Title: Multiplexed profiling of transcriptional regulators in plant cells

Authors: Simon Alamos[1,2,3,✣], Lucas Waldburger[1,2,4,5,✣], Amanda Dee[1,2,3], Lauren Owens[1,2,3], Rohan Rattan[4,6], Shirlyne Ong[3], Patrick M. Shih[1,2,3,7*]

Affiliations:
1 Joint BioEnergy Institute, 5885 Hollis Street, Emeryville, CA 94608, USA.
2 Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA.
3 Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.
4 Department of Bioengineering, University of California, Berkeley, CA 94720, USA.
5 Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94702, USA.
6 Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.
7 Innovative Genomics Institute, Berkeley, California, USA.
✣ These authors contributed equally.

*Correspondence should be addressed to Patrick M. Shih (pmshih@berkeley.edu).

Abstract

Transcriptional regulatory domains play key roles in plant growth, development, and environmental responses. Massively parallel reporter assays in yeast and animal cells have enabled large-scale studies of libraries of transcriptional regulators and other classes of proteins, however, similar technologies are limited in plants. Here, we developed ENTRAP-seq (Enrichment of Nuclear *Trans*-elements Reporter Assay in Plants), a high-throughput method that introduces protein-coding libraries into plant cells to drive a nuclear magnetic sorting-based reporter, enabling multiplexed measurement of regulatory activity from thousands of protein variants. Using ENTRAP-seq and machine learning, we screened 1,495 plant viruses and identified hundreds of novel transcriptional regulatory domains found in structural proteins and enzymes not previously associated with gene regulation. We further combined this approach with machine-guided design to provide the framework for semi-rational engineering of the regulatory activity of an *Arabidopsis* transcription factor. Our findings demonstrate how the development of scalable functional genomic platforms deployed *in planta* will enable the characterization of natural and synthetic coding diversity in plants.

Introduction

The accelerating pace of plant genomics has facilitated structure- and homology-based comparative approaches to infer the function of proteins regulating important traits; however, the functional annotation of this genetic variation has lagged behind. It remains very difficult if not impossible to predict gene function from sequence alone. Meeting this challenge is a stepping stone towards rationally designing more resilient crop varieties via gene editing[1]. This gap is particularly acute for transcriptional regulators, a class of proteins involved in virtually all aspects of plant biology[2]. Sequence-to-function annotation of transcriptional regulators is hindered by the high disorder and poor conservation of their regulatory domains[3–5], which limits the power of

comparative genomic algorithms to decipher how regulatory activity is encoded at the protein level. Moreover, mutations in regulatory domains are well tolerated or have subtle phenotypic effects, which has conspired against the use of classic random mutagenesis screens[6]. Directed mutagenesis studies are often guided by structural data and sequence conservation, but designing effective mutagenesis experiments becomes challenging when these constraints are absent and the search space is too large.

In recent years, techniques like STARR-seq[7] and DAP-seq[8] have enabled mapping the functional requirements of DNA sequences bound by plant transcription factors (TFs). However, large-scale characterization of TF regulatory activity remains a significant challenge. Given the limitations of traditional methods, the high-throughput study of regulatory domains has focused on exploring their protein sequence space using massively parallel reporter assays (MPRAs) in cultures of animal or microbial cells[9]. In this approach, TF libraries are transformed into cultured cells where they drive the expression of a reporter gene. Cells are then sorted according to the level of reporter expression and next-generation sequencing is used to determine the abundance of each transgene in each sorted bin, a proxy for their activity[10]. These assays empirically characterize transcriptional regulators at a genomic scale and yield the breadth of data necessary to train predictive models of the location and function of TF regulatory domains[11–13]. Such models provide a way to perform inference on protein sequences that are too costly or time-consuming to characterize experimentally[14].

Over the last decade, reporter-based MPRAs of protein libraries in *S. cerevisiae* (yeast) and animal cells have opened whole new avenues of inquiry and led to a number of predictive models to predict TF activity[11–13,15–21]. Although proteins of plant origin have been studied in a high throughput manner in yeast[13,17], these assays have so far not been conducted in plant systems. As a result, our view of plant transcriptional regulation in general and of plant TFs in particular is inherently biased by non-plant model organisms that diverged from plants more than 1 billion years ago[22]. These assumptions are inextricably embedded in current models to predict sequence features dictating the activity of plant TFs[13]. More broadly, MPRAs can be used to study other classes of *trans*-acting and protein-coding genetic elements beyond TFs such as immune effectors or metabolic enzymes whose activity in plants depends on other plant-specific proteins[23]. There is therefore a need for a high-throughput method to measure protein-coding gene libraries expressed in a plant host.

Here, we introduce ENTRAP-seq (Enrichment of Nuclear *Trans*-elements Reporter Assay in Plants), an MPRA based on magnetic sorting of isolated nuclei and the first MPRA capable of measuring *trans*-acting coding variants in plant cells. We deployed this method in *N. benthamiana* leaves expressing gene libraries delivered by *Agrobacterium tumefaciens*, allowing us to simultaneously measure the transactivation capacity of at least 2000 protein variants *in planta*. We combined this technology with a custom machine learning pipeline to predict and experimentally validate activation domains (ADs) from proteins across 1,495 plant viral genomes. We identified hundreds of viral ADs which exhibit positional conservation and tend to occur in structural proteins and enzymes not known to be involved in host gene regulation such as capsid proteins and RNA polymerases. Moreover, we demonstrate that many of these viral ADs can be

used as novel DNA parts for plant synthetic biology, opening the door to genome-scale mining of parts optimized for plant engineering. Finally, we show that ENTRAP-seq can be used to map and study the ADs of plant TFs. We used machine-based design to guide semi-rational mutagenesis of the central transcription factor in photoperiodic flowering CONSTANS, identifying sequence features critical for its function and enabling the generation of a new allelic series that includes repressors. ENTRAP-seq enables scalable functional genomic approaches to characterize natural and synthetic coding diversity in plants.

Results

Optimizing a magnetic nucleus sorting tag as a quantitative reporter assay

To measure the activity of multiple protein-coding elements in parallel, we sought to develop a MPRA based on separating nuclei by the level of expression of a reporter transgene (Fig. 1A). By placing this reporter under the control of a protein-coding library of regulators (*i.e.*, *trans*-regulatory elements), we aimed to sort nuclei into two fractions: a pulldown fraction enriched with regulators that activate the reporter, and a flow-through fraction enriched with repressors and/or inactive elements. Next-generation sequencing is used to quantify the relative enrichment of each *trans*-element in each fraction, which is proportional to their activity (*i.e.*, activators or repressors) and strength (Fig. 1A). As the expression platform, we used *Agrobacterium*-mediated transient expression in *N. benthamiana*, also known as agroinfiltration (Fig. 1A). Crucially, it was previously shown that transgene expression in this system follows Poisson-like statistics[24]. This makes it possible to minimize the number of cells expressing more than one *trans*-element from the library by infiltrating *Agrobacterium* cultures at a very low density. Single transgene expression can mitigate unwanted interactions that may occur when different *trans*-elements are expressed in the same cell.

To link reporter expression to nucleus sorting, we used a chimeric tag composed of a fusion of a biotin acceptor peptide to GFP and an <u>o</u>uter <u>n</u>uclear <u>e</u>nvelope targeting domain (hereafter ONEtag) which can be biotinylated by a coexpressed BirA transgene using endogenous biotin. Nuclei extracts incubated with streptavidin magnetic beads can then be sorted using magnets. Similar approaches have been used to purify nuclei from specific cell types in transgenic plants[25], but there are no reports of magnetic sorting of transiently transformed *N. benthamiana* tissues. We hypothesized that transiently transformed nuclei could be sorted into two fractions: a pulldown fraction enriched in T-DNAs encoding activators, and a flow-through fraction depleted of strong activators and enriched in T-DNAs coding for repressors and neutral elements (Fig. 1A). Importantly, for this approach to be feasible, magnetic nucleus sorting must be sensitive enough to detect quantitative differences in ONEtag expression levels among cells, rather than merely distinguishing between nuclei that do and do not express the ONEtag, as has been previously done to isolate cell type-specific nuclei[26]. To optimize magnetic nucleus sorting as a quantitative readout of transcriptional activity in agroinfiltrated *N. benthamiana*, we pursued two design goals. First, the transiently expressed ONEtag should localize to the nuclear envelope, as determined by the GFP subcellular localization. Second, as for any sensitive quantitative reporter, the level

134 of ONEtag accumulation must correlate with the transcriptional activity of the promoter driving its
135 expression.
136
137 As the first candidate ONEtag, we tested the AtRanGAP WPP domain, previously used to label
138 nuclei in specific tissues in *Arabidopsis*. To confirm the subcellular localization, we co-expressed
139 a nuclear marker (pUBQ::H2B-mScarlet, H2B-RFP). When transiently expressed in *N.*
140 *benthamiana*, the WPP tag did not localize to the nuclear envelope, as indicated by its failure to
141 encircle the H2B-RFP signal (Fig. 1B). A similar pattern was observed for the WPP domain of the
142 closest *N. benthamiana* RanGAP homolog. Since the C-terminus of WPP proteins has also been
143 used to target the nuclear envelope, we tested the C-terminal half of the *N. benthamiana*
144 homology of AtWIP1, which also failed to localize to the nuclear envelope. We next screened
145 seven domains known to localize to the outer nuclear envelope in their native contexts, including
146 the C-termini of two tomato WIP proteins. Several showed correct localization in *N. benthamiana*
147 (Fig. S1A), and we selected the C-terminal halves of WIP1 and WIP2 (SlWIP1/2-C) for further
148 characterization (Fig. 1B).
149
150 We next focused on the second design criterion, that is, ensuring that the degree of ONEtag
151 protein accumulation correlates with the transcriptional activity of the promoter driving its
152 expression. To measure promoter activity independent of ONEtag protein abundance, we used
153 an RFP-T2A-ONEtag fusion (Fig. 1C). RFP fluorescence intensity served as a proxy for promoter
154 activity while GFP fluorescence intensity provided a metric for ONEtag accumulation. To drive the
155 RFP-T2A-ONEtag reporter, we placed this transgene under the control of a minimal 35S promoter
156 with 6 upstream Gal4 DNA binding sites (UAS-min35S-RFP-T2A-ONEtag). To activate
157 expression of this reporter, we coexpressed it in *N. benthamiana* with previously generated
158 plasmids containing the Gal4 DNA binding domain alone (Gal4DBD) or fused to the strong VP16
159 activator (Gal4-VP16)[27]. As expected, the level of RFP fluorescence was higher in leaves
160 coexpressing the SlWIP1-C ONEtag and Gal4DBD-VP16 compared to Gal4DBD (Fig. 1D).
161 However, GFP fluorescence was comparable between Gal4DBD-VP16 and Gal4DBD, indicating
162 that the increase in transcriptional activity driven by the VP16 fusion was buffered
163 postranscriptionally and did not lead to an increase in SlWIP1-C ONEtag protein accumulation
164 (Fig. 1D). Cell microscopy confirmed that the ONEtag and RFP proteins are efficiently separated
165 after being produced from the same transcripts (Fig. S1B). While removal of BLRP did not relieve
166 buffering, removing SlWIP1-C was sufficient for GFP fluorescence to increase in response to
167 VP16, similar to RFP (Fig. S1C,D). We therefore replaced SlWIP1-C with SlWIP2-C. For this
168 ONEtag, GFP and RFP accumulated linearly in response to VP16 showing that the accumulation
169 of this tag closely correlates with the activity of the promoter driving it (Fig. 1E). Removing RFP-
170 T2A from the constructs made accumulation of SlWIP1-C ONEtag slightly responsive to Gal4-
171 VP16 but to a far lesser degree than SlWIP2-C (Fig. S1D) To further confirm that the UAS-
172 min35S::RFP-T2A-ONEtag construct containing SlWIP2-C acts as a quantitative reporter of
173 transcriptional activity, we coexpressed it with 8 previously generated Gal4DBD fusions spanning
174 a wide dynamic range in activity[27] and found RFP and GFP fluorescence to correlate well across
175 combinations (R^2 = 0.91, Fig. 1F, G, Table S1). Hence, accumulation of the SlWIP2-C ONEtag is
176 highly sensitive and quantitatively reflects the transcriptional activity of the promoter driving its
177 expression.

We previously determined that a low *Agrobacterium* OD of approximately 0.0025 is optimal for expressing libraries of *trans*-elements in *N. benthamiana* as it maximizes the number of transformed cells while minimizing the fraction of these cells expressing more than one transgene[24]. At this OD, approximately 10% of leaf cells are transformed and roughly 90% of these cells are transformed by a single strain. To determine the efficiency and yield of nuclei purified under these conditions, we infiltrated a SlWIP2-C UAS-min35S::H2B-RFP-T2A-ONEtag reporter at an OD of 0.0025. We coinfiltrated constitutively expressed Gal4DBD-VP16 and pUBQ::BirA both at a high OD of 0.1 which ensures that all cells express these two transgenes. To purify and sort nuclei, we developed a protocol based on pipelines optimized for other plant species (Materials and Methods)[26,28]. Using this protocol we were able to pull down approximately 1 million nuclei from about 20 fully infiltrated leaves, as determined using a hemocytometer (Materials and Methods). Virtually all the nuclei in the pulldown expressed RFP, reflecting an extremely small fraction of false positives in the purification (Fig. S2). The estimated number of transformed cells in the sample (approximately 30 million cells, see calculation in Fig. S2C) outnumbers that of the purified nuclei by about an order of magnitude, suggesting that purification enriches for nuclei that express the ONEtag at high levels (Fig. S2C).

Small-scale library validation of ENTRAP-seq

We next sought to validate our approach against an established measurement method[29]. To this end, we generated a small library composed of 26 synthetic TFs consisting of 53 previously characterized amino acid (aa) fragments – hereon referred to as 'tiles' – fused to Gal4DBD (Table S2)[17]. Based on previous measurements, we chose fragments expected to span a large dynamic range in activity, from repressor to strong activators[17]. We placed these Gal4DBD fusion tiles under the control of the constitutive pMas promoter and cloned them in an *Agrobacterium* T-DNA expression plasmid containing both the SlWIP2-C RFP-T2A-ONEtag construct under the control of UAS-min35S and a constitutive BirA transgene (Fig. 2A). All three cassettes were placed in the same T-DNA to ensure that biotinylated ONEtag expression occurs exclusively in cells that express a synthetic TF. We used this backbone for all subsequent experiments, unless specified (pSA676v2). To determine the activity of each tile, we infiltrated each construct individually at a final OD of 0.2 and measured the tissue RFP fluorescence normalized to a constitutive BFP strain. This ratio was then expressed as a function of the no-tile Gal4DBD control (Fig. 2B, Materials and Methods: Fluorescence plate reader reporter assays). The very low basal activity of the min35S promoter prevents detection of repression using fluorescence reporter assays. Hence, in the case of repressor tiles we coinfiltrated the tiles with a third construct containing a synthetic promoter with embedded Gal4 binding sites, which has a weak basal expression level[30].

Next, we combined these 26 *Agrobacterium* cultures carrying each of our control tiles in equal OD ratios and infiltrated the mix at a final OD of 0.0025. Nuclei purified from these leaves were subjected to magnetic sorting to obtain two fractions, pulldown and flow-through. Total DNA was extracted from each sample and amplicon sequencing of the T-DNA tile region was used to determine the abundance of each tile in each fraction. The level of enrichment strongly correlates with reporter fluorescence in the individual infiltrations (Spearman ρ = 0.91) indicating that

ENTRAP-seq can accurately and quantitatively measure TF regulatory strength (Fig. 2C). Overall, these results validate ENTRAP-seq as a way to measure *trans*-element activity in plant cells.

Accuracy and reproducibility in large libraries of viral protein tiles

We next sought to leverage ENTRAP-seq to discover novel transcriptional regulators. Plant pathogen attack is usually linked to widespread transcriptomic changes in the host through the action of pathogen-encoded transcriptional regulators[31]. There are well-known examples of such regulators delivered by pathogenic bacteria into plant cells, but how plant viruses manipulate host transcription is much less understood[32]. In the case of DNA viruses, transcription of the viral genome is part of the pathogenic cycle. Notably, many of the most widely used genetic parts in plant and mammalian genetic engineering – *e.g.*, the VP16 AD, the 35S promoter, and the SV40 NLS – originate from plant and animal viruses. Sourcing tools from plant viruses rather than animal viruses may enhance the biosafety profile of modified cultivars and facilitate regulatory approval. Plant viruses have very small genomes (typically in the order of 10-15 Kbp, Fig. S3) that code for just a few proteins (normally about 4, Fig. S3), none of which resemble known eukaryotic transcription factors[33]. Finally, a better molecular understanding of the viral transcriptional arsenal could help devise new crop protection strategies.

