators. Coactivators can function as adaptors between TFs and RNA polymerase II or the general transcription apparatus, whereas other coactivators modify chromatin to help transcription of chromatinized templates or help with unwinding DNA, all resulting in higher transcriptional output.[10–13,17–19] Coactivators interface between TFs and RNA polymerase but do not directly bind DNA, functionally separating them from TFs. Coactivators and chromatin regulators can contain ADs,[7,14,20] marking activator activity non-unique to TFs; still, there has been a dearth of high-throughput studies focused on identifying new coactivator candidates due to the multitude of mechanisms that coactivators use to promote transcription. Hence, the occurrence of ADs in nuclear non-TF genes could indicate that a given protein is involved in transcription and help annotate previously unknown transcription associated genes and coactivators.

Genome-scale studies characterizing TFs in plants have provided the foundational understanding of the complex regulation that underlies plant development, adaptation, and overall physiology.[21,22] However, transcriptional coactivators have been much less studied and leave a large blind spot in our understanding of their role in transcription. Unlike unicellular systems that are more readily tractable to screening massive libraries and cell sorting, the complex physiology and cell wall of plants have hindered the implementation of high-throughput methods for the characterization of ADs in plants. As a result, our understanding of plant ADs and the role of potential coactivators pales in comparison with other better-studied model eukaryotes (e.g., yeast). We previously reported that a machine learning model trained on data from a large library of synthetic activators from yeast can correctly predict ADs in plant TFs[9]; however, it is still unclear how applicable and scalable these models are in plant systems, necessitating further evaluation of more plant ADs predicted by yeast models. A larger set of validated plant ADs would allow the comparison of sequence features in plant ADs with observations in other well-studied eukaryotes. Moreover, studying pADs from non-TF genes can help us identify and annotate previously unknown proteins involved in transcriptional regulation and deepen our understanding of the features defining AD strength in plants.

Here, we assess the transcriptional activity of predicted pADs derived from non-TF proteins from yeast and plants. We generated a library of 18,000 synthetic TFs carrying predicted pADs from non-TF genes with pADs derived from *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Arabidopsis thaliana* (*A. thaliana*). We show that 753 (89%) of 846 parent genes in the library contain pADs capable of promoting transcription in yeast. Notably, pADs were not limited to nuclear genes, and many pADs were found in a wide range of protein families localized to other organelles. We find a positional distribution of key AAs that make large contributions to pAD activity, providing insight into sequence "grammar." Furthermore, we show how strong pADs from the library activate transcription in plants, marking them as universal pADs. Our large interspecies dataset provides both the foundational knowledge to explore the role of pADs in non-TFs and a large set of new pADs that can be readily integrated into genome engineering efforts across phylogenetically diverse eukaryotes.

RESULTS

Characterization of a library of non-TF ADs mined from yeast and plant proteomes

We aimed to systematically discover previously uncharacterized pADs derived from non-TF proteins in two model eukaryotic systems, *A. thaliana* and *S. cerevisiae*. In previous work, we have shown that AD predictors derived from fungal data can accurately predict ADs in plant TFs and that plant ADs function in yeast.[9] Here, we leveraged this result to predict activators in plant and yeast proteins, followed by high-throughput experimental validation in yeast.

To extract potential ADs from both proteomes, we utilized PADDLE, a neural network model capable of predicting acidic ADs in 53-AA-long peptides.[14] We computationally chopped each proteome in 53 AA tiles spaced every one residue, yielding 9,211,910 tiles in *A. thaliana* and 2,646,422 tiles in *S. cerevisiae* derived from 27,082 and 6,455 proteins, respectively (Figures 1A and 1B; Data S1). We used PADDLE to predict the potential of all tiles to activate transcription. We then used TF databases for both species—PlantTFDB v5.0 and Yeastract+—to remove all tiles derived from TF sequences.[23,24] We found that tiles from non-TF genes had a similar dynamic range of predicted activity as tiles from TF genes (Figures S1A and S1B), and in Arabidopsis, the strongest predicted tile occurred in a non-TF gene (AT5G07570.1). We defined all genes that we mined tiles from as parent genes because multiple tiles can come from a single protein sequence. We then selected 12,000 tiles from *A. thaliana* and 6,000 tiles from *S. cerevisiae* with the highest predicted activation score, yielding a 18,000-tile library derived from 447 Arabidopsis and 402 yeast parent proteins, respectively. We chose to include overlapping tiles to increase accuracy and resolution. To gauge the subcellular localization of parent proteins of the library, we utilized SUBA5 for Arabidopsis and YeastGFP/YPL+ to annotate localization.[25–27] There was a total of 214 parent proteins localized to the nucleus in Arabidopsis and 107 in yeast. Non-nuclear genes were localized throughout all subcellular locations in both species (Tables S1 and S2). This diversity of localization suggests that peptides predicted to be ADs occur throughout various organelles across both proteomes.

To experimentally characterize and validate our library, we used a previously established expression system utilizing synthetic TFs in yeast.[16] In this expression system, each tile is fused to a synthetic TF, consisting of (1) mCherry for normalization of TF concentration to generated reporter signal as a N-terminal

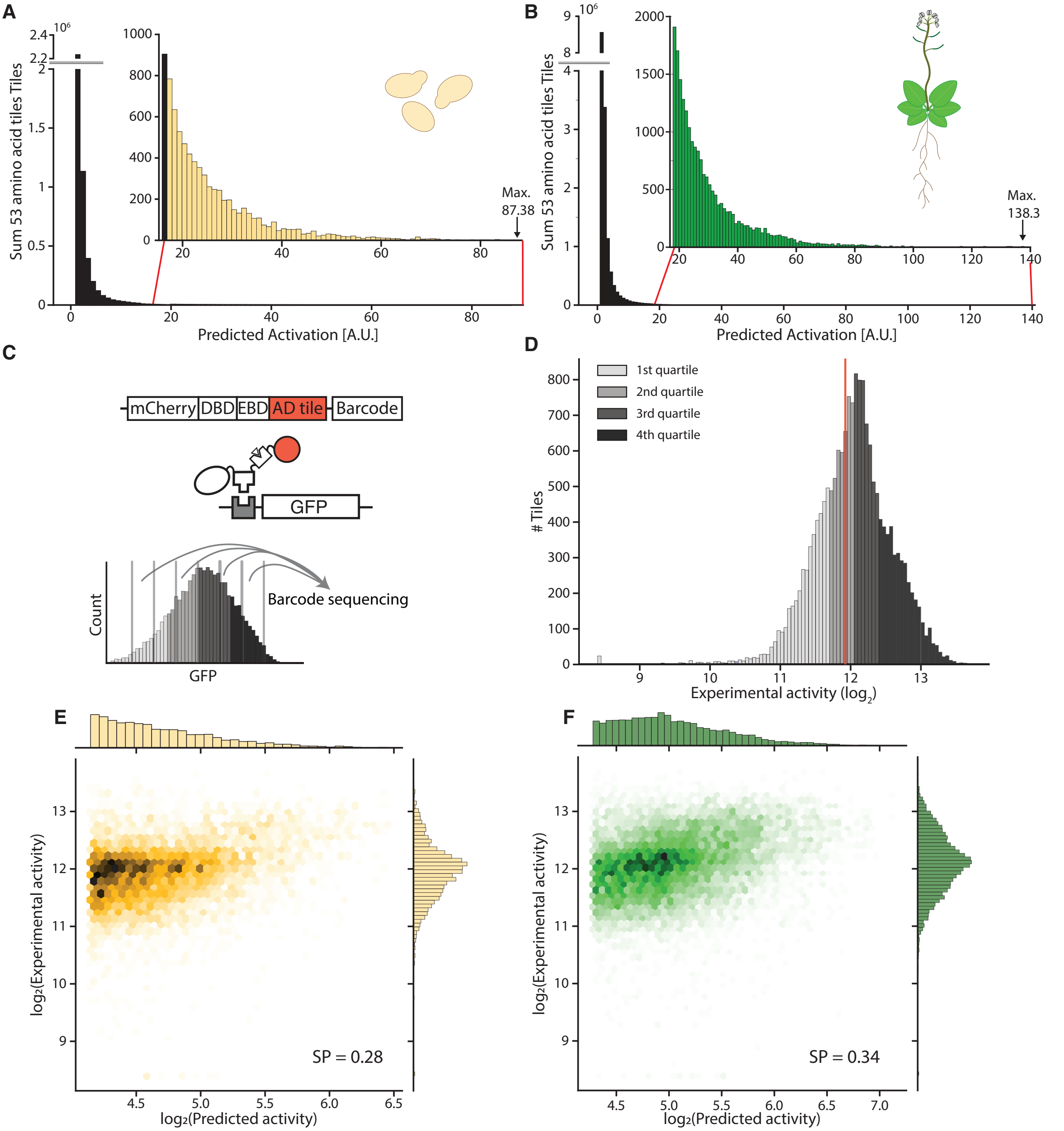

Figure 1. Proteome-wide characterization of pADs mined from non-TF plant and yeast proteins

(A and B) Histogram showing the PADDLE predicted activity of all 53 amino acid tiles in (A) *S. cerevisiae* and (B) *A. thaliana* proteome. Inlet figures show the magnified areas of the histogram the putative AD candidates for the libraries were chosen from (marked in red).

(C) The 12,000 strongest *A. thaliana* and 6,000 strongest *S. cerevisiae* tiles were characterized as a synthetic TF library in *S. cerevisiae*. Activator activity was calculated by abundance of barcodes in bins established during FACS sorting. DBD, DNA-binding domain; EBD, estrogen binding domain.

(D–F) (D) Activity of every tile as determined by FACS and consecutive barcode sequencing across eight bins. Red bar indicates activity of no-AD controls. Predicted activity vs. experimentally observed activity of all tiles from (E) *S. cerevisiae* and (F) *A. thaliana*. SP, Spearman's R.

fusion, (2) the orthogonal murine Zif268 DBD, (3) a human estrogen response domain to make the system inducible with ß-estradiol, (4) the 53-AA-long AD candidate, and (5) a unique barcode in the 3′ UTR marking candidate identity in the library (Figure 1C). The associated reporter consists of six copies of the Zif268 binding sites upstream of a modified GAL1 promoter driving GFP.[28]

Both the reporter and the synthetic TF were integrated into the genome of *S. cerevisiae* to reduce expression variability. We used fluorescence-activated cell sorting (FACS) to sort the library (see STAR Methods), and experimentally validated the activity of 17,553 tiles (97.5% of total library) from 846 parent genes with high reproducibility between replicates (Pearson's r = 0.82) (Figure S2; Tables S3 and S4). Multiple DNA barcodes were used for each tile to further measure the variability resulting from multiple integrations, thereby increasing accuracy of measurements as previously shown.[16]

The experimental activity of our library allowed us to evaluate the accuracy of PADDLE predictions. In the PADDLE training dataset, ~30% of TF-derived tiles showed activity, and the model achieved reliable qualitative and quantitative prediction of ADs (~10,400 tiles, Pearson's r = 0.81). In our library, tile activity ranged over three orders of magnitude with 56.5% of the library showing significant activity above no-TF control levels (Figure 1D). This was the largest fraction of active tiles we have observed using this system.[6,16] Parts of the library activated transcription equally or stronger than Gcn4-AD and -VP16-AD controls (Figures S3A and S3B). We found 89.0% of parent genes (753 out of 846) of tiles in our library to contain at least one tile with activator activity, demonstrating that tiles that can function as ADs are widespread throughout non-TF genes. This result shows how PADDLE can, in most cases, correctly localize pADs in proteins but its architecture, which predicts AD likelihood and then extrapolates activity predictions, is not rigorous enough to predict both qualitative and quantitative aspects of ADs. Overall, we identified 9,911 pADs derived from non-TFs, providing a rich resource for engineering efforts.

Single-AA tiling unravels positional effects and key residues dictating AD activity

Although the large fraction of active tiles supports the ability of PADDLE to localize pADs in protein sequences, the quantitative predictions of pAD strength did not correlate as strongly with our experimental results (Spearman's r = 0.35, Pearson's r = 0.33) (Figure S4). Notably, the PADDLE algorithm predicted activity with higher accuracy in *A. thaliana* than in *S. cerevisiae* AD populations with moderate Spearman correlation coefficients of 0.34 and 0.28, respectively (Figures 1E and 1F). We found that in 13% of parent genes, the strongest predicted tile precisely overlapped with the strongest experimental tile. Hence, PADDLE correctly identifies the general location of pADs but struggles to accurately predict the quantitative strength of the respective pAD and precise pAD boundaries. Our results support previous evidence that there are positional effects of AA residues dictating pAD activity and demonstrate that PADDLE cannot resolve these effects.

To further investigate the discrepancies between PADDLE predictions and observed activities, we examine how AA composition and positional context may play a role in defining AD activity. We recovered previously observed relationships between AA sequence and activity, i.e., that W, F, L, D, and E are associated with activity and K and R are not.[5,14,29,30] To compare tile populations, we split the entire library into four equal quartiles, ranging from weakest to strongest activity. Notably, the average AA composition of tiles in all four quartiles was nearly identical (Figure S5). This result emphasizes the power of this dataset to probe sequence grammar. We found that the enrichment of acidic-aromatic dipeptides in strong ADs was reproducible in our dataset (Figures S6A–S6D; Data S2).[5] We hypothesized that the positional distribution of functional AAs is key to AD activity.

To gauge the positional information encoded in each quartile, we measured the local density of all residues along each tile of every quartile, where density is the frequency of the respective AA in a 5 AA window. We then grouped AA groups linked to AD activity and computed the density at each position of the 53 AAs of every tile of each respective quartile. We found the density of W, F, Y, and L residues, which are closely linked to AD activity, to be overall higher in the highest activity quartile, and density was higher in the C terminus when compared with tiles in other quartiles (Figures 2A and S7A). This finding supports the acidic exposure model, which predicts that hydrophobic residues at the C terminus will be more exposed to solvent and make larger contributions to activity. All quartiles had a low density of hydrophobic residues in the N terminus, suggesting that PADDLE has partially learned this signal. Correspondingly, the fourth quartile displayed a weaker density of acidic residues in the C terminus and was more evenly distributed throughout the entire tile, whereas the weaker quartiles had a stronger enrichment of acidic residues in the C terminus and depletion in the N terminus (Figures 2B and S7B). These results support the hypothesis that the occurrence of key AAs—in this case, hydrophobic and acidic residues—alone does not correlate with activity but rather their distribution along the AD, further supporting the acidic exposure model.[6] The dip in acidic residues at the extreme C terminus in the fourth quartile was surprising to us because we and others have previously shown that acidic residues near or next to aromatic residues boost activity. We speculate that the C terminus is highly exposed, both by virtue of being on the end and because of the additional acidity of the C-terminal backbone.

We further studied the role of intrinsic disorder on activity in our library. Virtually all ADs are intrinsically disordered, and the acidic exposure model suggests that more disordered sequences are likely more active. To gauge the disorder of tiles fused to synthetic TFs in the respective quartile, we utilized the disorder predictor Metapredict V2.[31] We found that all quartiles displayed increased disorder in their N terminus, suggesting that initial disorder in the tile is important for activity (Figure 2C). In all quartiles, disorder dropped drastically in the C terminus and the fourth quartile showed increased disorder throughout the entire tile (Figure S8). The disorder in the N terminus implies that an entropic spacer or expanded linker between the estrogen binding domain and the AD increases activity. It is further possible that some sequences in this library may be interacting with the estrogen binding domain, which could drive collapse and decrease activity. The drop in predicted disorder at the C terminus is likely the consequence of the increased density of aromatic residues (Figure 2C). In consequence, current predictive models are still missing necessary positional information of key residues in ADs that need to be incorporated into future network training.

High-throughput studies usually scan protein sequences by tiling in step sizes of >10 AAs.[7,14] Here, we decided to tile at single-AAresolution, allowing us to study how single AA changes both from losing and gaining one AA during tiling