As of the time of writing, the NCBI database contained 1,495 viral genomes and 5,607 proteins. To maximize sequence diversity, we first clustered the protein sequences at 90% sequence identity, yielding 4,537 representative proteins. Each sequence was then tiled using 53 aa windows and 10 bp overlaps (Fig. 3A). To estimate the likelihood of each tile as a putative AD, we used PADDLE, a deep convolutional neural network that predicts acidic ADs[12]. The convolutional layer enables PADDLE to identify sequence-to-function relationships, especially within disordered sequences, where traditional structure- or homology-based comparative genomic algorithms fail. Although the model was trained on data from a yeast-based screen using yeast-derived protein sequences, acidic ADs are evolutionarily-conserved and we previously demonstrated that PADDLE can identify ADs from plants[17]. Only tiles with the highest predicted z-scores were selected for experimental testing (Fig. 3A). Finally, we further optimized protein diversity by selecting sequences that maximized the Hamming distance between binary protein encodings (Fig. 3A).

Although our nuclei enrichment method performed well on the control set of 26 tiles (Fig. 2C), the scalability of ENTRAP-seq to larger libraries was not immediately clear. To address this question, we created 4 nested libraries of 250, 500, 1K, and 2K viral tiles (Fig. 3A). All except for the 2K library also contained the 26 control tiles for which we had previously obtained single infiltration fluorescence data (Fig. 2C). For the 2K library, a new set of 26 viral tiles was measured individually in a fluorescence plate reader. These internal controls allowed us to validate the assay by comparing the activity of these tiles in terms of reporter fluorescence and tile enrichment. Across all libraries, we found that the activity of a given control tile as measured individually in a fluorescence plate reader strongly correlates with its enrichment following magnetic pulldown for two biological replicates of each library (Spearman ρ = 0.72-0.97) (Fig. 3B-E). These validation experiments allowed us to obtain a calibration value to classify tiles as activators based on their

sequencing enrichment. To this end, we fitted the validation data to a sigmoidal curve and calculated the enrichment value corresponding to the midpoint of this sigmoid plus its estimation error (Fig. 3B-E). This yielded an enrichment value that was used as a cutoff to classify viral tiles as activators (Fig. 3B-E).

When the enrichment of all tiles was considered, we found that biological replicates correlate well (Pearson r = 0.5-0.55) and this correlation increased substantially to r = 0.7-0.8 by applying even modest library coverage cutoffs (Fig. 3F). We found no consistent correlation between tile coverage in the library and their enrichment ratio (Fig. S3). Because of the nested nature of the libraries, it was also possible to compare tile enrichment across libraries. This revealed a high degree of reproducibility even across libraries (Pearson r = 0.59-0.8 without applying coverage cutoffs) (Fig. 3G). Thus, our assay is accurate and reproducible in libraries of up to 2K tiles.

Identification of activation domains in plant viral proteins

Our pipeline for discovering viral ADs combining machine learning with high-throughput measurements yielded 1,105 tiles capable of activating the reporter, confirming that PADDLE can predict acidic ADs in plants[17] (Fig. 4A). These tiles are found throughout the phylogenetic tree of plant viruses (Fig. 4A). Two large families of plant ssDNA and dsDNA viruses (*Geminiviridae* and *Caulimoviridae*, respectively*)* as well as multiple +ssRNA families (especially *Potyviridae* and *Betaflexiviridae*) were particularly rich in acidic ADs (Fig. 4A). We did not find any predicted acidic ADs in only 4 relatively rare families: *Amalgaviridae*, *Aspiviridae*, *Mitoviridae*, and *Nanoviridae* (-ssRNA, -ssRNA, +ssRNA, and ssDNA, respectively). This contrasts with recent global surveys of human viruses where ADs were found almost exclusively in a few families of dsDNA viruses[20,34]. Although our screen did not encompass all viral genomes, it identified approximately twice as many tiles as the more comprehensive studies of human viral activation domains, which reported 500–600 ADs. Thus, ADs appear to be much more widespread in plant viruses compared to their human counterparts. In general, our search algorithm based on PADDLE performed remarkably well in terms of predicting ADs; ~50% of the tested tiles qualify as activators based on our calibration thresholds, although there was substantial variation in predictive accuracy across taxa (Fig. 4A, Fig. S5).

To better understand the nature of these potential regulators, we performed GO term enrichment analysis. We found activator tiles in proteins from a wide range of functional annotations, although we note that due to the biased nature of our search not all classes are represented equally in our libraries (Fig. 4B, Fig. S6). The presence of ADs in structural and catalytic proteins reflects the known multifunctional nature of viral proteins and mirrors the fact that only about 10% of human viral ADs were found in regulatory proteins[19]. Four kinds of plant viral proteins showed a notable enrichment of activator tiles: geminiviral AL2-like proteins, RNA-dependent RNA Polymerases (RdRp), helicases, and helical capsid proteins (Fig. 4B). The AL2 protein, also known as TrAP (transcriptional activator protein), is involved in transactivating viral promoters in *Geminiviridae* and has been shown to regulate genomic promoters by physically interacting with host TFs[35,36]. AL2 lacks a DNA-binding domain but contains an C terminal acidic region that can transactivate a synthetic reporter when fused to Gal4DBD[37]. Thus, the recovery of strong ADs in AL2-like

310 proteins using ENTRAP-seq validates our framework for discovering new viral-encoded
311 transcriptional regulators. Our high-resolution tiling of these proteins allowed us to precisely map
312 the boundaries of their putative ADs (Fig. 4C). Consistent with the C-terminus of AL2 being
313 important for transactivation, we found that ADs are consistently located in three positions within
314 a highly disordered region in the C-terminus of diverse AL2 proteins (Fig. 4C). The positions of
315 RdRp activator tiles are also relatively conserved but unlike AL2 these putative ADs are clustered
316 in multiple distinct regions throughout the protein, most of which are predicted to fold into stable
317 structures (Fig. 4C). Similar to RdRp, putative activator tiles in viral helicases and helical viral
318 capsid proteins consistently fall in a few structured regions. In terms of prediction accuracy, the
319 large majority of predicted ADs in AL2 and helical capsid proteins turned out to be true activators
320 while the opposite was true for RdRp and helicases (Fig. 4C). These data suggest that a dedicated
321 regulatory protein like AL2 might activate transcription in a similar manner as canonical eukaryotic
322 TFs while multifunctional viral proteins may repurpose stably folded and conserved catalytic or
323 structural domains to accomplish this task.
324
325 The discovery of these viral regulators prompted us to consider their potential as novel plant
326 synthetic biology tools. To find potent and compact activators that could serve as alternatives to
327 existing field standards, we individually cloned 20 tiles that were among the most enriched in at
328 least one library. We then tested them individually using the same Gal4-UAS fluorescence
329 reporter assay described above, side by side with two potent and popular ADs used in the field,
330 VP16 and VPR[38]. All 20 tiles strongly activated the reporter (Fig. S7) and 4 performed as well or
331 better than VP16 and VPR (Fig. 4D). To determine if these ADs perform well in different genetic
332 contexts, we fused them to dCas9 and tested their activity upon recruitment to a synthetic
333 promoter driving GFP as a reporter (Fig. 4E)[39]. dCas9-VPR outperformed all other fusions and
334 one tile was stronger than VP16 when fused to dCas9 (Fig. 4E). Since VPR is considerably larger
335 than our viral tiles (553 vs 53 aa), this compact viral AD may be advantageous for vectors with a
336 limited size capacity. In addition to minimizing construct size, our viral tiles offer a diverse set of
337 genetic parts to reduce unintended consequences in synthetic programs, including cross-
338 interactions with modules sharing identical parts and gene silencing arising from context-
339 dependent effects.
340
341 **Partial conservation of AD function between distant eukaryotes**
342
343 Recently, massively parallel reporter assays in yeast have identified plant protein sequences
344 encoding ADs[13,17]. Deriving insight about plant ADs from these experiments in a distantly-related
345 host depends on the degree to which *trans*-activation mechanisms are conserved between these
346 two systems. It is well documented that there is substantial conservation of transcriptional
347 activation mechanisms between yeast and plants. Decades of insights about plant TFs derived
348 from yeast-based assays and our own results showing that a model trained in yeast successfully
349 predicts plant viral ADs support this idea[17] . Yet, it is not known the degree to which each model
350 organism has its own taxa-specific regulatory pathways and whether the quantitative activation
351 strength of ADs is consistent between yeast and plants. To systematically address these
352 questions, we took advantage of an existing dataset where thousands of protein tiles from
353 *Arabidopsis* TFs were tested in yeast[13]. We leveraged ENTRAP-seq to test a subset of these tiles

354 in *N. benthamiana* and compare their activity across model systems. We randomly selected 250
355 tiles while keeping the same distribution of yeast transcriptional strengths (referred to as PADI
356 scores[13]) from the original library. We then cloned this library in the same vector used for our viral
357 libraries and performed a nucleus enrichment experiment.
358
359 Comparing the rank ordering of activity between yeast and *N. benthamiana* revealed a very limited
360 correlation (Spearman $\rho = 0.18$, $p = 0.0075$) (Fig. 5A). Indeed, numerous strong yeast activators
361 were depleted from the ENTRAP-seq pulldown and multiple enriched tiles in *N. benthamiana*
362 were not found to be activators in yeast (Fig. 5A). Because yeast-based MPRAs are heavily
363 biased towards identifying acidic ADs, we hypothesized that tiles not found to be activators in
364 yeast might belong to other classes of ADs. To classify tiles as likely plant activators we used an
365 enrichment threshold of 1.2 — the minimum enrichment value for validated activators from the
366 viral libraries. Among the 40 tiles that had an ENTRAP-seq enrichment greater than 1.2 and a
367 PADI score lower than 1 (labeled 'quadrant 4' in Fig. 5A) we found two Q-rich sequences that
368 were not classified as activators in yeast. The *trans*-activation role of Q-rich stretches is conserved
369 in eukaryotes[40]. The first of these polyQ ADs is located in the middle region (MR) of ARF19 (Fig.
370 5B). The MR is important for *trans*-activation in activator ARFs[41] but the yeast assay mapped the
371 ADs of ARF19 and other activator ARFs to regions flanking the MR[13]. The second polyQ stretch
372 belongs in the C terminus of the AGAMOUS-LIKE gene ALG65, a region that is thought to contain
373 the AD in this family of TFs (Fig. 5C)[42].
374
375 An alternative hypothesis for why ADs can be active in plants but not in yeast is that these tiles
376 are not 'bona fide' ADs, namely, intrinsically disordered regions that contact the core
377 transcriptional machinery. Instead, these tiles may code for folded protein domains that bind plant-
378 specific TFs that in turn contain canonical ADs that activate transcription in yeast. To explore this
379 possibility, we calculated the predicted disorder of tiles in the library and compared this prediction
380 with their ENTRAP-seq enrichment. We found these two metrics to be completely uncorrelated
381 (Spearman $\rho = 0.13$, $p = 0.054$) and tiles that are active in plants (enrichment >1.2) are equally
382 likely to be structured as to be disordered (Fig. 5D). Thus, tiles with a plant-specific bias do not
383 necessarily form folded domains. However, among highly structured tiles we found cases that do
384 fit our initial hypothesis. For example, a tile in the NAC001 TF containing part of a structured
385 domain involved in homo- and hetero-dimerization within NAC family TFs was found to be highly
386 active in plants but not in yeast (Fig. 5E)[43]. The NAC TF family is plant-specific and contains acidic
387 ADs in their C-terminal half that are activators in yeast[44] (Fig. 5E). Thus, in *N. benthamiana*, this
388 NAC001 tile might simply recruit endogenous NAC activators which, in turn, contain ADs that are
389 active in both *N. benthamiana* and yeast. In yeast however, this NAC001 tile may be incapable of
390 activating the reporter because this organism lacks interacting NAC partners.
391
392 We confirmed that the polyQ ADs of ARF19 and AGL65 as well as the NAC001 AD activated the
393 reporter in our fluorescence plate reader assay (Fig. 5F). In this ENTRAP-seq experiment, we
394 found 14 tiles that had a PADI score > 1 and an enrichment value < 1.2 (labeled 'quadrant 1' in
395 Fig. 5A). To test if these correspond to yeast-specific activators, we individually cloned and
396 measured 12 of these tiles. Only 25% (3/12) were activators in our plant plate reader assay, the
397 rest were either repressors (5/12) or were statistically indistinguishable from Gal4DBD alone

398 (4/12) (Fig. 5G). These results demonstrate that plant-specific as well as yeast-specific AD tiles
399 are pervasive. The stark differences in activity between the two systems for some tiles is
400 consistent with there being fundamental differences in transcriptional activation mechanisms
401 between these two models, although other explanations are conceivable as well. For example,
402 the reporter in the yeast assay is chromosomally integrated while the ENTRAP-seq reporter is
403 transiently expressed. To better understand the degree of conservation of ADs between these
404 two model eukaryotes, future studies will be necessary to leverage ENTRAP-seq to measure all
405 tiles across all *Arabidopsis* TFs in an unbiased and systematic manner. Finally, these findings
406 show that it is wise to exercise caution when interpreting tile MPRA data of plant proteins tested
407 in yeast as both false positives and false negatives can arise when using one model system to
408 represent the biology of another.
409
410 **Rationally tuning the activity of a plant TF**
411
412 One of the implicit assumptions of AD recruitment assays is that TFs are largely modular[45]. If this
413 is the case, it may be possible to study and engineer plant TFs in high throughput by isolating,
414 mutating, and screening their ADs for altered functionality. To test this idea, we leveraged
415 ENTRAP-seq to characterize a mutant library of the *Arabidopsis* activator TF CONSTANS (CO)
416 with the goal of creating weaker and stronger activators than wild type as well as alleles that
417 switch CO from activation to repression of target genes. Natural variation in CO and its homologs
418 is associated with flowering time, a key agronomic trait[46,47]. Although *Arabidopsis* CO appears to
419 act exclusively as a transcriptional activator, related CO-like genes in this and other species can
420 act as repressors, suggesting that it is feasible to switch the CO regulatory role via coding
421 mutations[48–50]. In agreement with the role of CO in its native context and previous reports,
422 transient expression in *N. benthamiana* of a fusion between CO and a blue fluorescent protein
423 (BFP-CO) strongly activates expression of the FT promoter (1.9 kb upstream of the ATG,
424 hereafter FTpro) as measured by a GFP reporter assay (Fig. 6A)[51].
425
426 CO contains a poorly conserved, highly disordered middle region that is sufficient for
427 transactivation in a Gal4DBD recruitment assay (Fig. 6B)[51]. Although multiple CO alleles have
428 been identified in random mutagenesis screens, all lie outside this middle region and are loss-of-
429 function (Fig. 6B)[52]. To better define the position of the CO AD within this middle region, we
430 leveraged PADDLE as well as TADA, another model trained on datasets from yeast-based
431 screens that can predict acidic ADs[11,13]. Both models have a high degree of agreement and point
432 to a region around aa position 175 as the putative CO AD (Fig. 6B). We therefore selected a 69
433 aa window in this region as the putative AD (Fig. 6B). To explore the sequence space of the CO
434 AD, we generated a library of 350 tiles carrying 1-4 aa substitutions and including the wild-type
435 AD as a control. The substitutions were rationally designed to either increase or decrease AD
436 strength based on features important for acidic AD function in yeast and animal cells, namely 1)
437 depletion of positive residues (R and K), 2) enrichment of negative residues (D and E), 3)
438 enrichment of hydrophobic residues (particularly aromatics and leucine), 4) depletion of
439 continuous stretches of hydrophobic residues, and 5) lack of stable secondary structure[6,11,53,54].
440 Following the same rationale, we also mutated every residue to K, P, W, and D. As internal
441 controls we included 5 wt tiles. Finally, we introduced 10 random synonymous mutations in each