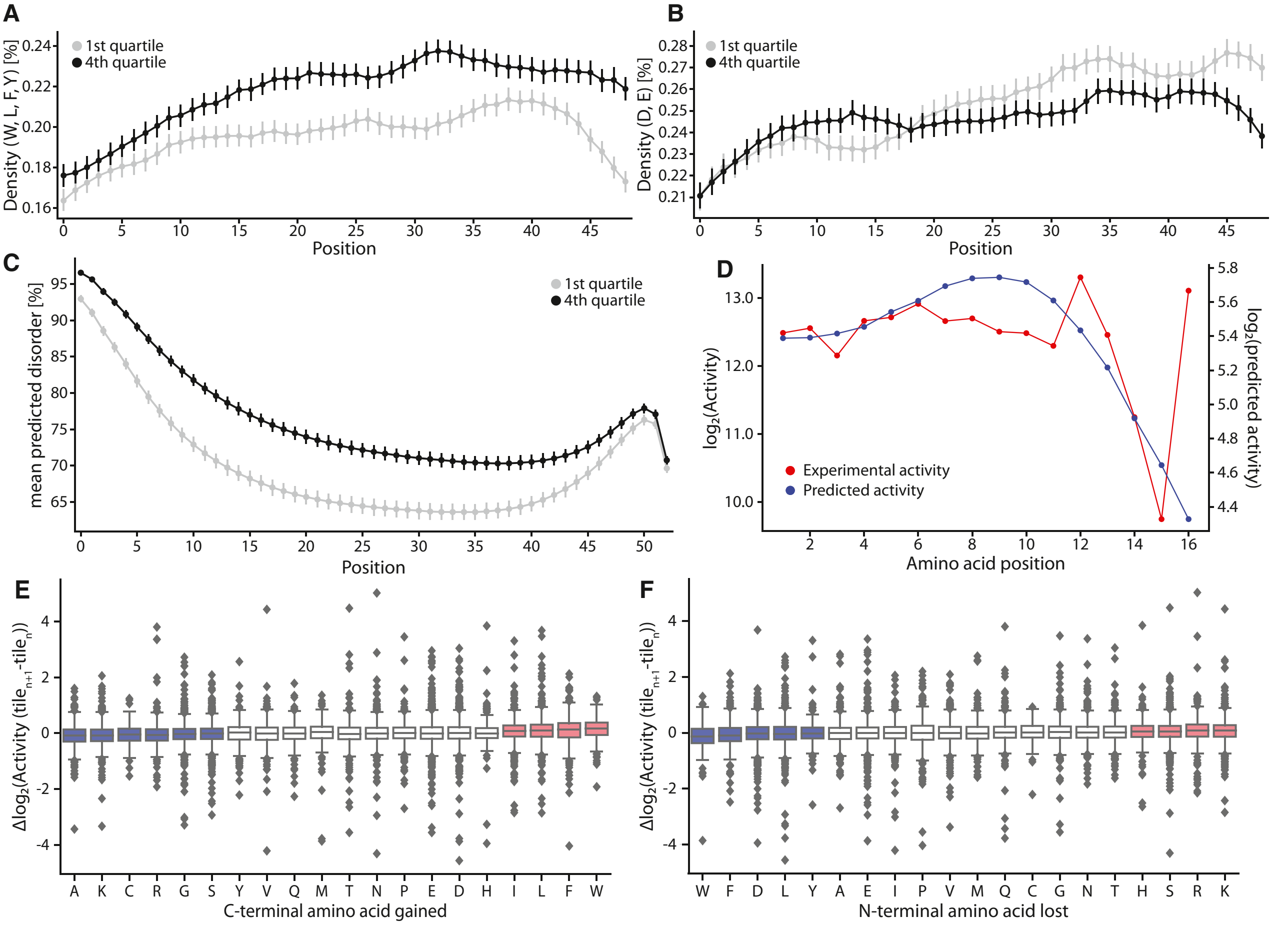

Figure 2. Tiling pADs with single-amino-acid resolution deciphers positional distribution of key amino acid groups

(A and B) Density of functional amino acids across every position of every tile in the quartile with the strongest and weakest activity (4,388 sequences per quartile). Density is calculated in a five amino acid window for each position along the pAD as the average of all (A) hydrophobic residues (W, L, F, and Y), (B) acidic residues (D and E). Error bars indicate the 95% confidence interval.

(C) Mean predicted disorder at every position of tiles fused to the synthetic TF in respective quartiles predicted by MetapredictV2. Error bars indicate the 95% confidence interval.

(D) Example of PADDLE predicted vs. measured activity in single-amino-acid resolution of regions of interest (AT1G14630).

(E and F) Tiling protein sequences with single-amino-acid resolution allows us to observe the effects on activity when (E) gaining a C-terminal amino acid or (F) losing an N-terminal amino acid. Blue colored boxes indicate amino acids significantly decreasing activity, red boxes indicate amino acids significantly increasing activity ($p < 0.05$) as measured by Mann-Whitney U test of the respective amino acid against M, T, N, and P.

directly affect AD activity because the rest of the tiles fully overlap (Figure 2D). At the C terminus, gaining the hydrophobic residues (W, F, and L) significantly enhanced activity as expected (Mann-Whitney U test, $p < 0.05$). Notably, isoleucine, which is not normally linked to enhancing activity, had a stronger positive effect on activity than acidic residues (Figure 2E). The enrichment of isoleucine was only observed in the C-terminal position tiles (Figure S9), suggesting an unknown role of this residue in AD activity. At the N terminus, all effects were smaller than for the C terminus, but losing hydrophobic residues or aspartate significantly decreased activity, while losing positively charged residues like arginine and lysine significantly increased activity, following known rules of activity (Figure 2F).[5,8] Overall, we show that the AA composition alone cannot fully explain AD activity, with the position and distribution of key AAs playing a critical role.

A genome-wide compendium of coactivator candidates in plants

Coactivators provide an interface between TFs and RNA polymerase and are essential for the activation of gene expression. Although there has been significant attention on characterizing TFs at a genome scale, only a limited number of coactivators have been characterized in plants, limiting our ability to fully understand how they regulate transcription. Artificially recruiting coactivators and chromatin regulators to DNA can directly modulate transcription.[3,32] Moreover, coactivators can contain ADs. We reasoned that pADs could occur in any gene involved in transcription because nuclear non-TF genes with pADs represent potential coactivators. Based on subcellular localization data, our library contains pADs from 107 and 200 nuclear non-TF genes in yeast and Arabidopsis, respectively, allowing us to explore their potential coactivator function (Figures 3A and 3B).

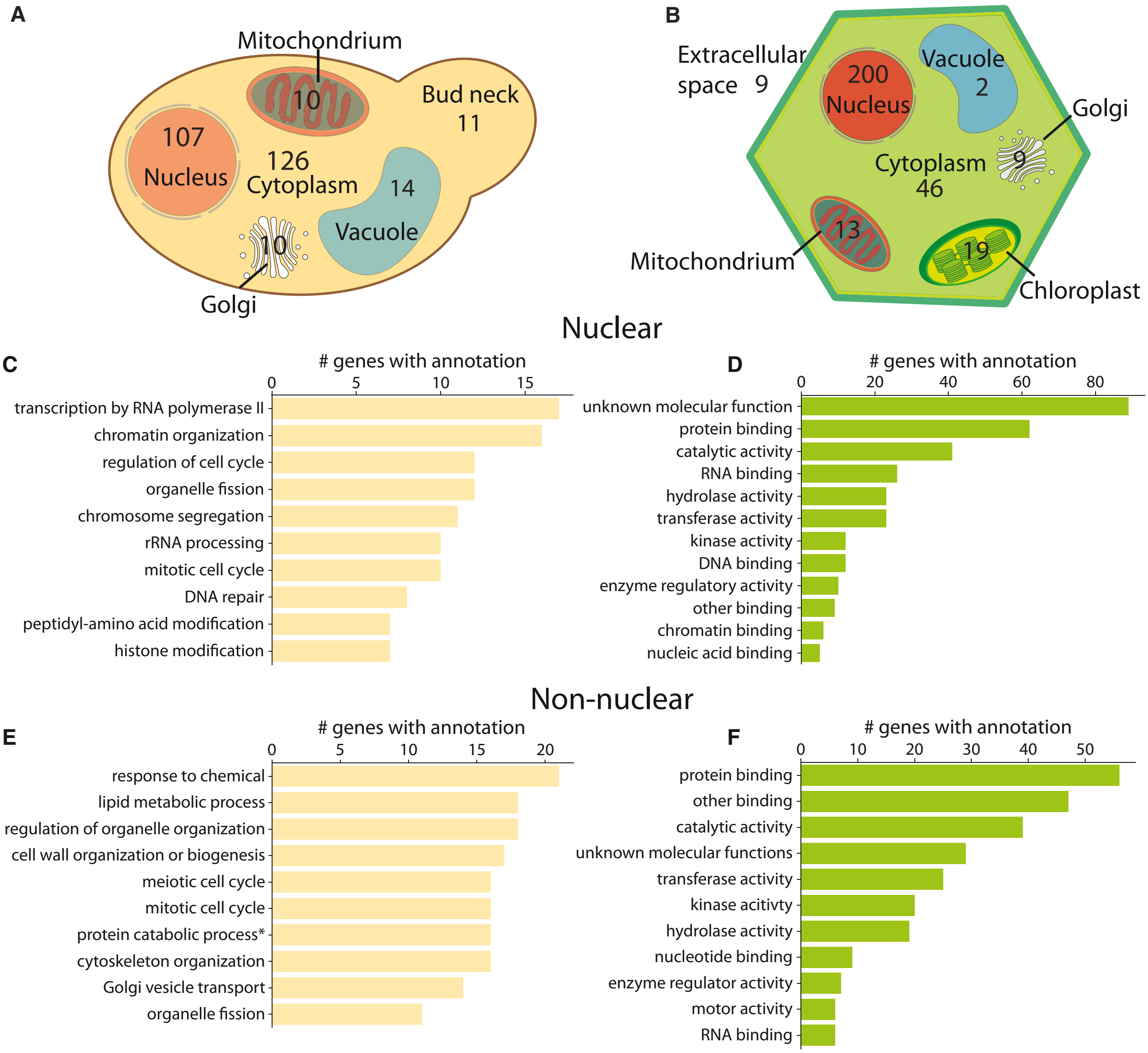

Figure 3. Non-nuclear proteins occurring throughout organelles have predicted pADs in both yeast and plants

Cell schematics with occurrence of parent genes with pADs in the different organelles and cellular structures in (A) *S. cerevisiae* and (B) *A. thaliana*. GO terms associated with nuclear genes in (C) *S. cerevisiae* and (D) *A. thaliana*. GO terms associated with non-nuclear genes in (E) *S. cerevisiae* and (F) *A. thaliana*.