442 tile in the library to better distinguish between highly similar tiles from sequencing. Following the
443 same strategy used for our previous experiments, we cloned these tiles as Gal4DBD fusions in
444 our magnetic sorting reporter and performed an ENTRAP-seq assay (Fig. 6C).
445
446 The enrichment of the CO AD tiles spanned a range of values comparable to other libraries and
447 the enrichment of WT tiles was highly reproducible (1.12 ± 0.02). Perhaps not surprisingly, most
448 tiles were loss-of-function mutants, *i.e.*, less enriched than WT (Fig. 6D). We found a moderate
449 agreement between tile activity as measured by ENTRAP-seq and the predictions from PADDLE
450 and TADA (Spearman ρ = 0.38 and 0.4, respectively) and the two models largely agreed with one
451 another (Pearson r = 0.89) (Fig. S8). This suggests that plant-specific predictive models could be
452 used to aid in the design of mutant plant AD screens. Comparing the activity of each tile to its
453 biochemical properties showed that, as expected for an acidic AD, negative charge and
454 hydrophobicity have a significant correlation with activity (Spearman ρ = -0.34 and 0.24,
455 respectively, Fig. S9). Throughout the AD, mutation of residues to K tended to decrease the
456 activity and most of the strong loss of function AD alleles are K substitutions (Fig. 6E, F). The
457 majority of the gain-of-function single mutants correspond to W or D substitutions (Fig. 6E, F).
458 Mutating residues to P, which severely disrupts protein secondary structure, rarely had strong
459 effects (Fig. 6E, F). These results are consistent with the CO AD behaving like a classic
460 intrinsically disordered acidic AD and highlights the power of combining predictive models with
461 semi-rational mutagenesis guided by known AD compositional biases.
462
463 To test the effect of these mutations in the context of the full-length CO protein, we chose 10
464 putative loss-of-function and gain-of-function alleles among the tiles at either end of the
465 enrichment value distribution. We then introduced the mutations found in each of these tiles into
466 the full-length BFP-CO expression construct and tested their effect individually using the FTpro
467 fluorescence reporter assay we previously established (Fig. 6G). As controls we also cloned three
468 previously isolated semi-dominant loss-of-function alleles corresponding to point mutations
469 outside the AD obtained through EMS mutagenesis. Out of 10 predicted loss-of-function CO
470 mutants, 7 were significantly weaker activators than WT and 2 of those were as weak as classic
471 alleles. The M186P mutant behaved as a gain of function allele, in contrast to its activity as a Gal4
472 DBD fusion tile (Fig. 6H). The EMS mutant controls as well as some of our isolated alleles
473 switched CO activity from an activator into a repressor, consistent with the EMS mutants being
474 semi-dominant in *Arabidopsis*[52]. Out of the 10 predicted gain-of-function mutant alleles, 5 were
475 stronger activators than WT (Fig. 6H). Among the rest, 4 behaved as activators but were not
476 statistically stronger than WT and one putative gain of function allele (Y194W) was a slightly
477 weaker activator than WT. We conducted an independent replicate of this whole experiment and
478 obtained similar results (Pearson r = 0.92, Fig. S10). Collectively, these data demonstrate that tile
479 activity in the context of an ENTRAP-seq assay can be predictive of tile function within the original
480 TF protein environment. This allowed us to rapidly engineer an entirely new allelic series of a
481 plant transcription factor, including switching its function from activator to repressor.
482

Discussion

ENTRAP-seq overcomes a major technical hurdle in plant functional genomics, enabling high-throughput, quantitative measurement of transcriptional regulatory activity from protein-coding sequences directly in plant cells. While methods like DAP-seq and STARR-seq have shed light on DNA-binding preferences and *cis*-regulatory landscapes, they do not capture the *trans*-regulatory activity of TFs and other regulatory proteins. This gap is particularly critical given that many transcriptional regulatory domains — such as ADs — are intrinsically disordered and poorly conserved, making them difficult to identify through structure- and homology-based algorithms. ENTRAP-seq addresses this challenge by systematically linking protein sequence to transcriptional output in a native plant context, allowing for the functional annotation of regulatory domains at scale. By enabling genome-wide characterization of *trans*-regulatory elements, ENTRAP-seq lays the groundwork for constructing more complete and mechanistic models of plant gene regulatory networks (GRNs). When integrated with ongoing efforts to map the *cis*-regulatory landscape, ENTRAP-seq will play a crucial role in advancing systems-level understanding of plant transcriptional regulation.

Beyond mapping transcriptional circuitry, ENTRAP-seq provides a powerful framework for dissecting and engineering protein function. Its parallelized, library-based design is well-suited to systematic exploration of sequence-to-function relationships, akin to deep mutational scanning approaches that have revealed critical residues and domains in other systems. By screening libraries of natural variants or synthetic constructs, ENTRAP-seq enables the direct measurement of how coding sequence variation impacts regulatory activity—an essential step in linking genotype to phenotype. This capability has broad implications for plant biology and crop improvement. For example, assessing allelic diversity across populations could reveal how *trans*-regulatory mutations contribute to traits such as stress tolerance, developmental timing, or yield. Furthermore, coupling ENTRAP-seq with machine learning models trained on plant-specific data offers a promising route toward the rational design of synthetic TFs with customized regulatory properties. As tools for genome editing and synthetic biology continue to evolve, ENTRAP-seq stands as a critical functional bridge—empowering both foundational research and translational advances in plant science.

Acknowledgments

First and foremost we would like to thank all members of the Shih lab. From weekly growing healthy plants to making *Agrobacterium* competent cells to discussing results, their support enabled this research. In particular, we are thankful to Niklas Hummel for introducing us to agroinfiltration and the study of TFs in plants and for sharing countless plasmids. Amy Lanclot and Maximo Kesselhaut provided insightful comments during the preparation of this manuscript. We also thank a number of colleagues with whom we discussed preliminary data and ideas during the development of this project. We had productive and inspiring discussions with Max Staller, Roger Deal, Jose Miguel Alvarez, Patricia Lang, Hernan Garcia, Christine Queitsch, and George Coupland. A number of scientists provided critical technical feedback as well. Yu Tang and Julia Bailey-Serres shared constructs for nuclear envelope targeting, Adam Deutschbauer gave us

advice for library cloning and sequencing and Rita Kuo and Noam Prywes guided us through the early stages of next-generation sequencing. Finally, we thank Shana McDevitt from QB3 Genomics, UC Berkeley, Berkeley, CA, RRID:SCR_022170 and Brian McCarthy from the UC Berkeley DNA sequencing facility for their advice and assistance. This work was part of the DOE Joint BioEnergy Institute (https://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

Data, code, and sequencing availability

All code and models related to this study are publicly available on GitHub (github.com/shih-lab/magnets). Raw sequencing reads are publicly available through the NCBI SRA Database under BioProject accession PRJNA1254296.

Competing Interests

A patent for characterizing gene libraries in plant cells was filed by the Lawrence Berkeley National Laboratory with S.A. and P.M.S and as inventors. P.M.S. has financial interests in BasidioBio.

Author Contributions

P.M.S provided strategic scientific guidance and funding for this project. P.M.S, L.W, and S.A conceived the research and wrote the manuscript. L.W and S.A analysed the data. L.W performed statistical inference, modeling, and phylogenetic analyses. S.A, A.D, L.O, R.R, and S.O contributed biological materials. S.A performed experiments.

Materials availability

All plasmids are available from the JBEI registry. The sequences of all plasmids used in this study are listed in table S4. The sequences of plasmids that were created for this study can be found in the public google drive folder:
https://drive.google.com/drive/folders/1ogtAr0cA7BEVsjFaDrkzEIUccGWJD2VS?usp=drive_link

Materials and Methods

Molecular cloning: Except for dCas9 fusions, all tiles were cloned in the pCambia1300-based backbone pSA676v2 containing 3 transgenes: 6xUAS-RFP-T2A-ONEtag, pUBQ10::BirA, and pMas::Gal4DBD-dropout. The dropout consists of a E.coli GFP reporter flanked by BsaI sites to allow negative GFP colony selection. Individual tiles were cloned in pSA676v2 using gibson assembly. These constructs were used for fluorescence reporter assays. Tiles fused to dCas9 were cloned into backbone pSA742 containing p35S::dCas9-dropout using gibson assembly. The sequence of pSA676v2 and pSA742 as well as other plasmids used in this study can be found in the drive folder listed above. For library cloning see Materials and Methods: Library cloning.

572
573 **Live cell Imaging:** Leaf samples 3 days post infiltration were mounted in water. All images were
574 acquired in a Zeiss LSM 710 confocal microscope using a 63X oil objective. Channels were
575 acquired sequentially using 20 z slices every 0.5 um and 1024 x 1024 pixel image size. The zoom
576 was 2.0 and the wavelengths were 568 nm excitation, 585-630 nm emission for RFP, and 488
577 nm excitation, 494-581 nm emission for GFP. The pinhole in both channels was set to 1 AU.
578
579 **Library cloning:** Single stranded oligo pools containing tiles of each library were obtained from
580 Twist Biosciences. The oligos were flanked by a shared overhang sequence containing BsaI cut
581 sites for cloning into the ENTRAP-seq backbone pSA676v2. The overhangs were 5'
582 ACGTCAGTGTGGTCTCT**AGGT** and TAA**GCTT**AGAGACCAAGCGCTTA 3' (BsaI sites
583 underlined, cloning overhangs in bold). Oligo pool DNA was resuspended in 150 ul of water and
584 used for a 50 ul PCR reaction using primers binding to the overhangs (tileLib-F and tileLib-R,
585 table S5). The PCR reaction contained 1 ul of resuspended oligos as template, 2.5 ul of each
586 primer at 10 uM, 19 ul of water, and 25 ul of 2x NEB Q5 high fidelity DNA polymerase master mix.
587 The reactions were run using the standard conditions provided by the Q5 manufacturer using a
588 melting temperature of 65C, an elongation time of 15 sec and 15 cycles. A control 10 ul reaction
589 was run for 35 cycles. Both reactions were run side by side on a 1% agarose gel to confirm correct
590 amplification. The PCR was purified using Zymoclean DNA clean and concentrator columns using
591 two washes with 600 ul of wash buffer and eluting in 30 ul of water. The dsDNA amplified oligo
592 pool was then digested overnight at 37C using the following conditions: 13 ul of water, 30 ul of
593 purified PCR, 5 ul of NEB 10X HF buffer, 2 ul of NEB BsaI-HF. The digestion was purified using
594 Zymoclean DNA clean and concentrator columns using two washes with 600 ul of wash buffer
595 and eluting in 20 ul of water. For the backbone, approximately 10 ug of pSA676v2 DNA were
596 obtained from a standard miniprep. The plasmid was digested overnight at 37C with NEB BsaI-
597 HF using the following reaction conditions: 50 ul plasmid miniprep DNA, 3 ul of restriction enzyme,
598 8 ul of 10X HF buffer, 19 ul of water. The entire digestion was run on a 1% agarose gel and the
599 band corresponding to the backbone was purified using the Zymoclean gel DNA recovery kit using
600 two washes with 600 ul of wash buffer and eluting in 20 ul. The digested dsDNA library was ligated
601 into digested pSA676v2 using NEB T7 ligase. 500 ng of backbone (approximately 6 ul) and
602 enough insert to achieve a ratio of 3:1 insert:backbone were added to 120 ul final reaction volume
603 ensuring that the total DNA concentration did not exceed 10ng/ul. The reaction was run using 3
604 cycles of 2 hr at 22C and 4 hr at 16C. The ligations were column-purified and eluted in 15 ul of
605 water.
606
607 **Library transformation into bacteria**: One tube of E. coli SIG10 MAX electrocompetent cells
608 (Millipore Sigma CMC0004) was electroporated with 2 ul of ligated library and recovered in LB at
609 37C for 1 hr. Following recovery, 1 ul of cells was plated in 50 µg/ml Kanamycin LB plates in
610 triplicate to assess the number of colony forming units (CFUs) per transformation. The minimum
611 number of CFUs used was 10 times the library size. The rest of the culture was inoculated in 50
612 ml of 50 µg/ml Kanamycin LB and grown overnight at 37C. A miniprep was run on 5 ml of the
613 overnight culture to obtain plasmid DNA for *Agrobacterium* transformation. Electrocompetent For
614 each library, *Agrobacterium* GV3101::pMP90 was electroporated using 10 ul of competent cells,
615 2 ul of library miniprep (approximately 500 ng of DNA) and 88 ul of water in each cuvette. Cells

616 were recovered in 1 ml of LB at 28C for 2 hr. 5 such transformations per library were used to
617 increase the number of CFUs. Following recovery, 1 ul of cells was plated in LB plates containing
618 50 µg/ml Rifampicin, 50 µg/ml Kanamycin, and 30 µg/ml Gentamicin in triplicate to assess the
619 number of CFUs per transformation. The minimum number of *Agrobacterium* CFUs used was 10
620 times the library size. Recovered *Agrobacterium* cells were inoculated in 100 ml of LB containing
621 the same antibiotic concentrations and grown for 2 days at 30C shaking. Glycerol cryostocks were
622 prepared from these cultures.
623
624 **Plant growth conditions:** *Nicotiana benthamiana* was grown in a controlled environment room
625 at 25 C of temperature, 60% humidity, 16/8 hr light/dark cycle, and a light intensity of 120
626 µmol/m^{2}s. Prior to agroinfiltration, plants were grown for 29 days (sowing at day 0) in individual
627 pots using Sunshine Mix #4 soil (Sungro) supplemented with 499 Osmocote 14-14-14 fertilizer
628 (ICL) at 5mL/L. Immediately after agroinfiltration plants were returned to the growth room for 3
629 days until sampling.
630
631 ***Agrobacterium* growth conditions:** For fluorescence plate reader assays *Agrobacterium*
632 glycerol stocks were streaked onto LB agar plates supplemented with antibiotics: 50 µg/ml
633 Rifampicin, 50 µg/ml Kanamycin, and 30 µg/ml Gentamicin. The day prior to infiltration, individual
634 colonies were inoculated into liquid LB medium containing the same antibiotics and incubated
635 overnight at 30°C, shaking. On the day of infiltration, cultures were diluted 1:10 into LB with
636 identical antibiotic concentrations and incubated further under the same conditions for 3 hours or
637 until an OD600 of ~1.0 was reached. For gene libraries, transformed cells previously recovered
638 for 2 hr in LB without antibiotics were transferred to 100 ml of LB containing the same antibiotic
639 concentrations listed above. Two days later, the OD of this culture was measured and the culture
640 was diluted 1:10 if the OD was higher than 1.0 prior to infiltration. Alternatively, approximately 100
641 ul of library glycerol cryostocks were inoculated into LB containing antibiotics the day before
642 infiltration. The day of infiltration, the culture was diluted 1:10 and grown until reaching an OD of
643 ~1.0.
644
645 **Agroinfiltration:** *Agrobacterium* liquid cultures with an OD600 between 0.5 and 1.0 were pelleted
646 by centrifugation at 4,000 × g for approximately 10 minutes and resuspended in an equal volume
647 of infiltration buffer (10 mM MES, pH 5.6; 10 mM $MgCl_2$; 150 µM Acetosyringone). The
648 resuspended cultures were incubated with shaking at room temperature for 1 hour before OD600
649 measurements were taken. Next, a 1:5 dilution of each culture was prepared in infiltration buffer,
650 and the OD600 of this dilution was determined using a spectrophotometer. Final OD600 dilutions
651 for infiltration were then made using the same buffer. The mixtures were infiltrated into the 6th
652 and 7th leaves (counting upward, with cotyledons considered as leaves 1 and 2) within 1 hour of
653 preparation. For fluorescence reporter assays 3 mixes were infiltrated per leaf at an OD of 0.2
654 each (total bacterial OD of 0.6). For nucleus purification infiltrations the culture was diluted to OD
655 0.0025.
656
657 **Fluorescence plate reader reporter assays:** For leaf fluorescence measurements of
658 transcriptional reporters, 4 6mm leaf disks were cut from each agroinfiltrated leaf using a single-
659 hole puncher and deposited on top of 350µL of tap water in a clear-bottom black 96-well plate. All