Coactivators have been more thoroughly studied in yeast than in plants; hence, we benchmarked the occurrence of pADs in nuclear non-TF genes from yeast to identify previously unannotated coactivator candidates, as well as provide potential regions with AD activity in known coactivators with ADs. To provide a more comprehensive list of genes with potential ADs, we included parent genes that yielded the 50% strongest pADs in the library. We used Gene Ontology (GO) terms to gauge the function of candidate genes and found most GO terms to be linked to transcription, such as "transcription by RNA polymerase II," "chromatin organization," "regulation of cell cycle," and "histone modification" (Figure 3C; Table S5). As expected, we characterized tiles derived from known coactivators in yeast, namely, IFH1, MED2, ROX3, and NRS1.[33–37] We also found pADs in chromatin regulators, namely, HFI1, CHD1, SFH1, STH1, SDS3, CTI6, and INO80.[35,38–43] Tiles from other proteins involved in transcription included transcription initiation factor eIF4G1, TATA binding factors TAF1, TAF14, and BDF1, and general TF TFG1.[44–47] Notably, we found pADs in two genes of unknown protein family and function (YBL029W and YML108W), which may function as potential coactivators. Candidate genes were also associated with the GO terms "rRNA processing" and "chromosome segregation," raising the question of what roles ADs might play in these proteins. Overall, previous observations of ADs in coactivator complexes and chromatin regulators were supported by our results. Hence, our approach can be used to generate a list of putative genes with regulatory functions for further characterization.

Coactivators in plants have been far less studied and mostly annotated based on homologs from other eukaryotes.[48,49] Hence, the parent genes of tiles from Arabidopsis contained far fewer hits in known transcription associated genes. Of the 211 nuclear non-TF hits, only four had previously been validated to be coactivators, highlighting the opportunity to discover putative plant coactivators. We found pADs in the coactivators MED13 and LNK1/LNK2/LNK3,[50–52] the chromatin regulators HAF2 and SCS2A/B,[53,54] and four transcription elongation factors from family S-II. We also found pADs in three members of the VQ family of suspected transcriptional coregulators that interact with WRKY family TFs during abiotic stress response and four CCT-motif-containing proteins that have been linked to transcriptional elongation in other eukaryotes.[55] Notably, only 23 genes have GO terms linked to transcription with terms such as "chromatin binding," "nucleic acid binding," "DNA binding." The most abundant GO term was "unknown molecular functions" with 89 associated genes, highlighting putative coactivators that have not yet been characterized (Figure 3D, Table S6). Other nuclear genes with pADs were either not previously associated with transcription or have never been studied before, suggesting that there may be plant-specific coactivators that cannot be annotated purely based on sequence homology to other eukaryotes. Our results supply an extensive list of putative coactivators in Arabidopsis, which should accelerate the proper characterization of such proteins, ultimately providing useful targets to modulate and engineer key traits in plants.

Non-nuclear proteins throughout all organelles contain pADs

Non-nuclear proteins can contain ADs and influence transcription via relocation to the nucleus as exemplified in the examples of Notch1 and beta-catenin.[10,11] We investigated the prevalence of pADs that occur in proteins outside of the nucleus. We again only focused on parent genes harboring tiles of the 50% strongest experimentally validated pADs, yielding 136 Arabidopsis and 207 yeast non-nuclear genes (Figures 3A and 3B). We found 46 and 126 cytosolic genes in Arabidopsis and yeast, respectively. These genes are candidates for relocalization into the nucleus, similar to Notch1 and beta-catenin. Besides these candidates in Arabidopsis, we found AD containing genes in the chloroplast (19), plasma membrane (15), mitochondria (13), Golgi (9), endoplasmic reticulum (9), peroxisome (2), vacuole (2), and extracellular space (9). In yeast, there were candidates in the endoplasmic reticulum (12), vacuole (14), bud neck (11), mitochondria (10), the vacuolar membrane (3), and multiple non-nuclear organelles (26). Overall, 90 Arabidopsis and 101 yeast non-nuclear genes with ADs are non-cytosolic and are targeted to a specific organellar compartment, raising the question of whether they can be relocated to the nucleus to modulate transcription.

To gauge the role of all non-nuclear candidate genes, we studied their GO terms in both species. In yeast, GO terms were unrelated to transcription and included both metabolic terms, such as "lipid metabolic process," and signaling terms, such as "response to chemical." Many GO terms were linked to architecture of the cell such as "meiotic/mitotic cell cycle," "cytoskeleton organization," and "organelle fission" (Figure 3E; Table S7). In Arabidopsis, the two most abundant molecular function terms were linked to "protein binding" and "general binding," followed by "catalytic activity" and "unknown molecular function" (Figure 3F; Table S8). The inclusion toward GO terms linked to binding interactions in Arabidopsis raises the question of whether AD-like peptides have been co-opted in other organelles to facilitate protein-protein interactions outside of the nucleus. In principle, isolating proteins in topologically separate compartments enables the reuse of the same protein-protein interactions without crosstalk. This highlights the diverse functionalities of proteins with AD-like sections in both species and suggests that AD-like peptides may perform other protein-protein interactions outside the nucleus.

A set of universal eukaryotic ADs

Our library gave us the unique opportunity to validate the transferability of yeast-derived pADs into plants and establish potential universal activators that function in phylogenetically divergent eukaryotes. We characterized 51 of the strongest pADs in the library—31 derived from Arabidopsis and 20 from yeast—in the plant *Nicotiana benthamiana* using an agroinfiltration-mediated transient expression system that we previously established (Table S9).[9] In this system, each tile is fused to the yeast GAL4-DBD and localized to a synthetic minimal promoter with five concatenated GAL4 binding sites to drive GFP (Figure 4A), modulating GFP expression. A constitutively expressed dsRed is used to normalize the signal. To benchmark the potency of the tiles, we also tested the strong activator VP16 and VPR (a fusion of the three strong activators VP64, p65, and RTA) as GAL4 fusions. Of 51 pADs tested, we found 45 (88.2%) to significantly increase GFP expression over the reporter only control, five were stronger than VP16, and four were statistically indistinguishable from VPR (Mann-Whitney-U test, $p < 0.05$, $n = 48$) (Figure 4A; Table S10). Notably, our tested pADs span the entire dynamic range of possible activities *in planta*, from no observed to very strong activity, highlighting the importance of remeasuring pADs *in planta*. Overall, we discovered short, universal pADs from non-TF proteins that perform similarly to longer state-of-the-art ADs, such as VPR, that are readily available for further eukaryotic engineering efforts. Shorter and stronger ADs are in demand for space-limited synthetic biology engineering strategies, such as viral delivery methods.

Our agnostic approach to mapping pADs in non-TF genes allowed us to mine strong ADs from proteins that have not previously been associated with transcription, and we show that these ADs function in plants. As an example, we localized and validated pADs in known plant coactivators, namely, one CCT-motif containing protein from Arabidopsis (AT1G04500), coactivator LNK3 (AT3G12320), and SAGA complex subunit 2A (AT2G19390), which showed activity similar to the VP16 control. Furthermore, the strongest pAD in plants was derived from an Arabidopsis uncharacterized 2Fe-2S ferredoxin-like superfamily protein (AT1G50780). The second strongest pAD was derived from a hypothetical protein (AT2G29920). Overall, pADs in non-TF proteins involved in transcription and in non-nuclear proteins function *in planta*.

Eukaryotic TFs utilize conserved general transcription machinery (e.g., Mediator) to facilitate transcription, making new TF parts a potential resource to develop tools for the control of transcription across eukaryotes. For our plant experiments,

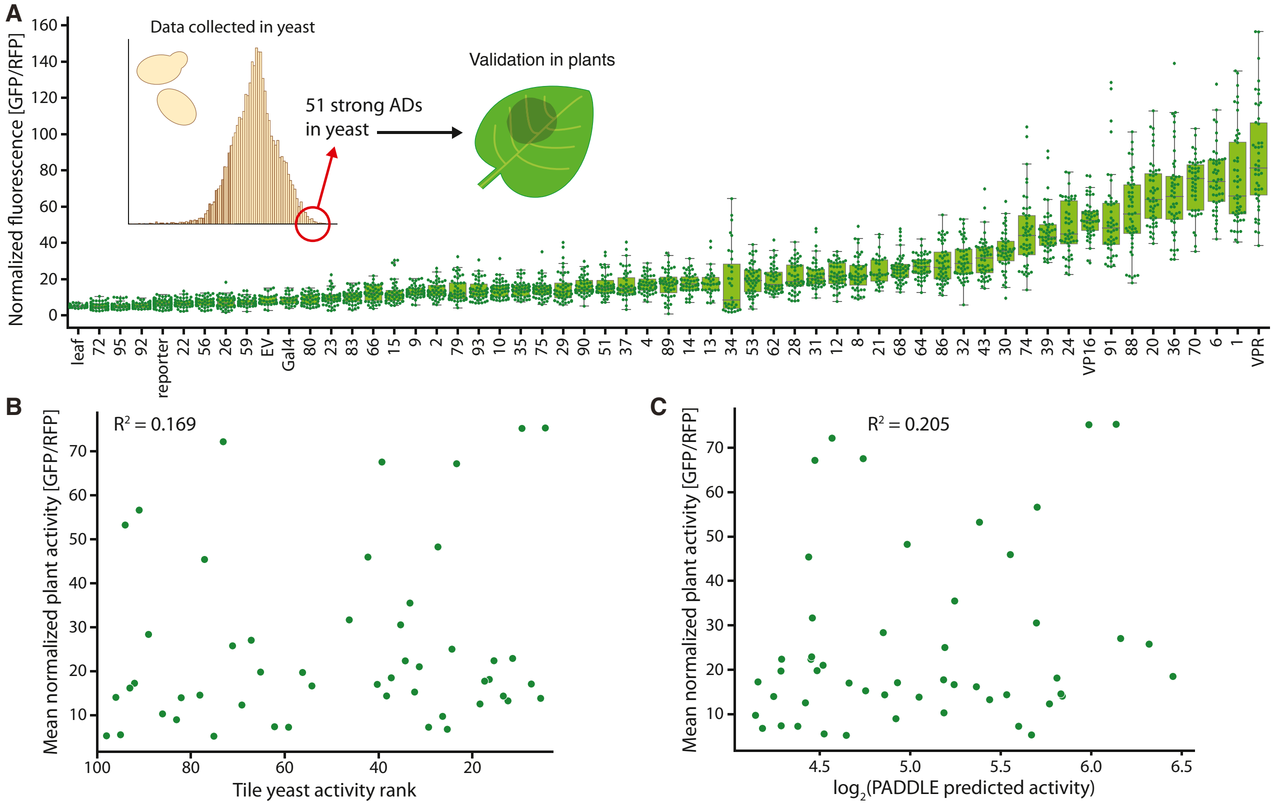

Figure 4. Cross-validation of strong pADs yields candidates with similar activities to gold standard activators

(A) Activity of 51 of the strongest tiles measured in yeast in *N. benthamiana*. Each data point represents the expression of GFP normalized by a constitutively expressed dsRed. VP16 and VPR serve as positive AD controls. EV, empty vector control; Gal4, Gal4-DBD without a tethered AD.