660 reporter constructs were cloned in the ENTRAP-seq construct backbone (pSA676v2). The DBD-
661 only control contains a stop codon instead of a tile (pSA676v2s). For normalization a pMAS::BFP
662 strain was coinfiltrated (pSA639). Both strains were diluted to OD 0.4 and mixed in equal ratios
663 to a final OD of 0.2 each. Three days post infiltration the plates were read on a BioTek Synergy
664 H1 plate reader using 488 nm excitation and 520 nm emission for GFP, 587 nm excitation and
665 615 nm emission for RFP, 400 nm excitation and 454 nm emission for BFP. The normalized
666 reporter activity reported throughout this work corresponds to a double normalization to BFP and
667 to Gal4DBD alone. First, the RFP fluorescence of each leaf punch was divided by its BFP
668 fluorescence. Next, this value was divided by the mean RFP/BFP fluorescence of Gal4DBD.
669
670 **Nucleus Purification Buffer (NPB):** NPB was prepared following the recipe from Wang & Deal
671 2015: 20 mM MOPS pH 7, 40 mM NaCl, 90 mM KCl, 2 mM EDTA pH 8, 0.5 mM EGTA pH 8, 0.5
672 mM spermidine, 0.2 mM spermine, 1 complete protease inhibitor tablet per 100 ml. A 5x stock
673 solution was prepared adding everything except for fresh ingredients: spermidine, spermine and
674 protease inhibitors. The fresh ingredients were added to the 1x solution up to one day prior to
675 nucleus pulldown. NPBt was prepared following the same recipe and protocol except that 0.1 %
676 (v/v) Triton X-100 was added up to one day before nucleus pulldown.
677
678 **Preparing tissue for nucleus sorting:** *Agrobacterium* cultures containing gene libraries were
679 diluted to a final OD of 0.0025 in IB and infiltrated into 8-10 4-week old *N. benthamiana* plants
680 (see Materials and Methods: Agroinfiltration). The full area of true leaves 6-8 of each plant was
681 infiltrated. 3 days post infiltration, leaves were harvested and the mid vein was removed. A ~5mm
682 perimeter around the edge of the leaves was discarded as well. 10 g of leaf material was then
683 chopped into ~1cm^2 pieces using a razor blade. The chopped leaves were submerged in 50 ml
684 of nucleus purification buffer (NPB) in a beaker and placed in a vacuum chamber for 10 min to
685 infiltrate NPB into the tissue (see Materials and Methods: Nucleus Purification Buffer (NPB)).
686 Further rounds of vacuum filtration were applied until the tissue was completely soaked in NPB.
687 Next, the tissue was washed 3 times with tap water and tapped dry with paper towels. Finally, the
688 leaf material was flash frozen in liquid nitrogen, ground into a very fine powder using mortar and
689 pestle and stored at -80 C.
690
691 **Nucleus sorting:** Frozen ground tissue (see Materials and Methods: Preparing tissue for nucleus
692 sorting) was resuspended in 100 ml of ice-cold NPB and passed through a 70 um strainer into 2
693 50 ml tubes to separate nuclei from cells and other larger leaf fragments, keeping the material on
694 ice at all times. This filtrate was centrifuged at 500 G for 10 minutes at 4 C. The pellets from both
695 tubes were combined by resuspending them in 20 ml of NPB. The resuspension was passed
696 through a 100 um strainer to remove remaining pieces of tissue. In parallel, 100 ul of MyOne
697 streptavidin dynabeads were washed by resuspending them in 1.5 ml of NPB, separating them in
698 a magnetic rack and resuspending in 100 ul of NPB. The washed dynabeads were then added to
699 the resuspended nuclei and the mix was incubated for 30 minutes shaking gently in a cold room.
700 All following steps were carried out in a cold room. The nuclei were diluted in NPBt to 100 ml final
701 split into 2 50 ml tubes. The samples were then placed in a magnetic rack for 30 minutes to pull
702 down nuclei. The supernatant was removed and centrifuged for 10 min at 500 G, resuspended in
703 25 ml of NPB per tube and centrifuged again at 500 G for 10 min to obtain a pellet corresponding

704 to the flowthrough sample. The pulldown from the first magnetic separation was washed twice by
705 resuspending it in 50 ml of NPBt and placing it in the magnetic rack for 20 min. The washed nuclei
706 were next resuspended in 15 ml of NPBt in a 15 ml tube and placed in the magnetic rack again
707 for 20 min. The pellet from this separation was resuspended in 3 ml of NPB and 10 ul of this
708 sample were taken for imaging. The rest of the sample was centrifuged at 1000 G for 10 min to
709 obtain a pellet corresponding to the pulldown sample. The pellets corresponding to the
710 flowthrough and pulldown samples were resuspended in 1 mL and 200 ul of CTAB, respectively
711 (100 mM Tris-Cl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl, 2% (w/v) cetyltrimethyl ammonium
712 bromide, 1% PVP 40,000). These samples were stored at -20 C until DNA extraction.
713
714 **Nucleus DNA extraction:** Purification of DNA from the flowthrough and pulldown samples was
715 carried out based on a previously published protocol[55]. CTAB-resuspended pellets were
716 incubated at 65 C for 30 minutes vortexing every 10 minutes. A volume approximately equal to
717 the pellet volume of phenol:chloroform:isoamyl alcohol 25:24:1 v/v was added per tube and tubes
718 were vortexed thoroughly. Samples were centrifuged at 20K G for 2 min at room temperature.
719 The top aqueous layer was removed and mixed with an equal volume of CTAB in a new tube.
720 Samples were vortexed and incubated at 65 C for 20 minutes. A volume of chloroform:isoamyl
721 alcohol 96:4 v/v equal to the volume of CTAB added in the last step was added to each tube.
722 Samples were vigorously vortexed and centrifuged at 20K G at room temperature for 2 minutes.
723 The top aqueous layer was transferred to a new tube and 70% of this volume of isopropanol was
724 gently added to the wall of each tube. Samples were mixed gently by inverting 5-10 times and
725 placed and -20 C for 20 min. The DNA was pelleted by centrifugation at 20K G for 15 min. The
726 supernatant was removed and the pellets were washed with ice-cold 70% ethanol. Ethanol was
727 removed and the pellets were air dried on a bench top prior to resuspending them in either 54 ul
728 (pulldown) or 108 ul (flow-through) of water.
729
730 **Sequencing library preparation:** For each sample, a 480 ul PCR reaction was prepared
731 containing: 240 ul of Q5 2x master mix, 168 ul of water, 24 ul of nucleus DNA extraction, and 24
732 ul of a 10uM stock of each primer. Primers were designed to anneal to regions immediately
733 adjacent to the tile sequences and contained Nextera-style sequencing adapters and 8 bp
734 indexes. Throughout this work we used the same number of PCR cycles for all samples, which
735 was determined as follows. We ran qPCR reactions using the pulldown and flow-through DNA
736 samples of 2 replicates of the viral 250 tile library, using the same set of primers used for preparing
737 sequencing libraries. The qPCR Ct value of these reactions was 22-23 cycles. We added 2 more
738 cycles to obtain extra DNA for sequencing, resulting in 25 cycles used for library preparation PCR.
739 The library PCR were run on a 2% agarose gel and bands of the expected size were excised and
740 purified using a Zymo gel purification column following manufacturer instructions. Libraries were
741 eluted in 50 ul of water and run in a fragment analyzer to confirm that the size distribution matched
742 the expected amplicon size. Whenever necessary libraries were size selected using SPRI bead
743 cleanup. Samples from the 26 control strains experiments shown in Figure 2 were PCR amplified
744 using the same protocol except that primers did not contain any extensions. Adapters were added
745 to the purified PCR products using the Nanopore Native Barcoding Kit 96 V14 (SQK-NBD114.96).
746

DNA sequencing: The libraries prepared from the control set of 26 mixed strains shown in Figure 2 were sequenced using the entire whole flow cell of a Nanopore minION for each replicate. The rest of the samples were run on an Illumina NextSeq 2000 for 100 cycles using 30% PhiX and single-end reads covering the 5' end of the sequencing library amplicon.

Mapping reads to tiles: The 2FAST2Q[56] package version 2.6.0 was run using the following parameters. Feature length = 76 bp, allowed mismatches = 2, minimal Phred score = 20, upstream search sequence = TGGATCTGGAGGT, mismatches in upstream sequence = 0, minimal upstream sequence Phred score = 30.

Calculation of tile enrichment: For each tile, the number of mapped reads in the pulldown sample was divided by the total number of reads mapped to all the library tiles in this sample to obtain a number corresponding to the proportion of pulldown reads mapping to each specific tile. The same calculation was performed for each tile in the flowthrough sample. The ratio between these two values corresponds to the tile enrichment.

Calculation of tile coverage: We define tile coverage as the percentage of all reads in both pulldown and flowthrough samples that map to a given tile**.** To obtain this percentage the number of mapped reads in the pulldown and flowthrough samples are added and the reads that match each tile are divided by the total number of reads.

Designing viral libraries: A total of 1,495 plant virus genomes were downloaded from NCBI, yielding 5,607 unique proteins. These proteins were clustered at 90% sequence identity using CD-HIT[57], resulting in 4,537 representative proteins. Each protein was fragmented into 53 aa tiles with a 10 aa stride, generating 190,520 tiles. PADDLE was used to perform inference on each tile. To construct a diverse subset for experimental screening, the top 2,000 tiles with the highest predicted z-score were binary encoded, and a greedy selection algorithm was applied to maximize the pairwise Hamming distance between selected encodings. This approach prioritized sequences with the most dissimilar binary representations, thereby enhancing diversity in the DNA library.

Machine-based predictions of activation domains: In addition to PADDLE, TADA was used to predict transactivation strength, and Metapredict[58] was used to identify disorder regions in proteins. Because PADDLE and TADA use different input sizes (53 aa and 40 aa, respectively), CONSTANS and CO variants were tiled with a 1 aa stride, and tile-level predictions were aggregated into residue-level profiles by averaging the scores of all overlapping tiles at each position. Residues not covered by any tile were assigned a score of zero. PADDLE and TADA predictions were independently min–max normalized to scale scores between 0 and 1 for comparison between models.

Calibration of ENTRAP-seq from fluorescence data: To model the relationship between ENTRAP-seq and fluorescence measurements, we fit a sigmoid function to the enrichment ratios as a function of logRFP for control tiles. The standard deviation of residuals from the fitted curve

790 was used to estimate confidence bounds. We defined the activation threshold as the upper
791 confidence bound at the midpoint of the sigmoidal fit.
792
793 **References**
794

795 1. Borowsky, A. T. & Bailey-Serres, J. Rewiring gene circuitry for plant improvement. *Nat.*
796 *Genet.* 56, 1574–1582 (2024).
797 2. Gonzalez, D. H. *Plant Transcription Factors: Evolutionary, Structural and Functional*
798 *Aspects*. 434 (Academic Press, 2015).
799 3. Udupa, A., Kotha, S. R. & Staller, M. V. Commonly asked questions about transcriptional
800 activation domains. *Curr. Opin. Struct. Biol.* 84, 102732 (2024).
801 4. Soto, L. F. *et al.* Compendium of human transcription factor effector domains. *Mol. Cell* 82,
802 514–526 (2022).
803 5. Ferrie, J. J., Karr, J. P., Tjian, R. & Darzacq, X. “Structure”-function relationships in
804 eukaryotic transcription factors: The role of intrinsically disordered regions in gene
805 regulation. *Mol. Cell* 82, 3970–3984 (2022).
806 6. Staller, M. V. *et al.* A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic
807 Transcriptional Activation Domain. *Cell Syst.* 6, 444-455.e6 (2018).
808 7. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-
809 seq. *Science* 339, 1074–1077 (2013).
810 8. O’Malley, R. C. *et al.* Cistrome and epicistrome features shape the regulatory DNA
811 landscape. *Cell* 165, 1280–1292 (2016).
812 9. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence-
813 Function Relationships. *Annu. Rev. Genomics Hum. Genet.* 20, 99–127 (2019).
814 10. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to
815 characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl*
816 *Acad Sci USA* 107, 9158–9163 (2010).
817 11. Erijman, A. *et al.* A High-Throughput Screen for Transcription Activation Domains Reveals
818 Their Sequence Features and Permits Prediction by Deep Learning. *Mol. Cell* 78, 890-
819 902.e6 (2020).
820 12. Sanborn, A. L. *et al.* Simple biochemical features underlie transcriptional activation domain
821 diversity and dynamic, fuzzy binding to Mediator. *eLife* 10, (2021).
822 13. Morffy, N. *et al.* Identification of plant transcriptional activation domains. *Nature* 632, 166–
823 173 (2024).
824 14. Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *Curr. Opin.*
825 *Struct. Biol.* 72, 145–152 (2022).
826 15. Arnold, C. D. *et al.* A high-throughput method to identify trans-activation domains within
827 transcription factor sequences. *EMBO J.* 37, (2018).
828 16. Ravarani, C. N. *et al.* High-throughput discovery of functional disordered regions:
829 investigation of transactivation domains. *Mol. Syst. Biol.* 14, e8190 (2018).
830 17. Hummel, N. F. C., Markel, K., Stefani, J., Staller, M. V. & Shih, P. M. Systematic
831 identification of transcriptional activation domains from non-transcription factor proteins in
832 plants and yeast. *Cell Syst.* 15, 662-672.e4 (2024).
833 18. Tycko, J. *et al.* High-Throughput Discovery and Characterization of Human Transcriptional

Effectors. *Cell* 183, 2020-2035.e16 (2020).
19. DelRosso, N. *et al.* Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* 616, 365–372 (2023).
20. Ludwig, C. H. *et al.* High-throughput discovery and characterization of viral transcriptional effectors in human cells. *Cell Syst.* 14, 482-500.e8 (2023).
21. Alcantar, M. A., English, M. A., Valeri, J. A. & Collins, J. J. A high-throughput synthetic biology approach for studying combinatorial chromatin-based transcriptional regulation. *Mol. Cell* 84, 2382-2396.e9 (2024).
22. Voichek, Y., Hristova, G., Mollá-Morales, A., Weigel, D. & Nordborg, M. Widespread position-dependent transcriptional regulatory sequences in plants. *Nat. Genet.* 56, 2238–2246 (2024).
23. Arndell, T. *et al.* Pooled effector library screening in protoplasts rapidly identifies novel Avr genes. *Nat. Plants* 10, 572–580 (2024).
24. Alamos, S. *et al.* Quantitative dissection of Agrobacterium T-DNA expression in single plant cells reveals density-dependent synergy and antagonism. *Nat. Plants* (2025) doi:10.1038/s41477-025-01996-w.
25. Wang, D. & Deal, R. B. Epigenome profiling of specific plant cell types using a streamlined INTACT protocol and ChIP-seq. *Methods Mol. Biol.* 1284, 3–25 (2015).
26. Deal, R. B. & Henikoff, S. The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nat. Protoc.* 6, 56–68 (2011).
27. Hummel, N. F. C. *et al.* The trans-regulatory landscape of gene networks in plants. *Cell Syst.* 14, 501-511.e4 (2023).
28. Reynoso, M. A. *et al.* Nuclear transcriptomes at high resolution using retooled INTACT. *Plant Physiol.* 176, 270–281 (2018).
29. Belcher, M. S. *et al.* Design of orthogonal regulatory systems for modulating gene expression in plants. *Nat. Chem. Biol.* 16, 857–865 (2020).
30. Cai, Y.-M. *et al.* Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res.* 48, 11845–11856 (2020).
31. Tehrani, N. & Mitra, R. M. Plant pathogens and symbionts target the plant nucleus. *Curr. Opin. Microbiol.* 72, 102284 (2023).
32. Boch, J., Bonas, U. & Lahaye, T. TAL effectors--pathogen strategies and plant resistance engineering. *New Phytol.* 204, 823–832 (2014).
33. Hull, R. *Plant Virology*. 1118 (Academic Press, 2013).
34. Liu, X. *et al.* Human virus transcriptional regulators. *Cell* 182, 24–37 (2020).
35. Guerrero, J. *et al.* Manipulation of the plant host by the geminivirus AC2/C2 protein, a central player in the infection cycle. *Front. Plant Sci.* 11, 591 (2020).
36. Wang, L. & Lozano-Durán, R. Manipulation of plant RNA biology by geminiviruses. *J. Exp. Bot.* 74, 2311–2322 (2023).
37. Hartitz, M. D., Sunter, G. & Bisaro, D. M. The tomato golden mosaic virus transactivator (TrAP) is a single-stranded DNA and zinc-binding phosphoprotein with an acidic activation domain. *Virology* 263, 1–14 (1999).
38. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* 12, 326–328 (2015).
39. Zinselmeier, M. H. *et al.* Optimized dCas9 programmable transcriptional activators for