(B) Mean normalized *in planta* activity of pADs in (A) sorted by activity strength in yeast.

(C) Mean normalized *in planta* activity of pADs in (A) relative to predicted activity.

we chose only the strongest tiles from our yeast experiments, expecting that, if they utilize general conserved transcription machinery, activities in plants should be similarly strong. However, we observed a poor correlation between the rank order of yeast pAD vs. the rank order in plants (Figure 4B). PADDLE predictions correlated worse with observed pAD activity in plants than in yeast (Figure 4C). Our results suggest that although PADDLE can localize ADs in parent genes in both plant and yeast proteins with 89.0% accuracy, there are mechanistic features of plant transcription not fully captured by PADDLE that prevent accurate prediction of AD strength. We conclude that future work is needed to generate independent plant AD datasets to train models that can predict the strength of plant ADs with higher accuracy to enable the full potential of mining plant proteomes.

DISCUSSION

High-throughput studies have largely focused on ADs found in TFs and protein classes known to be involved in transcription, which has partly biased our understanding of the biological role of such peptides. By mining proteomes for pADs from non-TF genes and demonstrating their activity in yeast and plants, we reveal that pADs frequently occur across entire proteomes and outside the nucleus, going beyond the canonical description of ADs in TFs that mediate nuclear transcription. Studying nuclear non-TF pADs from the well-studied model yeast expands our understanding of which genes contain AD-like peptides and where they are localized. We found a direct correlation between nuclear genes containing pADs and their likelihood to function as coactivators in yeast. Our dataset provided the motivation to extrapolate this observation to plants, and we annotated over 200 putative coactivators that may be involved in many facets of plant transcriptional regulation. Due to the throughput limitations of our experimental setup, we focused on the strongest 18,000 tiles from both species, leaving a far larger sequence space of medium or weak pADs unstudied. Future work will focus on experimentally validating larger sets of predicted pADs in both species and help understand how frequent pADs occur throughout proteomes.

The recent establishment of large experimental datasets of ADs in yeast has led to the development of multiple neural networks that attempt to localize and predict the activity of ADs from protein sequences.[5,14] In this study we utilized one of these models PADDLE to build and test our library.[14] We found that PADDLE can correctly localize pADs throughout entire proteomes; however, the capabilities of PADDLE to predict the quantitative activity of pADs fell short in comparison with the high correlation value that was reported in the original study. We further show that plant tiles that functioned as strong ADs in yeast, indeed, largely functioned in plants but with divergent degrees of activity. This discrepancy indicates that, although general eukaryotic mechanisms for the regulation of transcription between plants and yeast are conserved, there are intricacies in plants that models trained on yeast data cannot resolve. These intricacies parallel leucine-dependent ADs from metazoans that are missing in yeast.[15] It further highlights that the flexible positional and compositional sequence requirements of ADs need to be explored further in their native context.

Recently, there has been significant interest in utilizing genome engineering approaches in non-model eukaryotes that have traditionally been recalcitrant to genetic studies. Such efforts are constrained by the dearth of characterized genetic parts that reliably function across phylogenetically diverse eukaryotes, restraining the application of high-throughput genetic screening methods, such as CRISPR activation (CRIPSRa). Our results highlight how computational models for predicting ADs are still in their infancy. Sequences with very similar AA composition can largely differ in activity based on AA arrangement. Future work is needed to further unravel this sequence grammar to better understand AD function and guide the construction of next-generation predictive models. In the long term, we anticipate building models for species-specific activity. Such knowledge will help establish design principles for ADs, which will ease the implementation of new synthetic biology tools and genome-scale activation assays, such as CRISPRa screens in plants and other non-model organisms.

At their core, ADs enable protein-protein interactions with transcriptional machinery to facilitate transcription. We observed an abundance of peptides with AD-like properties throughout entire proteomes in two distantly related eukaryotes, suggesting three possible roles of these peptides. (1) Their parent proteins moonlight into the nucleus to facilitate transcription. (2) The broad sequence space that allows AD activity has led to statistical occurrence of peptides with AD-like properties. (3) ADs are an instance of a larger class of protein-protein interaction domains that perform many functions. Option one is supported by anecdotal evidence of proteins normally localized to the cytosol and mitochondria being imported into the nucleus to promote transcription during signaling cascades.[10,11,56,57] Option two is supported by large AD screens that show that up to 1% of random sequences have AD activity when localized to promoters.[58,59] We believe that option three entails the most logic. ADs do not rely on structure to facilitate binding, they form multiple weak interactions with coactivators.[60,61] We believe that compartmentalization allows the "recycling" of AD-like interactions in different organelles. Of note, the GOs predicted from physically separated plant organelles (e.g., mitochondria and chloroplast) are highly enriched in proteins involved in protein binding, supporting that these interactions could be used for different functions independent of transcription. We speculate that the versatile nature of ADs extends their role beyond nuclear transcription and blurs the distinction of ADs as a feature unique to TFs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL DETAILS
  - ○ *N. benthamiana* growth conditions
  - ○ Bacterial and yeast growth conditions
- METHOD DETAILS
  - ○ PADDLE prediction of every 53 amino acid tile in the proteome of A. thaliana and S. cerevisiae
  - ○ Plasmid library construction
  - ○ Yeast Library Construction and measurement
  - ○ Fluorescence Activated Cell Sorting and library preparation
  - ○ Plant experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Analysis of barcodes and inferring activity
  - ○ Data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2024.05.007.

ACKNOWLEDGMENTS

This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. M.V.S. is supported by NIH grant R35GM150813, Simons Foundation grant 1018719, and NSF grant 2112057. M.V.S. is a Chan Zuckerberg Biohub—San Francisco investigator.

AUTHOR CONTRIBUTIONS

Project design, N.F.C.H., M.V.S., and P.M.S.; experimental work, N.F.C.H., K.M., J.S., and M.V.S.; analysis, N.F.C.H., J.S., M.V.S., and P.M.S.; visualization, N.F.C.H.; financial support, M.V.S. and P.M.S.; supervision, M.V.S. and P.M.S.; writing—original draft, N.F.C.H., M.V.S., and P.M.S.; writing—review & editing, N.F.C.H., M.V.S., and P.M.S.

DECLARATION OF INTERESTS

P.M.S. and N.F.C.H. have a patent pending for the library of synthetic TFs and all derived parts under US patent application number 63/579,836.

Received: November 7, 2023
Revised: April 26, 2024
Accepted: May 22, 2024
Published: June 11, 2024

REFERENCES

1. Latchman, D.S. (2007). Eukaryotic Transcription Factors, Fifth Edition (Elsevier).
2. Tycko, J., DelRosso, N., Hess, G.T., Aradhana, Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K., et al. (2020). High-Throughput

Discovery and Characterization of Human Transcriptional Effectors. Cell *183*, 2020–2035.e16.

3. Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147–151.

4. Arnold, C.D., Nemčko, F., Woodfin, A.R., Wienerroither, S., Vlasova, A., Schleiffer, A., Pagani, M., Rath, M., and Stark, A. (2018). A high-throughput method to identify trans-activation domains within transcription factor sequences. EMBO J. *37*, e98896. https://doi.org/10.15252/embj.201798896.

5. Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2020). A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. Mol. Cell *78*, 890–902.e6.

6. Staller, M.V., Ramirez, E., Kotha, S.R., Holehouse, A.S., Pappu, R.V., and Cohen, B.A. (2022). Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. Cell Syst. *13*, 334–345.e5.

7. DelRosso, N., Tycko, J., Suzuki, P., Andrews, C., Aradhana, Mukund, A., Liongson, I., Ludwig, C., Spees, K., Fordyce, P., et al. (2023). Large-scale mapping and mutagenesis of human transcriptional effector domains. Nature *616*, 365–372.

8. Ravarani, C.N., Erkina, T.Y., De Baets, G., Dudman, D.C., Erkine, A.M., and Babu, M.M. (2018). High-throughput discovery of functional disordered regions: investigation of transactivation domains. Mol. Syst. Biol. *14*, e8190.