878 plants. *Plant Biotechnol. J.* 22, 3202–3204 (2024).
879 40. Atanesyan, L., Günther, V., Dichtl, B., Georgiev, O. & Schaffner, W. Polyglutamine tracts as
880 modulators of transcriptional activation from yeast to mammals. *Biol. Chem.* 393, 63–70
881 (2012).
882 41. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Activation and repression of transcription by auxin-
883 response factors. *Proc Natl Acad Sci USA* 96, 5844–5849 (1999).
884 42. Cho, S. *et al.* Analysis of the C-terminal region of Arabidopsis thaliana APETALA1 as a
885 transcription activation domain. *Plant Mol. Biol.* 40, 419–429 (1999).
886 43. Ernst, H. A., Olsen, A. N., Larsen, S. & Lo Leggio, L. Structure of the conserved domain of
887 ANAC, a member of the NAC family of transcription factors. *EMBO Rep.* 5, 297–303 (2004).
888 44. Olsen, A. N., Ernst, H. A., Leggio, L. L. & Skriver, K. NAC transcription factors: structurally
889 distinct, functionally diverse. *Trends Plant Sci.* 10, 79–87 (2005).
890 45. Brent, R. & Ptashne, M. A eukaryotic transcriptional activator bearing the DNA specificity of
891 a prokaryotic repressor. *Cell* 43, 729–736 (1985).
892 46. Rosas, U. *et al.* Variation in Arabidopsis flowering time associated with cis-regulatory
893 variation in CONSTANS. *Nat. Commun.* 5, 3651 (2014).
894 47. Takahashi, Y., Teshima, K. M., Yokoi, S., Innan, H. & Shimamoto, K. Variations in Hd1
895 proteins, Hd3a promoters, and Ehd1 expression levels contribute to diversity of flowering
896 time in cultivated rice. *Proc Natl Acad Sci USA* 106, 4555–4560 (2009).
897 48. Tan, J. *et al.* OsCOL10, a CONSTANS-Like Gene, Functions as a Flowering Time
898 Repressor Downstream of Ghd7 in Rice. *Plant Cell Physiol.* 57, 798–812 (2016).
899 49. Steinbach, Y. The Arabidopsis thalianaCONSTANS-LIKE 4 (COL4) - A Modulator of
900 Flowering Time. *Front. Plant Sci.* 10, 651 (2019).
901 50. Nemoto, Y., Nonoue, Y., Yano, M. & Izawa, T. Hd1,a CONSTANS ortholog in rice, functions
902 as an Ehd1 repressor through interaction with monocot-specific CCT-domain protein Ghd7.
903 *Plant J.* 86, 221–233 (2016).
904 51. Tiwari, S. B. *et al.* The flowering time regulator CONSTANS is recruited to the FLOWERING
905 LOCUS T promoter via a unique cis-element. *New Phytol.* 187, 57–66 (2010).
906 52. Robson, F. *et al.* Functional importance of conserved domains in the flowering-time gene
907 CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. *Plant J.* 28,
908 619–631 (2001).
909 53. Staller, M. V. *et al.* Directed mutational scanning reveals a balance between acidic and
910 hydrophobic residues in strong human activation domains. *Cell Syst.* 13, 334-345.e5 (2022).
911 54. Sigler, P. B. Transcriptional activation. Acid blobs and negative noodles. *Nature* 333, 210–
912 212 (1988).
913 55. Weigel, D. & Glazebrook, J. *Arabidopsis: A Laboratory Manual*. 448 (Cold Spring Harbor
914 Laboratory Press, 2002).
915 56. Bravo, A. M., Typas, A. & Veening, J.-W. 2FAST2Q: a general-purpose sequence search
916 and counting program for FASTQ files. *PeerJ* 10, e14041 (2022).
917 57. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
918 or nucleotide sequences. *Bioinformatics* 22, 1658–1659 (2006).
919 58. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-
920 use predictor of consensus disorder and structure. *Biophys. J.* 120, 4312–4319 (2021).

Main Figures

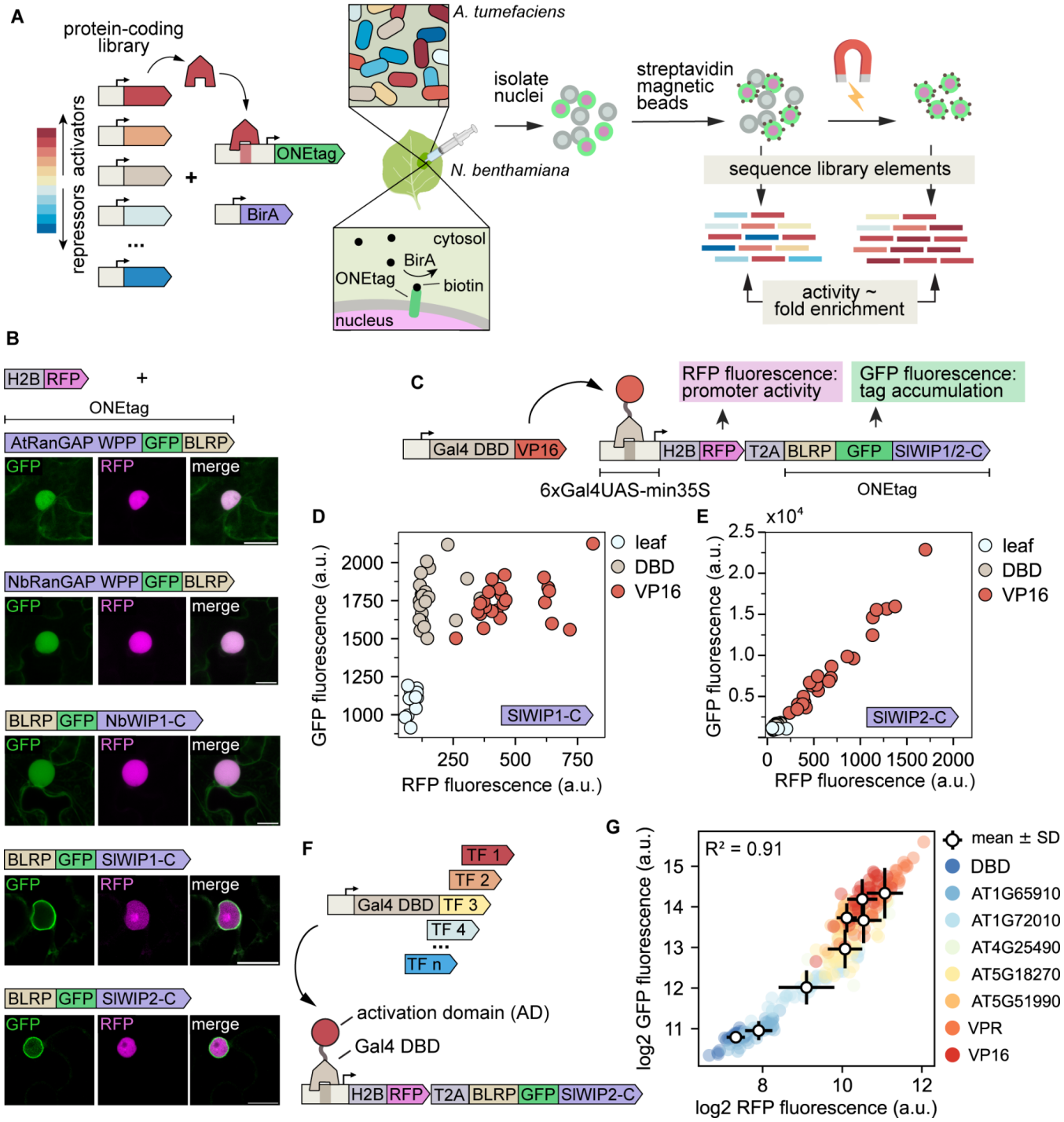


Figure 1: Design and optimization of an assay to measure the activity of *trans*-acting elements in parallel in plant cells. (A) Overview of the ENTRAP-seq assay. A library of n different *trans*-acting elements such as transcription factors (TFs) of varying strengths is cloned in binary vectors for *Agrobacterium*-mediated transformation. These *trans*-elements regulate a

929 promoter driving the expression of an outer nuclear envelope tag (ONEtag) containing a biotin
930 acceptor peptide. A transgene for the constitutive expression of the biotin ligase BirA is also
931 included to biotinylate the ONEtag. The library as well as the ONEtag reporter and BirA are
932 transformed into *Agrobacterium* and mixed cultures are infiltrated into leaves of *N. benthamiana*.
933 The ONEtag is then biotinylated by constitutively expressed BirA using endogenous biotin. Next,
934 nuclei are isolated from infiltrated leaves and incubated with magnetic streptavidin-coated beads,
935 allowing magnetic sorting. DNA is extracted from the unsorted and sorted nuclei samples and the
936 abundance of each of the *trans*-elements in the library is determined by sequencing the T-DNA.
937 The enrichment of a given *trans*-element in the sorted vs unsorted population correlates with its
938 activation strength. **(B)** Live cell fluorescence microscopy images of *N. benthamiana* epidermis
939 cells expressing a nuclear marker (H2B-RFP) and different ONEtag candidate versions fused to
940 GFP. Bar = 20 um. Shown on top of each set of images is a schematic of the ONEtag variant
941 tested. **(C)** Schematic of the construct used in (D) and (E) to compare promoter activity (RFP
942 fluorescence) with ONEtag accumulation (GFP fluorescence). **(D)** Scatter plot showing RFP (x
943 axis) and GFP (y axis) fluorescence of leaf punches expressing the construct in (C) containing
944 the SlWIP1-C ONEtag combined with a second strain carrying Gal4DBD or Gal4-VP16. **(E)** Same
945 as (D) for a ONEtag containing SlWIP2-C. **(F)** Schematic of the experiment to test the sensitivity
946 of the SlWIP2-C ONE tag. Gal4DBD fusions to effector domains of Arabidopsis TFs of varying
947 strengths were used to drive the expression of the reporter. **(G)** Scatter plot of the GFP (x axis)
948 and RFP (y axis) fluorescence expressing the constructs in (F). In (D), (E), and (G) N = 24 (3
949 plants with 8 leaf punches per plant).

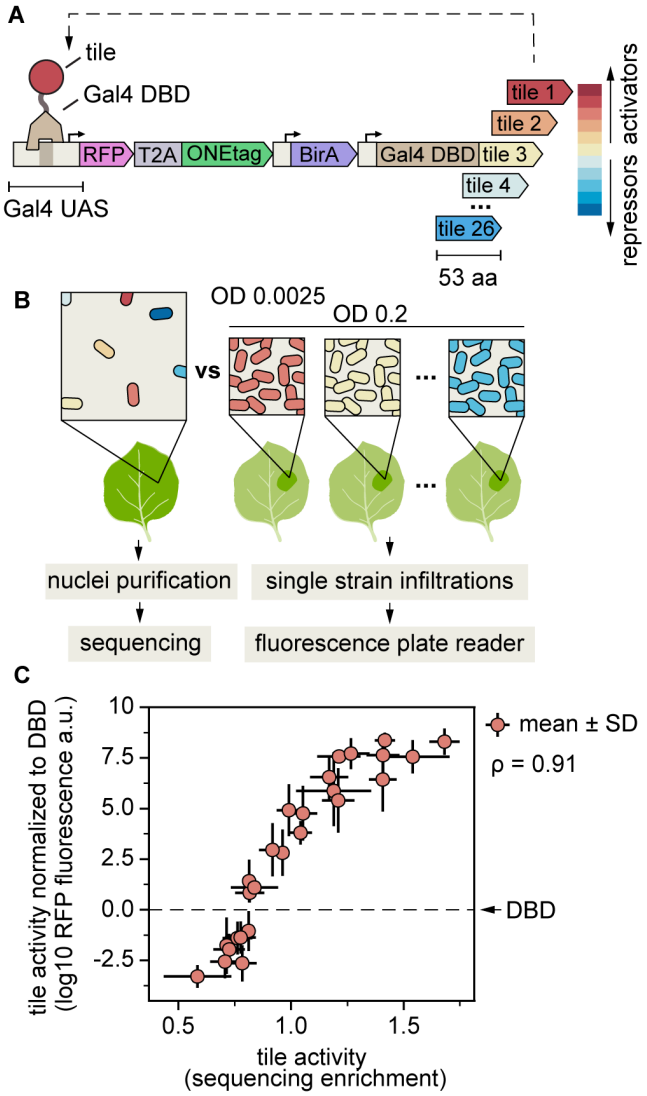


Figure 2: Assay validation comparing pooled sequencing with single-infiltration reporter fluorescence. (A) T-DNA constructs used to validate the parallel reporter assay**.** A synthetic promoter containing 6x Gal4 UAS sites followed by the minimal 35S promoter is used to drive the ONEtag fused to RFP by a self-cleaving T2A peptide. Downstream of the ONEtag, the plasmid contains a constitutive BirA expression cassette and a constitutively-expressed fusion between the Gal4 DBD and variable 53 aa tiles of various strengths. **(B)** Experimental strategy. Left: strains carrying the 26 constructs from (A) are grown separately and mixed in equal OD ratios prior to leaf infiltration at a final total OD of 0.0025. ENTRAP-seq is then used to determine the activation strength of each tile based on its enrichment in the pulldown fraction. Right: each of the strains in (A) is infiltrated individually at an OD of 0.2. The activity of each tile is determined by measuring the RFP fluorescence of infiltrated leaves in a plate reader. **(C)** Y-axis: Tile transcriptional activity measured individually using a fluorescence plate reader. X-axis: transcriptional activity of each tile based on ENTRAP-seq (enrichment of each tile T-DNA in the nuclei pulldown fraction compared to the flow-through sample). N for nuclei pulldown and sequencing = 4 biological

966 replicates. N for plate reader assays = 24 (3 plants with 8 leaf punches per plant). ρ = Spearman
967 correlation coefficient.

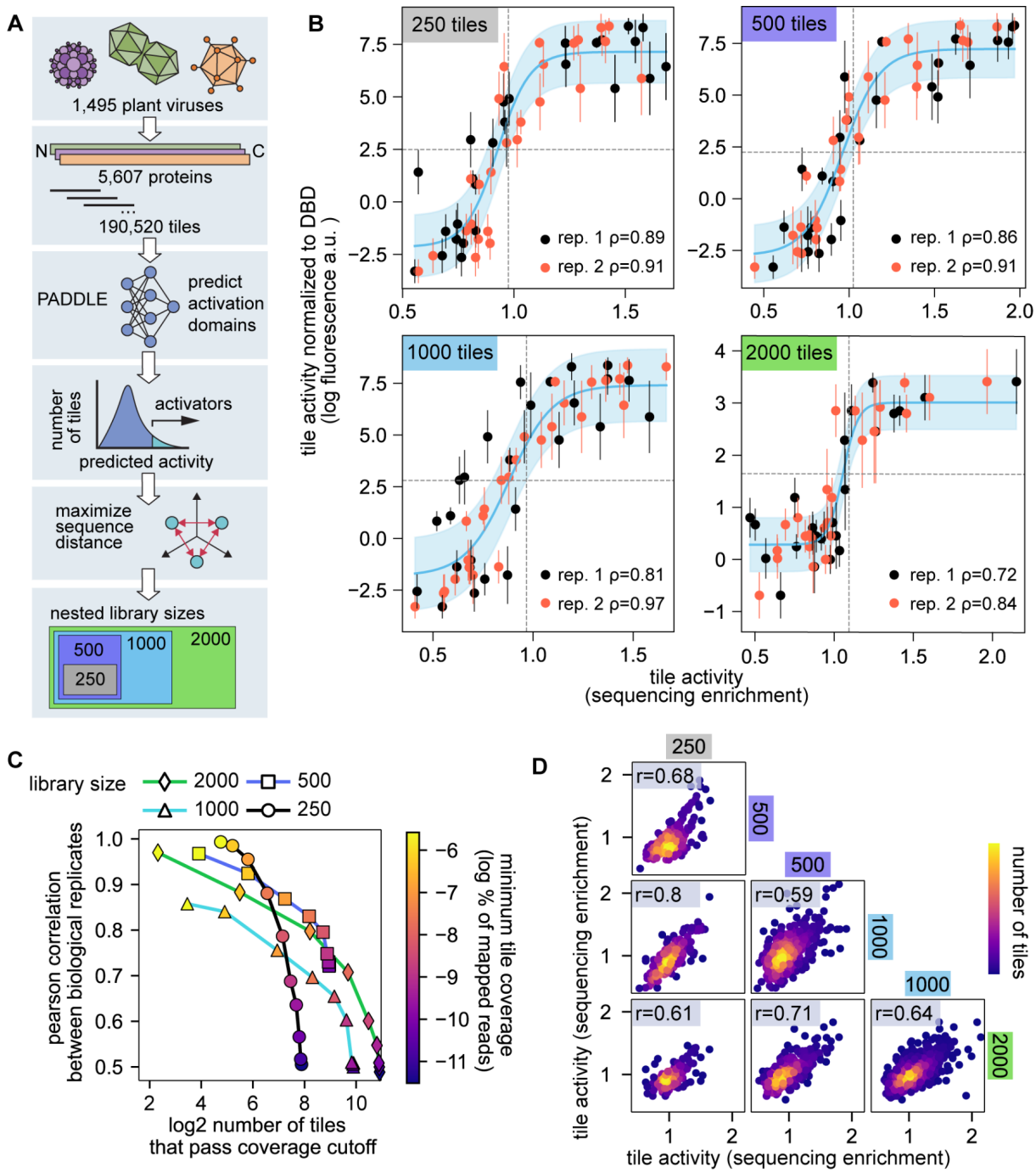
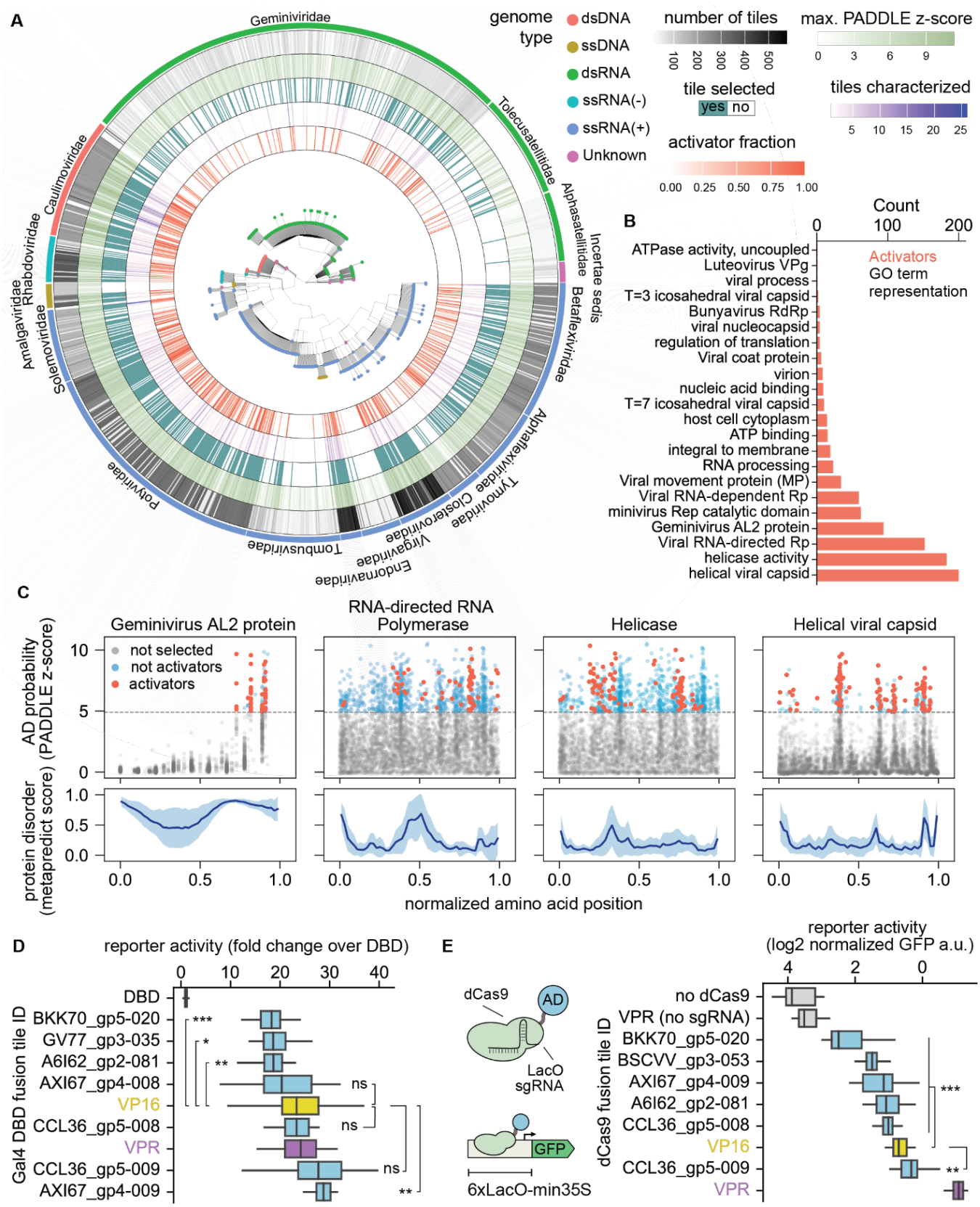


Figure 3: Generation of nested libraries of putative viral activation domains to test assay scale and reproducibility. (A) Schematic of the computational pipeline to generate libraries of putative transcriptional activation domains from plant viruses. From the top: protein ORF sequences from genomes of all plant viruses in NCBI were deduplicated and clustered using 90% sequence identity then fragmented into 53 aa tiles with 10 bp overlap. PADDLE was used to predict the transactivation strength of these tiles. The top 2,000 tiles with the highest predicted activity were selected (PADDLE z-score > 4.9). Protein sequences were binary encoded, and sequence diversity maximization algorithm was applied to iteratively select sequences that

maximized the average Hamming distance from those already included in the subset. This approach enriched for sequence-level diversity in the selected proteins to generate nested libraries of 250, 500, 1K, and 2K tiles. **(B)** Tile activity as measured by ENTRAP-seq (x axis) plotted against the mean ± 1 s.d. tile activity measured individually using a fluorescence reporter assay (y axis) for two biological replicates of a library of 250, 500, 1K, and 2K viral tiles. Blue lines show a sigmoidal fit to the data with the shaded blue area indicating ± 1 s.d. of residuals as an empirical confidence bound. Horizontal dashed lines indicate the middle point of the sigmoid fits. Vertical dashed lines indicate the enrichment value corresponding to the sigmoid midpoint + 1 s.d. This value was taken as the lower bound of enrichment to classify tiles as activators. **(C)** Reproducibility between biological replicates (Pearson correlation in enrichment, y axis) as a function of the absolute number of tiles that pass a threshold number of mapped reads (x axis). The set of tiles counted on the x axis was generated by imposing coverage cutoffs such that only tiles with more counts than an arbitrary and variable threshold (color bar) were considered to calculate the Pearson correlation. **(D)** Reproducibility across different libraries. The activity of each tile was first averaged across both biological replicates of each library of a given size. Scatter plots show the activity of tiles that are shared between libraries of different sizes. The Pearson correlation (r) between libraries is shown.


A
Geminiviridae
Tolecusatellitidae
Alphasatellitidae
Incertae sedis
Betaflexiviridae
Alphaflexiviridae
Tymoviridae
Closteroviridae
Virgaviridae
Endornaviridae
Tombusviridae
Polyviridae
Solemoviridae
Amalgaviridae
Rhabdoviridae
Caulimoviridae
genome type
dsDNA
ssDNA
dsRNA
ssRNA(-)
ssRNA(+)
Unknown
number of tiles
100 200 300 400 500
tile selected
yes no
activator fraction
0.00 0.25 0.50 0.75 1.00
max. PADDLE z-score
0 3 6 9
tiles characterized
5 10 15 20 25
B
Count
0 100 200
Activators
GO term
representation
ATPase activity, uncoupled
Luteovirus VPg
viral process
T=3 icosahedral viral capsid
Bunyavirus RdRp
viral nucleocapsid
regulation of translation
Viral coat protein
virion
nucleic acid binding
T=7 icosahedral viral capsid
host cell cytoplasm
ATP binding
integral to membrane
RNA processing
Viral movement protein (MP)
Viral RNA-dependent Rp
minivirus Rep catalytic domain
Geminivirus AL2 protein
Viral RNA-directed Rp
helicase activity
helical viral capsid
C
Geminivirus AL2 protein
RNA-directed RNA Polymerase
Helicase
Helical viral capsid
not selected
not activators
activators
AD probability (PADDLE z-score)
protein disorder (metapredict score)
normalized amino acid position
D
reporter activity (fold change over DBD)
Gal4 DBD fusion tile ID
DBD
BKK70_gp5-020
GV77_gp3-035
A6I62_gp2-081
AXI67_gp4-008
VP16
CCL36_gp5-008
VPR
CCL36_gp5-009
AXI67_gp4-009
***
*
**
ns
ns
ns
**
E
dCas9
AD
LacO sgRNA
GFP
6xLacO-min35S
reporter activity (log2 normalized GFP a.u.)
dCas9 fusion tile ID
no dCas9
VPR (no sgRNA)
BKK70_gp5-020
BSCVV_gp3-053
AXI67_gp4-009
A6I62_gp2-081
CCL36_gp5-008
VP16
CCL36_gp5-009
VPR
***
**

997 **Figure 4: Conserved and potent transcriptional activation domains are found in diverse**
998 **proteins across plant virus taxa. (A)** Phylogenetic tree of plant viruses showing the number of
999 tiles selected from each viral genome (gray), the maximum PADDLE predicted z-score (light
1000 green), whether a tile from this viral genome was selected for characterization (dark green), the
1001 number of tiles that were experimentally measured (purple), and the proportion of measured tiles
1002 that were activators (orange). **(B)** Gene ontology (GO) enrichment analysis of tiles measured as
1003 activators across all libraries. **(C)** Detailed view of tiles from the proteins most enriched in activator
1004 tiles in (B)**.** Top: AD prediction (PADDLE z-score) of each tile along the length of all proteins
1005 belonging to each given class. The amino acid position of each tile was normalized by dividing by
1006 the length of the protein it originates from. The horizontal dashed line shows the z-score cutoff
1007 used to select tiles for testing. Blue dots show tiles that were experimentally tested but did not
1008 pass our activator enrichment cutoff. Orange dots show tiles that were tested and had an
1009 enrichment higher than the cutoff. Bottom: average ± 1 s.d. protein disorder as predicted by
1010 Metapredict. **(D)** Selected strong activator viral tiles tested individually in *N. benthamiana* using a
1011 Gal4DBD recruitment fluorescence reporter assay. Shown is the GFP fluorescence normalized
1012 to the Gal4DBD alone (Materials and Methods: Fluorescence plate reader reporter assays). **(E)**
1013 A set of tiles tested in (D) were tested by recruiting them to a synthetic reporter driving GFP using
1014 dCas9 fusions**.** Shown is the normalized leaf GFP fluorescence. In (D) and (E) brackets indicate
1015 two-sided t-test comparisons against VP16 with a Bonferroni multiple comparison correction. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns not significant. In (D) and (E) box plots show the median ±
1017 IQR, whiskers correspond to minimum and maximum values excluding outliers.

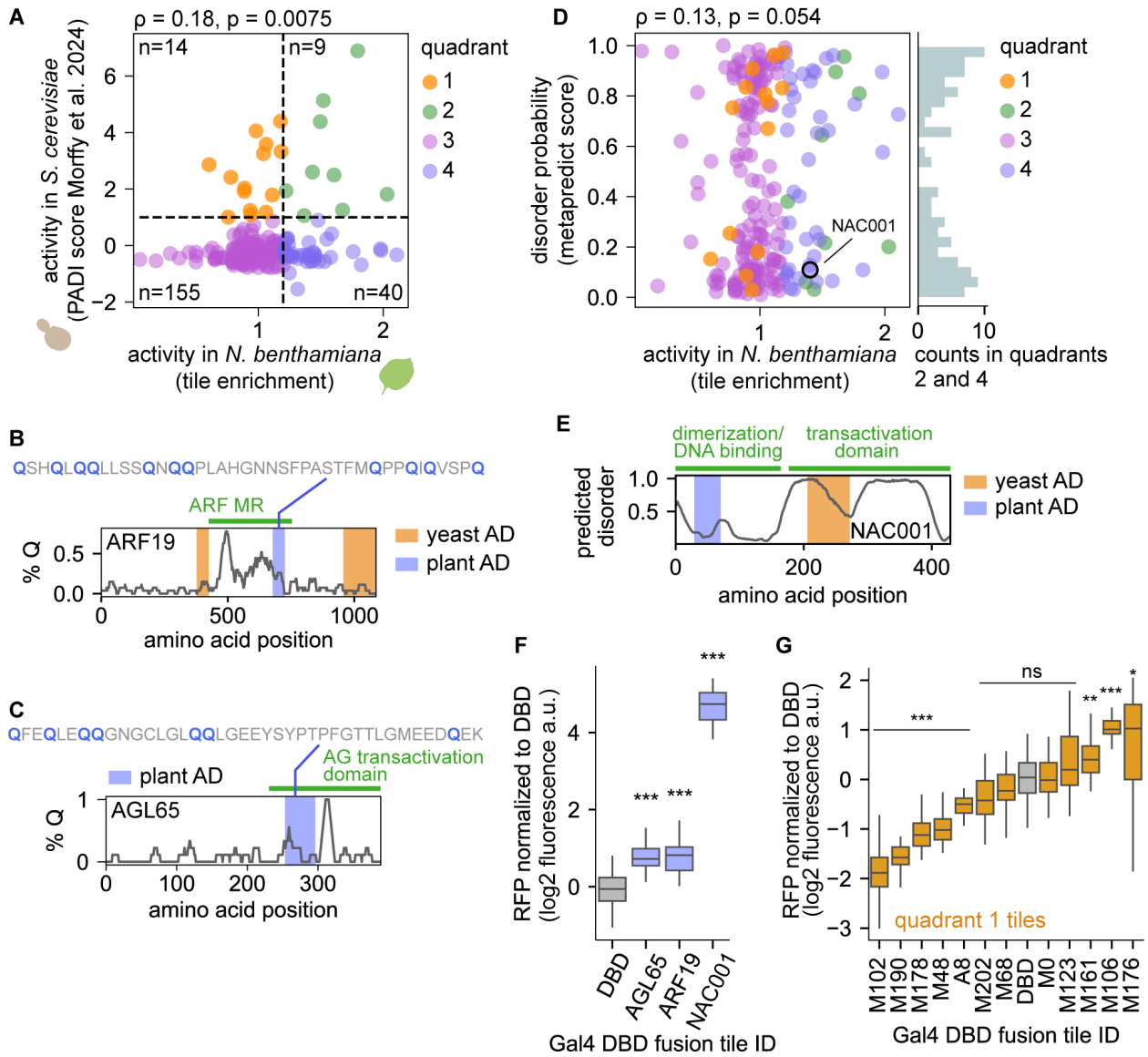


Figure 5: Systematic comparison of tiles from *Arabidopsis* TFs expressed in yeast and *N. benthamiana*. (A) Comparison of tile transcriptional activity between yeast and *N. benthamiana*. A set of peptide tiles from *Arabidopsis* TFs previously measured in yeast was tested in *N. benthamiana*. Scatterplot shows the activity of each tile in yeast (PADI score, y axis) and *N. benthamiana* (tile enrichment, x axis). The horizontal dashed line indicates the PADI threshold value used to classify tiles as activators in yeast. The vertical dashed line shows the enrichment value below which we never detected validated activator tiles in Fig. 3B. **(B)** Glutamine enrichment along the ARF19 TF protein shown as percentage of Q residues in a moving window of 20 aa. The green bar indicates the position of the ARF activation domain known as middle region (MR). Orange bars show the position of ADs according to the yeast assay. The blue bar shows the position of the AD in plants according to ENTRAP-seq. **(C)** Same as (B) for the ALG65 TF. No ADs were identified for this TF in yeast. **(D)** The degree of structural disorder of each tile is plotted against its ENTRAP-seq activity in *N. benthamiana*. The histogram on the right includes tiles likely

1033 to be activators in both yeast and *N. benthamiana* as well as *N. benthamiana* only. **(E)** Predicted
1034 disorder in the NAC001 TF. The green bar indicates the NAC dimerization domain. The blue bar
1035 shows the AD identified by ENTRAP-seq and the orange bar shows the position of the ADs active
1036 in yeast. **(F)** Validation of the tiles highlighted in blue in (B), (C), and (E) using a fluorescence
1037 reporter assay. **(G)** A random selection of tiles that were activators in yeast but de-enriched in
1038 tobacco ('quadrant 1' in (A)) tested in a fluorescence reporter assay against the Gal4DBD alone
1039 as a control. In (F) and (G) * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns not significant in a two-sided
1040 t-test against Gal4DBD. In (F) and (G) box plots show the median ± IQR, whiskers correspond to
1041 minimum and maximum values excluding outliers.