9. Hummel, N.F.C., Zhou, A., Li, B., Markel, K., Ornelas, I.J., and Shih, P.M. (2023). The trans-regulatory landscape of gene networks in plants. Cell Syst. *14*, 501–511.e4.

10. Jarriault, S., Brou, C., Logeat, F., Schroeter, E.H., Kopan, R., and Israel, A. (1995). Signalling downstream of activated mammalian Notch. Nature *377*, 355–358.

11. Behrens, J., von Kries, J.P., Kühl, M., Bruhn, L., Wedlich, D., Grosschedl, R., and Birchmeier, W. (1996). Functional interaction of beta-catenin with the transcription factor LEF-1. Nature *382*, 638–642.

12. Coates, J.C., Laplaze, L., and Haseloff, J. (2006). Armadillo-related proteins promote lateral root development in Arabidopsis. Proc. Natl. Acad. Sci. USA *103*, 1621–1626.

13. van de Wetering, M., Cavallo, R., Dooijes, D., van Beest, M., van Es, J., Loureiro, J., Ypma, A., Hursh, D., Jones, T., Bejsovec, A., et al. (1997). Armadillo coactivates transcription driven by the product of the Drosophila segment polarity gene dTCF. Cell *88*, 789–799.

14. Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman Aiden, E., Dror, R.O., and Kornberg, R.D. (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. eLife *10*, e68068. https://doi.org/10.7554/eLife.68068.

15. Kotha, S.R., and Staller, M.V. (2023). Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence. Genetics *225*, iyad131. https://doi.org/10.1093/genetics/iyad131.

16. Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. Cell Syst. *6*, 444–455.e6.

17. Glenn, D.J., and Maurer, R.A. (1999). MRG1 binds to the LIM domain of Lhx2 and may function as a coactivator to stimulate glycoprotein hormone alpha-subunit gene expression. J. Biol. Chem. *274*, 36159–36167.

18. Ogryzko, V.V., Schiltz, R.L., Russanova, V., Howard, B.H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell *87*, 953–959.

19. Malik, S., Guermah, M., and Roeder, R.G. (1998). A dynamic model for PC4 coactivator function in RNA polymerase II transcription. Proc. Natl. Acad. Sci. USA *95*, 2192–2197.

20. Liu, Z., and Myers, L.C. (2015). Fungal mediator tail subunits contain classical transcriptional activation domains. Mol. Cell. Biol. *35*, 1363–1375.

21. O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. Cell *165*, 1280–1292.

22. Gaudinier, A., Rodriguez-Medina, J., Zhang, L., Olson, A., Liseron-Monfils, C., Bågman, A.M., Foret, J., Abbitt, S., Tang, M., Li, B., et al. (2018). Transcriptional regulation of nitrogen-associated metabolism and growth. Nature *563*, 259–264.

23. Monteiro, P.T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., Galocha, M., Godinho, C.P., Martins, L.C., Bourbon, N., et al. (2020). YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. Nucleic Acids Res. *48*, D642–D649.

24. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2020). PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. *48*, D1104–D1113.

25. Natter, K., Leitner, P., Faschinger, A., Wolinski, H., McCraith, S., Fields, S., and Kohlwein, S.D. (2005). The spatial organization of lipid synthesis in the yeast Saccharomyces cerevisiae derived from large scale green fluorescent protein tagging and high resolution microscopy. Mol. Cell. Proteomics *4*, 662–672.

26. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

27. Hooper, C., Millar, H., Black, K., Castleden, I., Aryamanesh, N., and Grasso, S. (2022). Subcellular Localisation Database for Arabidopsis Proteins Version 5 (The University of Western Australia). https://doi.org/10.26182/8dht-4017.

28. McIsaac, R.S., Gibney, P.A., Chandran, S.S., Benjamin, K.R., and Botstein, D. (2014). Synthetic biology tools for programming gene expression without nutritional perturbations in Saccharomyces cerevisiae. Nucleic Acids Res. *42*, e48.

29. Broyles, B.K., Gutierrez, A.T., Maris, T.P., Coil, D.A., Wagner, T.M., Wang, X., Kihara, D., Class, C.A., and Erkine, A.M. (2021). Activation of gene expression by detergent-like protein domains. iScience *24*, 103017.

30. Erkine, A.M. (2018). "nonlinear" biochemistry of nucleosome detergents. Trends Biochem. Sci. *43*, 951–959.

31. Emenecker, R.J., Griffith, D., and Holehouse, A.S. (2022). Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.06.494887.

32. Keung, A.J., Bashor, C.J., Kiriakov, S., Collins, J.J., and Khalil, A.S. (2014). Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. Cell *158*, 110–120.

33. Cherel, I., and Thuriaux, P. (1995). The IFH1 gene product interacts with a fork head protein in Saccharomyces cerevisiae. Yeast *11*, 261–270.

34. Myers, L.C., Gustafsson, C.M., Bushnell, D.A., Lui, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R.D. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. Genes Dev. *12*, 45–54.

35. Sterner, D.E., Grant, P.A., Roberts, S.M., Duggan, L.J., Belotserkovskaya, R., Pacella, L.A., Winston, F., Workman, J.L., and Berger, S.L. (1999). Functional organization of the yeast SAGA complex: distinct components involved in structural integrity, nucleosome acetylation, and TATA-binding protein interaction. Mol. Cell. Biol. *19*, 86–98.

36. Bourbon, H.-M., Aguilera, A., Ansari, A.Z., Asturias, F.J., Berk, A.J., Bjorklund, S., Blackwell, T.K., Borggrefe, T., Carey, M., Carlson, M., et al. (2004). A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II. Mol. Cell *14*, 553–557.

37. Tollis, S., Singh, J., Palou, R., Thattikota, Y., Ghazal, G., Coulombe-Huntington, J., Tang, X., Moore, S., Blake, D., Bonneil, E., et al. (2022). The microprotein Nrs1 rewires the G1/S transcriptional machinery during nitrogen limitation in budding yeast. PLoS Biol. *20*, e3001548.

38. Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates, J.R., and Grant, P.A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature *433*, 434–438.

39. Cairns, B.R., Lorch, Y., Li, Y., Zhang, M., Lacomis, L., Erdjument-Bromage, H., Tempst, P., Du, J., Laurent, B., and Kornberg, R.D. (1996). RSC, an essential, abundant chromatin-remodeling complex. Cell *87*, 1249–1260.

40. Czaja, W., Mao, P., and Smerdon, M.J. (2014). Chromatin remodelling complex RSC promotes base excision repair in chromatin of Saccharomyces cerevisiae. DNA Repair (Amst) *16*, 35–43.

41. Vannier, D., Damay, P., and Shore, D. (2001). A role for Sds3p, a component of the Rpd3p/Sin3p deacetylase complex, in maintaining cellular integrity in Saccharomyces cerevisiae. Mol. Genet. Genomics *265*, 560–568.

42. Papamichos-Chronakis, M., Petrakis, T., Ktistaki, E., Topalidou, I., and Tzamarias, D. (2002). Cti6, a PHD domain protein, bridges the Cyc8-Tup1 corepressor and the SAGA coactivator to overcome repression at GAL1. Mol. Cell *9*, 1297–1305.

43. Shen, X., Mizuguchi, G., Hamiche, A., and Wu, C. (2000). A chromatin remodelling complex involved in transcription and DNA processing. Nature *406*, 541–544.

44. Lanker, S., Müller, P.P., Altmann, M., Goyer, C., Sonenberg, N., and Trachsel, H. (1992). Interactions of the eIF-4F subunits in the yeast Saccharomyces cerevisiae. J. Biol. Chem. *267*, 21167–21171.

45. Matangkasombut, O., Buratowski, R.M., Swilling, N.W., and Buratowski, S. (2000). Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. Genes Dev. *14*, 951–962.

46. Tora, L. (2002). A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. Genes Dev. *16*, 673–675.

47. Henry, N.L., Campbell, A.M., Feaver, W.J., Poon, D., Weil, P.A., and Kornberg, R.D. (1994). TFIIF-TAF-RNA polymerase II connection. Genes Dev. *8*, 2868–2878.

48. Autran, D., Jonak, C., Belcram, K., Beemster, G.T.S., Kronenberger, J., Grandjean, O., Inzé, D., and Traas, J. (2002). Cell numbers and leaf development in Arabidopsis: a functional analysis of the STRUWWELPETER gene. EMBO J. *21*, 6036–6049.

49. Bäckström, S., Elfving, N., Nilsson, R., Wingsle, G., and Björklund, S. (2007). Purification of a plant mediator from Arabidopsis thaliana identifies PFT1 as the Med25 subunit. Mol. Cell *26*, 717–729.

50. Xie, Q., Wang, P., Liu, X., Yuan, L., Wang, L., Zhang, C., Li, Y., Xing, H., Zhi, L., Yue, Z., et al. (2014). LNK1 and LNK2 are transcriptional coactivators in the Arabidopsis circadian oscillator. Plant Cell *26*, 2843–2857.

51. Kidokoro, S., Konoura, I., Soma, F., Suzuki, T., Miyakawa, T., Tanokura, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2023). Clock-regulated coactivators selectively control gene expression in response to different temperature stress conditions in Arabidopsis. Proc. Natl. Acad. Sci. USA *120*, e2216183120.