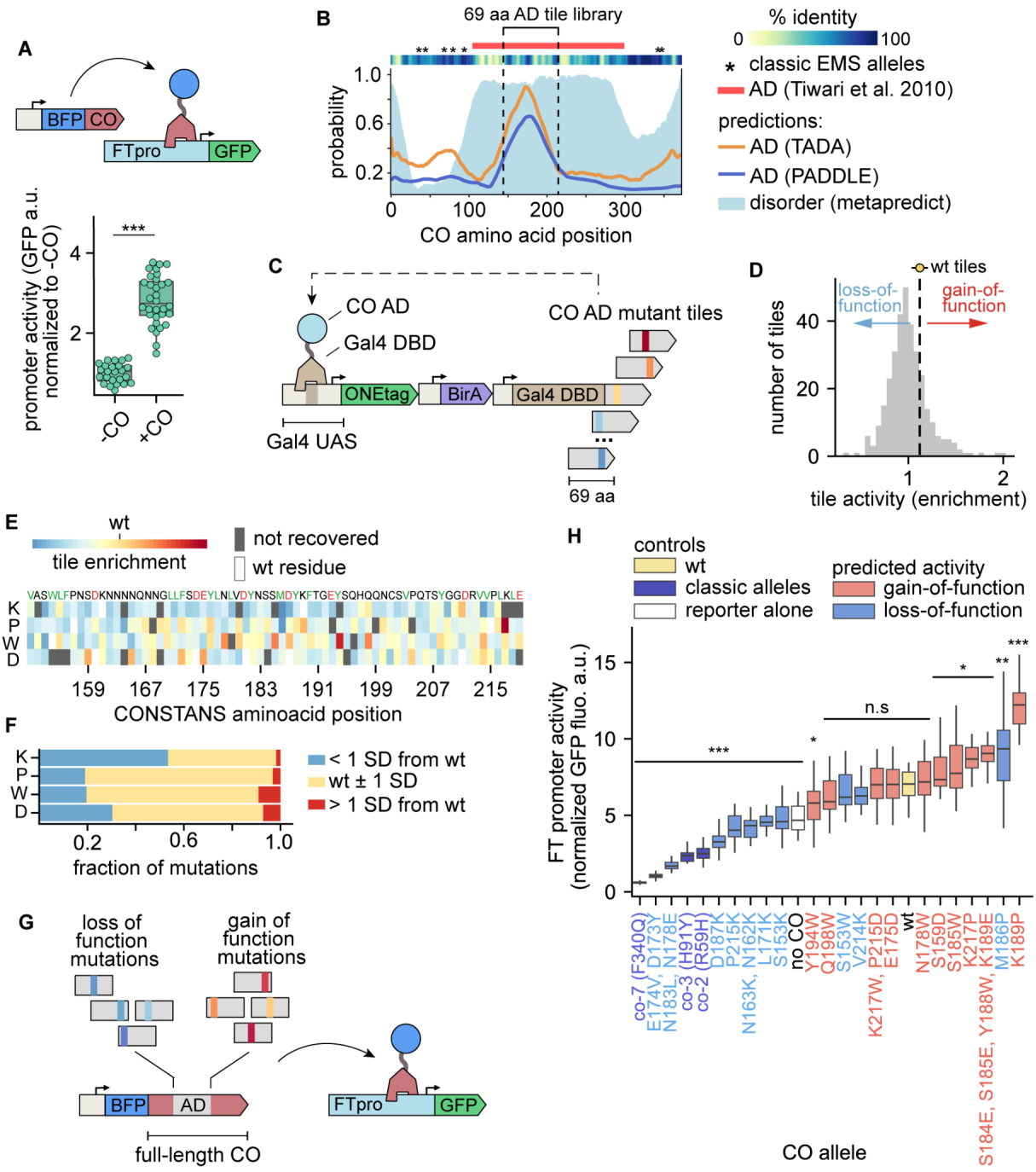


Figure 6: Large scale screening of mutant activation domains to create an allelic series of the CO transcription factor. (A) Transactivation assay for CO function in *N. benthamiana*. A strain carrying GFP driven by the *Arabidopsis* FT promoter (FTpro) was infiltrated with a strain carrying pUBQ-BFP-CO or an empty vector (EV). pUBQ-mCherry was used as a normalization. Box plot shows the mean ± IQR, whiskers are minimum and maximum values. Shown p is from a two-sided t-test. **(B)** Location of the CO activation domain (AD). Plot shows the probability of residues being part of an acidic AD according to the ADpred software. Bars on top show the empirically determined AD. The blue box indicates the 69 aa region used to build a tile library. **(C)**

1052 Screening of CO the AD tile library using nuclei enrichment. Mutant tiles are fused to the Gal4
1053 DBD and used to drive the expression of the magnetic sorting reporter. **(D)** Distribution of tile
1054 enrichment values. The dashed line shows the mean enrichment of 6 wt tiles. **(E)** Tile activity as
1055 a function of biochemical properties. Each dot corresponds to one tile. The dashed red lines are
1056 linear fits. Shown on top is the correlation coefficient (R2) and the p-value of the linear regression.
1057 **(F)** Heatmap showing the activity of tiles carrying single aa substitutions. **(G)** Transactivation
1058 assay to test the effect of mutations in the CO AD in the context of the full-length protein. Mutations
1059 identified in (C)-(F) as being weaker (loss-of-function) or stronger (gain-of-function) than the wt
1060 AD were introduced into the CO coding sequence and used to drive expression of a FTpro GFP
1061 reporter. **(H)** FT promoter activity in the presence of various CO mutant alleles. Predicted activity
1062 is based on the enrichment of the corresponding AD tile with respect to the wt tile. Box plots show
1063 the median ± IQR, bars correspond to minimum and maximum values excluding outliers. *** $p <$
1064 0.001, ** $p < 0.01$, * $p < 0.05$, n.s not significantly different using a t-test against wt with a
1065 Bonferroni multiple comparison correction.

Supplementary Figures

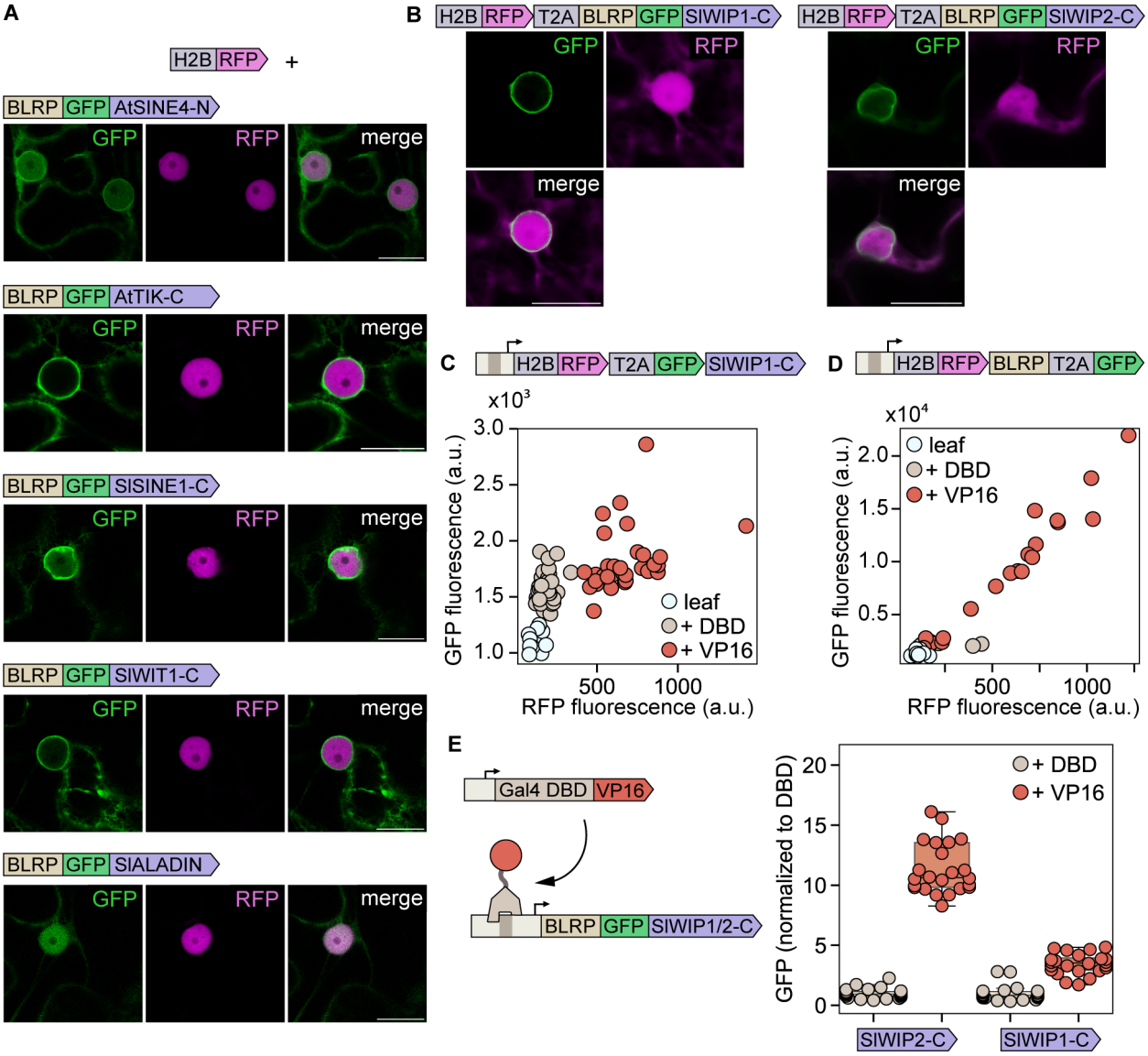


Figure S1: Additional ONEtag optimization details. (A) Fluorescence microscopy images showing the subcellular localization of ONEtag variants tagged with GFP and coexpressed with an RFP nuclear marker. From top to bottom the corresponding plasmid names are pSA266, pSA267, pSA268, pSA270, pSA272. **(B)** Cells expressing the SlWIP1-C and SlWIP2-C ONEtag fused to RFP by a self-cleaving T2A peptide (plasmids pSA271 and pSA91, respectively). **(C)** Top: schematic of the construct. BLRP was removed from the construct shown in Fig. 1C and the responsiveness of this reporter to VP16 was tested in a fluorescence plate reader. **(D)** Top: schematic of the construct. SlWIP1-C was removed from the construct shown in Fig. 1C and the responsiveness of this reporter to VP16 was tested in a fluorescence plate reader. In (C) and (D) each circle corresponds to one leaf punch. **(E)** Left: schematic of the experiment. The RFP-T2A tag was removed from both the SlWIP1 and SlWIP2 constructs shown in Fig. 1C and their

1081 responsiveness to VP16 was tested in a fluorescence plate reader (plasmids SA752 and pSA753,
1082 respectively). Right: leaf GFP fluorescence with or without VP16 driving reporter expression. Box
1083 plots show the median ± IQR, whiskers correspond to minimum and maximum values excluding
1084 outliers.

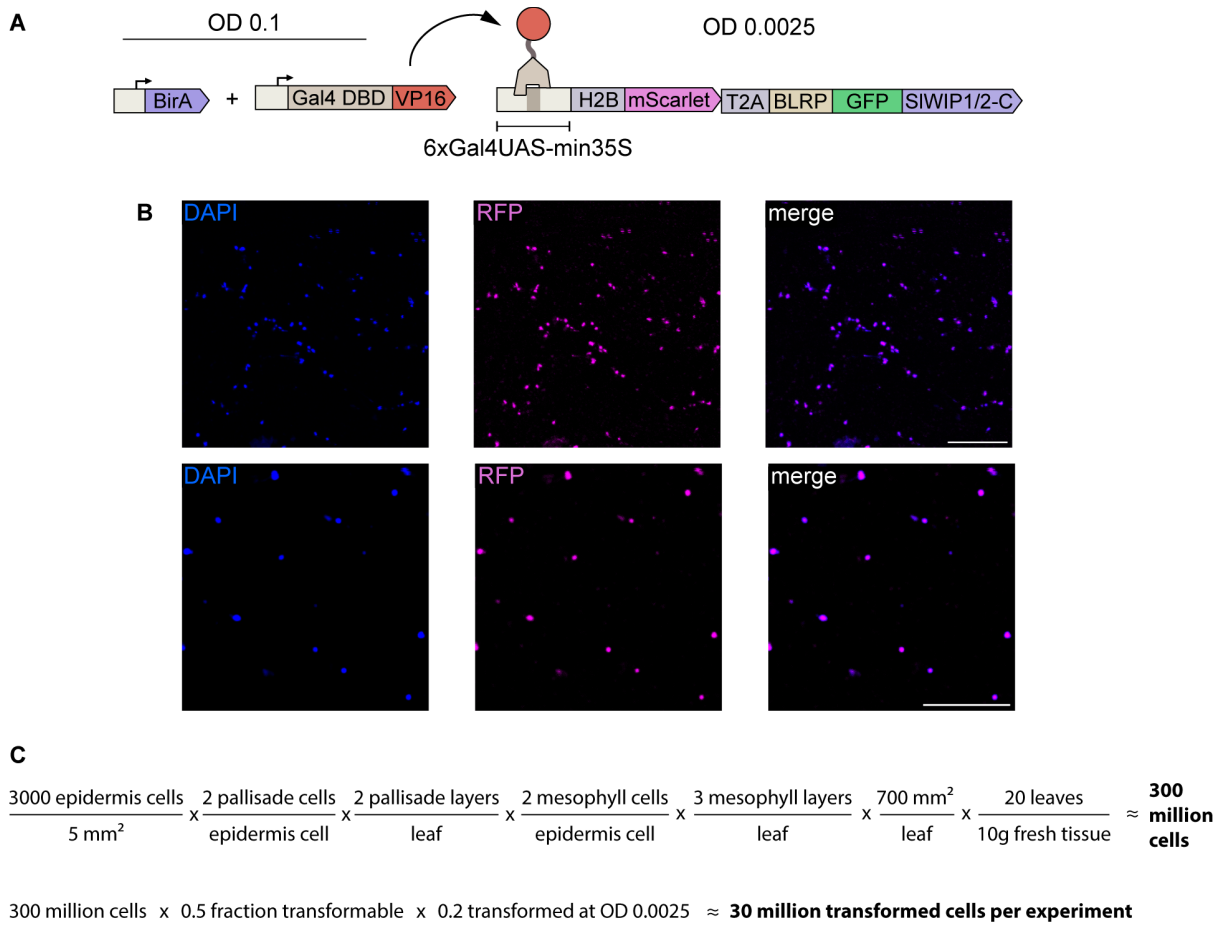


Figure S2: Efficiency of nucleus pulldown. (A) schematic of the experiment used to test the yield and purity of tobacco nucleus pulldowns. Transgenes for the constitutive expression of BirA and Gal4-VP16 were each delivered by a different *Agrobacterium* strain at an OD of 0.1. The culture mix also included a strain carrying the ONEtag reporter infiltrated at an OD of 0.0025. **(B)** Images of purified nuclei stained with DAPI and mounted in a hemocytometer. Bar = 200 um. **(C)** Calculation of the number of cells transformed by a strain carrying the ONEtag based on the numbers obtained from[24].