52. Gillmor, C.S., Park, M.Y., Smith, M.R., Pepitone, R., Kerstetter, R.A., and Poethig, R.S. (2010). The MED12-MED13 module of Mediator regulates the timing of embryo patterning in Arabidopsis. Development *137*, 113–122.

53. Wu, C.-J., Liu, Z.-Z., Wei, L., Zhou, J.-X., Cai, X.-W., Su, Y.-N., Li, L., Chen, S., and He, X.-J. (2021). Three functionally redundant plant-specific paralogs are core subunits of the SAGA histone acetyltransferase complex in Arabidopsis. Mol. Plant *14*, 1071–1087.

54. Lee, K., and Seo, P.J. (2018). The HAF2 protein shapes histone acetylation levels of PRR5 and LUX loci in Arabidopsis. Planta *248*, 513–518.

55. Lai, Z., Li, Y., Wang, F., Cheng, Y., Fan, B., Yu, J.-Q., and Chen, Z. (2011). Arabidopsis sigma factor binding proteins are activators of the WRKY33 transcription factor in plant defense. Plant Cell *23*, 3824–3841.

56. Monaghan, R.M., and Whitmarsh, A.J. (2015). Mitochondrial proteins moonlighting in the nucleus. Trends Biochem. Sci. *40*, 728–735.

57. Erkine, A.M., and Gross, D.S. (2003). Dynamic chromatin alterations triggered by natural and synthetic activation domains. J. Biol. Chem. *278*, 7755–7764.

58. Abedi, M., Caponigro, G., Shen, J., Hansen, S., Sandrock, T., and Kamb, A. (2001). Transcriptional transactivation by selected short random peptides attached to lexA-GFP fusion proteins. BMC Mol. Biol. *2*, 10.

59. Ma, J., and Ptashne, M. (1987). A new class of yeast transcriptional activators. Cell *51*, 113–119.

60. Tuttle, L.M., Pacheco, D., Warfield, L., Luo, J., Ranish, J., Hahn, S., and Klevit, R.E. (2018). Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. Cell Rep. *22*, 3251–3264.

61. Warfield, L., Tuttle, L.M., Pacheco, D., Klevit, R.E., and Hahn, S. (2014). A sequence-specific transcription activator motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. Proc. Natl. Acad. Sci. USA *111*, E3506–E3513.

62. Moffat, L., and Jones, D.T. (2021). Increasing the Accuracy of Single Sequence Prediction Methods Using a Deep Semi-Supervised Learning Framework. Bioinformatics *37*, 3744–3751.

63. Erdős, G., Pajkos, M., and Dosztányi, Z. (2021). IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. *49*, W297–W303.

64. Emenecker, R.J., Griffith, D., and Holehouse, A.S. (2021). Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Biophys. J. *120*, 4312–4319.

65. Daniel Gietz, R., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol. *350*, 87–96.

66. Amberg, D.C., and Strathern, J.N. (2005). Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual (CSHL Press).

67. Belcher, M.S., Vuu, K.M., Zhou, A., Mansoori, N., Agosto Ramos, A., Thompson, M.G., Scheller, H.V., Loqué, D., and Shih, P.M. (2020). Design of orthogonal regulatory systems for modulating gene expression in plants. Nat. Chem. Biol. *16*, 857–865.

68. Piskacek, S., Gregor, M., Nemethova, M., Grabner, M., Kovarik, P., and Piskacek, M. (2007). Nine-amino-acid transactivation domain: establishment and prediction utilities. Genomics *89*, 756–768.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Agrobacterium fabrum strain GV3101	Joint BioEnergy Institute	N/A
Chemicals, peptides, and recombinant proteins		
ß-estradiol	Sigma-Aldrich	50-28-2
Acetosyringone	Sigma-Aldrich	D134406
Critical commercial assays		
Monarch PCR and DNA kit	NEB	T1030S
Zymo YeaSTAR kit	Zymo Research	#D2002
Deposited data		
Illumina sequencing of barcoded library	GEO	GSE247147
PADDLE predictions and code for visualization	Zenodo	Zenodo: 11151016
Experimental models: Organisms/strains		
DHY211	Staller et al.[6]	Staller et al.[6]
Saccharomyces cerevisae strain FY5	Staller et al.[6]	Staller et al.[6]
Oligonucleotides		
Table S12	N/A	N/A
Recombinant DNA		
pms6370	https://registry.jbei.org	JBx_082980
pMVS219	RRID	https://www.addgene.org/99049/).Addgene_99049
Software and algorithms		
Python v3.9.5	Python.org	N/A
S4Pred	Moffat and Jones[62]	https://bio.tools/psipred
IUPRED	Erdős et al.[63]	https://iupred3.elte.hu/
PADDLE	Sanborn et al.[14]	Sanborn et al.[14]
Metapredict v2	Emenecker et al.[64]	Emenecker et al.[64]

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Patrick M. Shih (pmshih@berkeley.edu).

Materials availability

All plasmid materials and bacterial strains will be made available through the Inventory of Composable Elements (https://registry.jbei.org/). Sequences and raw data are available as supplemental information.

Data and code availability

- The Illumina sequencing data have been deposited under GEO accession GSE247147 and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. All flow cytometry data reported in this paper will be shared by the lead contact upon request.
- All code used for data analysis and associated data files are available on Zenodo: 11151016
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL DETAILS

N. benthamiana growth conditions

Wild type *N. benthamiana* plants were obtained from the in-house seed bank at the Joint BioEnergy institute. *N. benthamiana* plants were grown in SunGro Horticulture Professional Growing Mix #1 for four weeks in Percival-Scientific growth chambers at

25°C in 16/8-hour light/dark cycles and 60% humidity at ~100 μmol of photons m^{-2} s^{-1}. Plants were fertilized two weeks after germination with MiracleGro®. Post infiltration N. benthamiana plants were maintained in the same growth conditions.

Bacterial and yeast growth conditions

Agrobacterium fabrum strain GV3101 was obtained from the Inventory of Composable Elements (ICE) at the Joint BioEnergy Institute. Generated binary vectors were transformed into *A. fabrum* strain GV3101 and selected on LB plates (50 μg/mL kanamycin, 30 μg/mL gentamicin, and 100 μg/mL rifampicin). Selected transformants were inoculated in liquid LB media with the same antibiotic concentrations. Yeast strains were grown in synthetic complete glucose media with G418 (200 μg/ml) and/or NAT (100 μg/ml) at 30°C.

METHOD DETAILS

PADDLE prediction of every 53 amino acid tile in the proteome of A. thaliana and S. cerevisiae

We predicted the AD activity of all proteins of the reference proteome of *A. thaliana* (Colombia ecotype) and *S. Cerevisiae* (strain S288C) which we obtained from TAIR (Araport11) and SGD (S288C Genome release 64-3-1), respectively. Both proteomes with associated predictions are available in Data S1 and can be loaded using Load_predictions_SI_data1.ipynb. We predicted the secondary structure of every full-length protein using S4PRED and their structural disorder with IUPRED3 (long and short mode).[62,63] We then tiled the protein sequences and structural predictions into consecutive 53 amino acid tiles and predicted their AD activity using the PADDLE API for Python as described.[14] We ran all predictions in Python v3.9.5 with associated APIs and our pipeline is available in the supplemental data package. As we wanted to focus on tiles from non-TF genes, we utilized the TF databases PlantTFDB v5.0 and Yeastract+ to filter out any tiles derived from TFs. We selected tiles from genes that achieved a PADDLE predicted activation >30, yielding 12,000 *A. thaliana* tiles and 6,000 *S. cerevisiae* tiles with a dynamic range of PADDLE predicted activation strength between 17 and 138.

Plasmid library construction

The library of both Arabidopsis and Saccharomyces ADs were generated by mapping the tiles back to their native DNA sequence in the respective reference genomes, retrieved from TAIR and SGD (all sequences in Table S1). 18,000 unique DNA oligos coding for 53 amino acid long putative activators were synthesized in one oligo pool by Twist Bioscience. Each oligo contains a 24 bp upstream primer (GCGGGCTCTACTTCATCGGCTAGC), 159 bp encoding the activator candidate, a 21 bp primer (TGATAACTAG CTGAGGGCCCG) with four stop codons in 3 frames and the ApaI site. Specifically, we used 75 ng of template and 12 rounds of PCR in 16 parallel 50 μL reactions using primers LC3.P1_Lib_Hom_up_5', which adds homology arms and YL_randBCs_R1_3' which adds random 11 nt barcodes and downstream homology arms (NEB Q5 polymerase Tm=70C). The PCR product was pooled and cleaned using the Monarch PCR and DNA kit, followed by product visualization on a 1% Agarose gel. Vector pMVS219 was linearized using NheI, AscI and PacI and used for library assembly. The assembly was performed using 100 ng of linearized backbone and 7.5 ng of PCR product using NEB Hifi DNA Assembly Master Mix in 8, 10 μL reactions. Assemblies were electroporated into DH5β cells (NEB C3020K), and we recovered >1,000,000 colonies.

The plasmid sequence of the library assembly vector pMVS219 is available on addgene (https://www.addgene.org/99049/).

Yeast Library Construction and measurement

To ensure singular constructs per cell, we introduced our library into the URA3 locus of strain DHY211 (*MATa, MKT1(30G,) RME1(INS-308A) TAO3(1493Q), CAT5(91M) MIP1(661T), SAL1*$^{+}$ *HAP1*$^{+}$). Employing the established yeast transformation method,[65] we subjected the transformation to 30 minutes at 30 °C followed by 60 minutes at 42 °C. To minimize potential PCR errors, we performed SalI and EcoRI digestion on the plasmid library, releasing the section encompassing the ACT1 promoter, the synthetic TF, and the KANMX marker. Simultaneously, PacI digestion was conducted to cleave plasmids devoid of an activation domain variant and barcode insert, thereby reducing the occurrence of transformants with inactive TFs. Directed integration into the URA3 locus was guided by 500 bp upstream homology spanning the URA3 and ACT1 promoters, along with a corresponding 500 bp downstream homology region spanning the TEF and URA3 terminators. These regions were PCR amplified from pMVS 295 (Strader 6161) and pMVS 296 (Strader 6768), a generous gift from Nick Moffy and Lucia Starder. Transformation utilized a molar ratio of 1:3 for linearized library to homology arms, with 28 μmol of linearized library per reaction. The transformed library was plated on YPD, followed by an overnight incubation at 30°C, and subsequent replica-plating onto freshly prepared SC G418 plates. Employing this process across 80 transformation reactions yielded an estimated >1,000,000 individual colonies. Subsequently, the transformants were collectively mated with an FY5 strain containing the reporter integrated into the uncertain ORF, YBR032w. Diploids were selected on YPD with G418 (200 μg/ml) and NAT (100 μg/ml) (strain MY436 YBR032w::P3_GFP NAT S288C), resulting in prototrophic diploids. These 110,000 yeast transformants were mated in batches, and prior to the final experiment, batches were pooled, and multiple aliquots were frozen.

Fluorescence Activated Cell Sorting and library preparation

Each sorting experiment was preceded by thawing a frozen glycerol stock, followed by overnight growth in SC+G418+NAT. Cultures were cultivated in synthetic complete (SC) dextrose media at 30 °C.[66] Prior to fluorescence-activated cell sorting (FACS), overnight cultures were diluted (1:5) into SC+ 1 μM ß-estradiol and incubated for 3.5-4 hours at 30 °C. We sorted the yeast library on a Aria-fusion cell sorter at the UC Berkeley Flow Cytometry core facility. We used the parent yeast strain with the reporter and a TF lacking an activation domain as a negative control to determine autofluorescence and baseline mCherry levels. We sorted 1 million cells of the synthetic TF library into 8 bins with each bin roughly covering 11 % of the entire observable population in the GFP channel. To test reproducibility, we sorted another 500,000 cells from each bin.

Sorted cells were grown overnight in SC at 30 °C and gDNA was extracted with the Zymo YeaSTAR (#D2002) kit. Barcodes were amplified by PCR (CP21.P14: TCCTCATCCTCTCCCACATC, CP17.P12: GGACGAGGCAAGCTAAACAG, NEB Q5 for 20 cycles, Tm 67 °C). We added phasing nucleotides as well as overhangs for indexing primers using primer mixtures SL5.F[1-4] and SL5.R[1-4] (NEB Q5 for 20 cycles, Tm 62 °C). We finally added dual indexing primers using the i5 and i7 system from Illumina (NEB Q5 for 20 cycles, Tm 65 °C). We then performed a bead cleanup. We sequenced the library on an Illumina Novaseq 6000 system with 2x150 bp paired end reads.

We assessed library performance against known ADs from GCN4 and VP16 on a BD Accuri^{TM} C6 flow cytometer (BD Biosciences). All strains were grown in SC+G418+NAT at 30 overnight and diluted (1:5) into SC+/- 1 μM ß-estradiol and incubated for 3.5-4 hours at 30 °C. Samples were washed with cold 1x PBS (137 mmol NaCl, 2.7 mM KCl, 1.8 mM KH_2PO_4, 10 mM Na_2HPO_4) once before measurement. Per sample 100,000 events were recorded and analyzed using the Python fcsparser package.

Plant experiments

Generated binary vectors were transformed into *Agrobacterium fabrum* strain GV3101. Selected transformants were inoculated in liquid media with appropriate selection the night before the experiment. *A. fabrum* strains were grown until OD_{600} between 0.8 and 1.2 and were mixed equally (final OD_{600} = 0.5 for each strain) with the strain harboring the assay reporter construct to a final OD_{600} = 1.0. Cultures were centrifuged for 10 min at 4000 g and resuspended in infiltration buffer (10 mM $MgCl_2$, 10 mM MES, and 200 μM acetosyringone, pH 5.6). Cultures were induced for 2 h at room temperature on a rocking shaker. Leaves 6 and 7 of 4-week-old *N. benthamiana* plants were syringe infiltrated with the *A. fabrum* suspensions. Post infiltration *N. benthamiana* plants were maintained in the same growth conditions as described above. Leaves were harvested three days post infiltration and 16 leaf disks from two leaves and 3 plants total per construct were collected. The leaf disks were floated on 200 μL of water in 96 well microtiter plates and GFP (Ex. λ = 488 nm, Em. λ = 520 nm) and RFP (Ex. λ = 532 nm, Em. λ = 580 nm) fluorescence measured using a Synergy 4 microplate reader (Bio-tek). The reporter construct for the screen was pms6370 containing GFP and dsRed expression cassettes. GFP expression was driven by a fusion of five previously characterized GAL4 binding sites with the core WUSCHEL promoter.[67] GFP expression was normalized using dsRed driven by the constitutive MAS promoter on the same plasmid.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of barcodes and inferring activity

After demultiplexing samples, we kept only the reads that contained a perfect match to a designed tile. For each set of 8 sorted samples, we performed two normalizations. We first normalized the reads by the total number of reads in each bin. Then, for each designed tile, we normalized across the 8 bins to calculate a relative abundance. We then converted relative abundances to an activity score for each tile by taking the dot product of the relative abundance with the median fluorescence value of each bin **(**Table S11**)**. This computation is a weighted average. Tiles with less than 10 reads were not included in the final dataset. Later, post hoc analysis suggested that tiles with at least 1000 reads were well measured.

During plasmid library construction we added random barcodes to the designed tiles. To build a map linking designed tiles to barcodes, we combined all the sequencing data from the 16 sorted samples. We use this map to compare two modes of analysis. First, for the primary analysis used in the manuscript, we used only the tile sequences, effectively combining all the barcodes together and ignoring independent transformations. Second, we repeated the analysis for each AD+barcode combination, in effect measuring the activity of each independent transformant of each tile. The methods largely agreed (Figure S10). We determined statistical significance thresholds to infer the number of tiles with AD activity. We calculated the statistical difference between each individual tile with the mean of no-AD control using a one sample t-test and corrected p-values using Benjamini-Hochberg false discovery rates (5%).

Data analysis

We analyzed and visualized the data and underlying sequences of the tiles using the following APIs in Python v3.9.5: pandas, seaborn, matplotlib, numpy and scipy. All associated Jupyter Notebooks for producing all Figures are available on GitHub (doi: 10.5281/zenodo.11151016). We sorted the library by activity and split it into four equal sized quartiles with 4388 tiles per quartile. To gauge the composition of each tile in each quartile, we calculated the amino acid frequencies of all amino acids in each tile. For the amino acid density analysis, we applied a sliding window size 5 along every position of each tile, averaging the frequencies of amino acid occurrence of each amino acid for each quartile. We chose the amino acid window size to be 5 to not bias the analysis for short AD motifs like the 9aaTAD.[68] We then grouped the amino acid frequencies based on functional groups which we defined as follows: acidic (D, E) and hydrophobic (W, L, F, Y).

To gauge the disorder of tiles we utilized the disorder predictor MetapredictV2 which integrated the outcomes of multiple independent disorder predictors.[64] We predicted disorder of tiles when fused to the synthetic TF and in their endogenous context. Confidence intervals were calculated using the seaborn pointplot function.

Dipeptide frequencies were calculated by splitting tiles into quartiles as described before. We calculated the total occurrence of every amino acid in the respective quartiles. We measured the frequency of every dipeptide upstream and downstream, meaning if the first amino acid is an alanine, we accounted for all XA and AX dipeptides, where X is any of 20 twenty amino acids. The total occurrence of dipeptides was then normalized to the occurrence of the first amino acid in the quartile. We calculated dipeptide frequencies with spacers of up to 8 amino acids between amino acid one and two.

We provide figures of all parent genes with annotated location of tiles with their respective predicted and experimental activity as a resource in Data S3.

We utilized the single amino acid resolution of our tiling experiments to gauge the effect on AD activity when one C-terminal amino acid is gained, or one N-terminal amino acid is lost. We generated a subset of tiles only including tiles that had at least one consecutive neighboring tile, meaning a pair of identical tiles with only one amino acid difference in the C- and N-terminus. From this subset we calculated the change of AD activity between consecutive pairs of tiles and associated the lost and gained amino acid during the step. The analysis was performed for the entire library independent of whether a tile was defined as an AD or not.