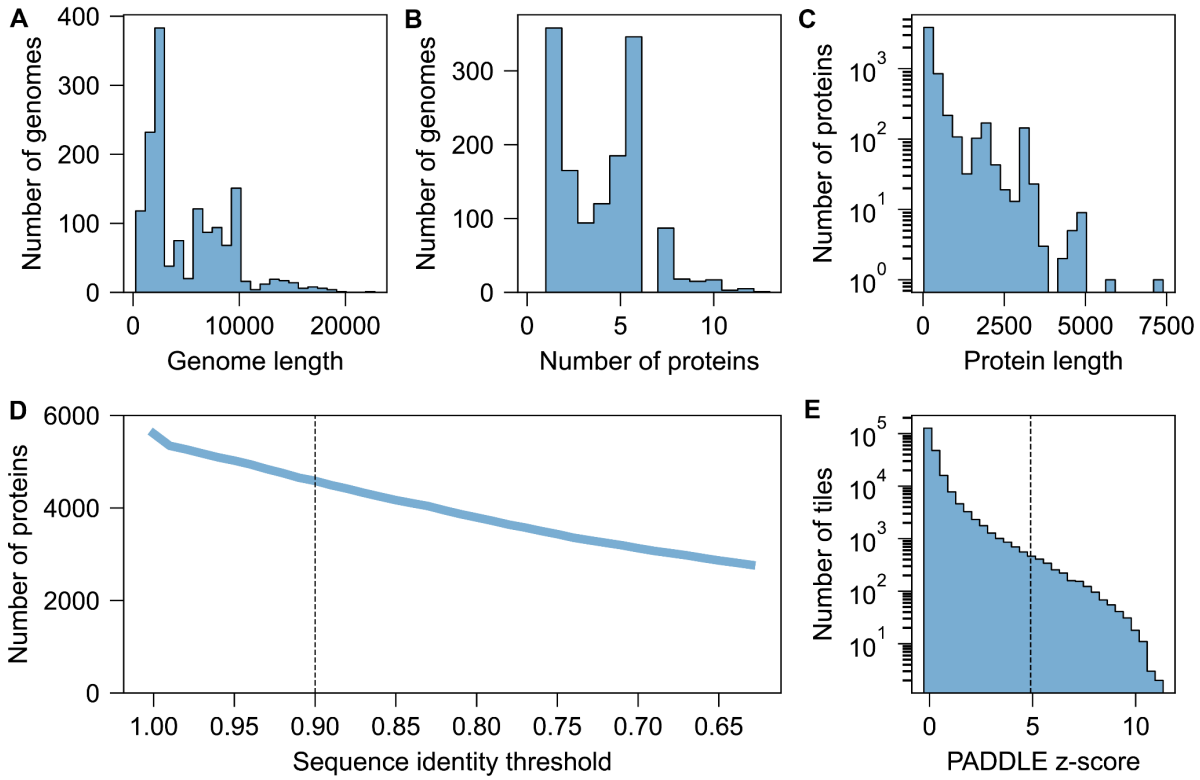


Figure S3: Overview of plant virus genomes and tile selection. Distribution of viral genome lengths. **(B)** Number of proteins in each viral genome. **(C)** Distribution of protein lengths. **(D)** Sequences were clustered at 90% identity (dashed line) to maximize protein diversity, and representative proteins were tiled for activator prediction using PADDLE. **(E)** The top 2,000 tiles with the highest PADDLE predicted z-score, corresponding to a threshold of 4.9 (dashed line), were selected for experimental characterization.

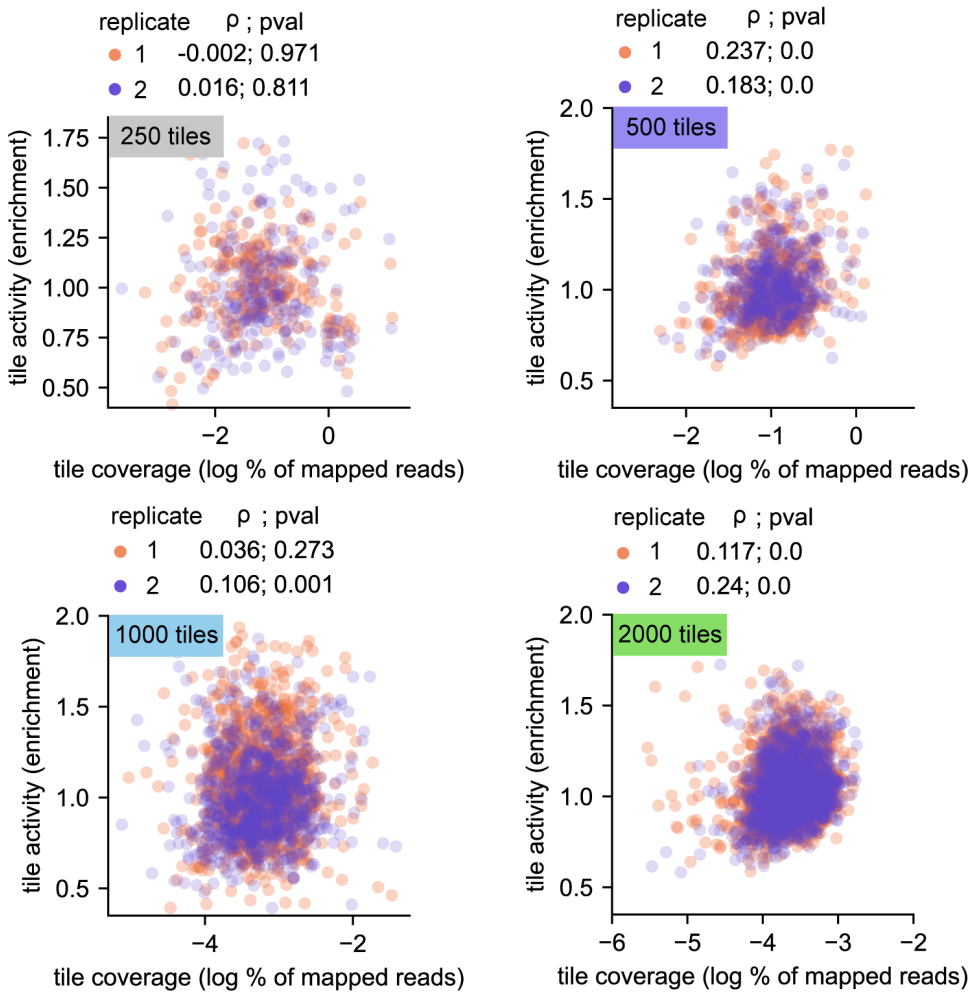


Figure S4: No correlation between tile coverage in the library and its enrichment ratio. Scatter plots showing the transcriptional activity of each tile as measured by its enrichment value (y axis) as a function of the abundance of this tile in the library (x axis).

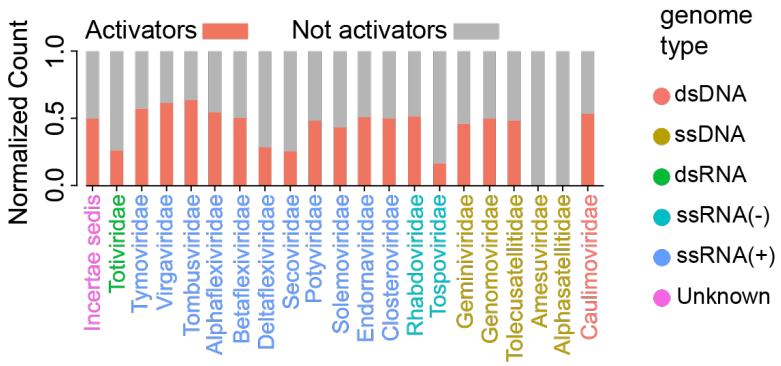


Figure S5: Fraction of selected tiles identified as activators across viral families. For each virus family, the number of activator tiles is normalized by the total number of tiles selected for the corresponding family.

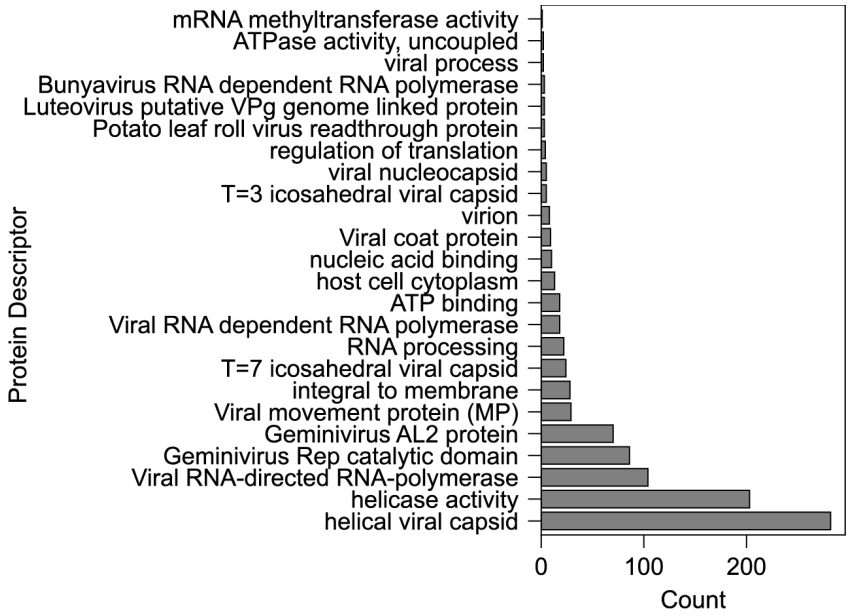


Figure S6: Functional annotation of proteins in the viral libraries. GO term analysis revealed enrichment of specific protein domains across library sequences, highlighting overrepresentation of functional classes among the tested viral protein sequences.

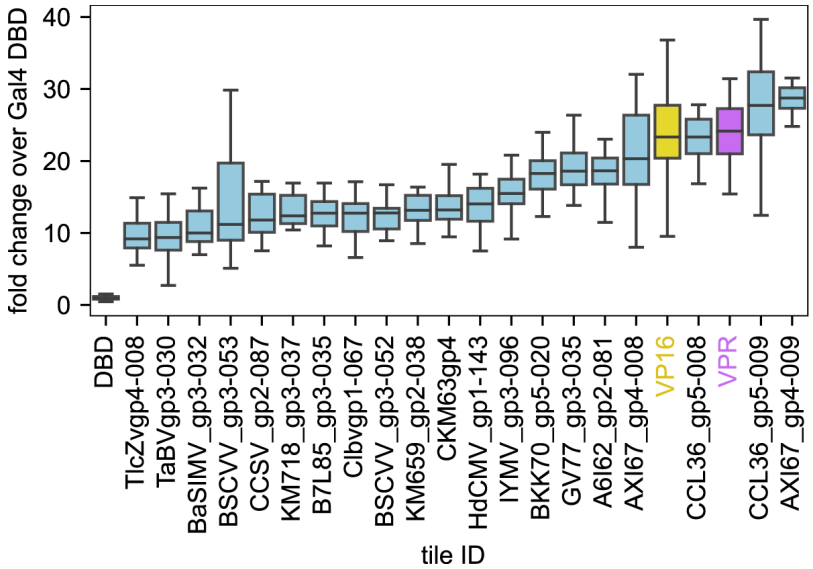


Figure S7: Strong activator tiles from plant viruses. 20 tiles that were among the most enriched in at least one library were cloned individually as Gal4DBD fusions and tested using a fluorescence reporter assay. Shown is the reporter fluorescence as a function of the activity of Gal4DBD alone. Fusions of Gal4DBD to the VP16 and VPR activators were included for comparison.

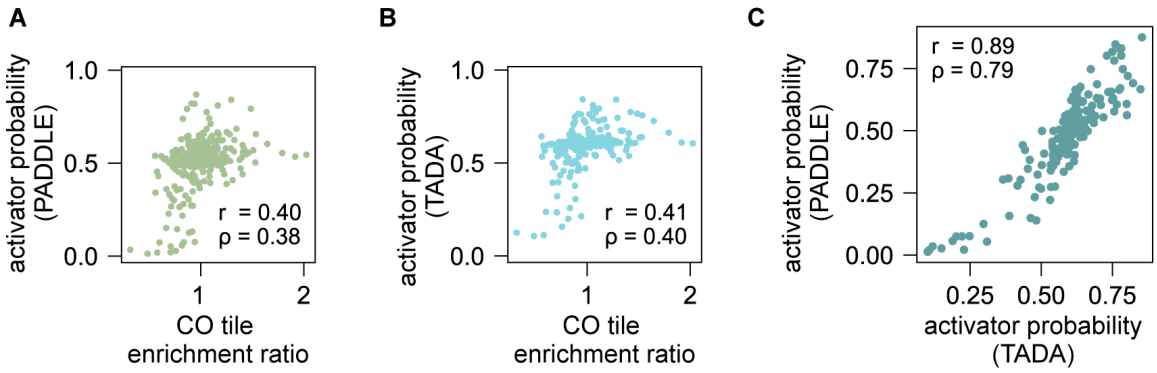


Figure S8: CO variant inference using PADDLE and TADA. (A) Median activator predictions from PADDLE and **(B)** TADA for each CO tile correlates with enrichment ratio. **(C)** Median PADDLE and TADA predictions for each CO tile also show strong concordance with respect to one another.

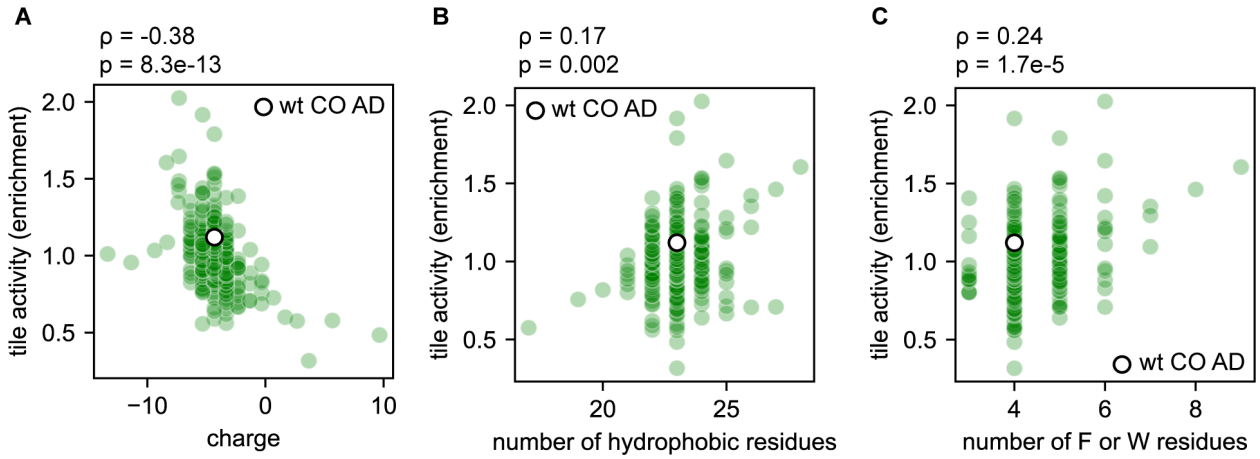


Figure S9: Biochemical properties of the CO acidic AD. The transcriptional activity of each tile in the CO AD library (y axis) is plotted as a function of different biochemical and sequence features. The wt tile is shown as a white circle. **(A)** Tile charge calculated as isoelectric point at pH 7.0. **(B)** number of hydrophobic residues. (C) Number of selected hydrophobic residues, phenylalanine (F) and tryptophan (W) and ρ = Spearman correlation coefficient.

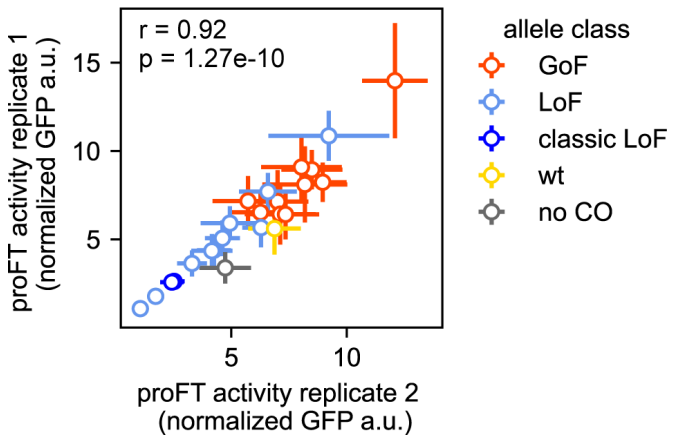


Figure S10: Replicate of CO alleles regulating the FT promoter driving GFP. The experiment shown in Fig. 6H was replicated on a different week. The FT promoter activity driven by a given CO allele on the first replicate (y axis) is plotted against the results from the same allele from the second replicate (x axis).

Supplementary Tables

All supplementary tables can be accessed from the public google drive folder https://drive.google.com/drive/folders/1DawbzdYHvlYebiBnAtV0fqYvYBUnKdQ8?usp=drive_link

Table S1: Compiled data for viral libraries.
Table S2: Compiled data for yeast library.
Table S3: Compiled data for CO library.
Table S4: List of plasmids.
Table S5: List of primers.