


START domains generate paralog-specific regulons from a single network architecture

Received: 27 February 2024

Accepted: 1 November 2024

Published online: 14 November 2024

 Check for updates

Ashton S. Holub^{1,2}, Sarah G. Choudury^{1,3}, Ekaterina P. Andrianova⁴, Courtney E. Dresden^{1,3,5}, Ricardo Urquidi Camacho^{1,3}, Igor B. Zhulin^{1,3} & Aman Y. Husbands^{1,3} 

Functional divergence of transcription factors (TFs) has driven cellular and organismal complexity throughout evolution, but its mechanistic drivers remain poorly understood. Here we test for new mechanisms using CORONA (CNA) and PHABULOSA (PHB), two functionally diverged paralogs in the CLASS III HOMEODOMAIN LEUCINE ZIPPER (HD-ZIP III) family of TFs. We show that virtually all genes bound by PHB (~99%) are also bound by CNA, ruling out occupation of distinct sets of genes as a mechanism of functional divergence. Further, genes bound and regulated by both paralogs are almost always regulated in the same direction, ruling out opposite regulation of shared targets as a mechanistic driver. Functional divergence of CNA and PHB instead results from differential usage of shared binding sites, with hundreds of uniquely regulated genes emerging from a commonly bound genetic network. Regulation of a given gene by CNA or PHB is thus a function of whether a bound site is considered ‘responsive’ versus ‘non-responsive’ by each paralog. Discrimination between responsive and non-responsive sites is controlled, at least in part, by their lipid binding START domain. This suggests a model in which HD-ZIP III TFs use information integrated by their START domain to generate paralog-specific transcriptional outcomes from a shared network architecture. Taken together, our study identifies a mechanism of HD-ZIP III TF paralog divergence and proposes the ubiquitously distributed START evolutionary module as a driver of functional divergence.

Life depends on careful control of cellular decisions. At the heart of this control is the regulation of gene expression by transcription factors (TFs). Intricacy of gene regulation scales with organismal complexity^{1–6}, as evidenced by the expansion in both the number and diversity of TFs during plant and animal evolution⁷. This expansion was mediated primarily by gene duplications which afforded organisms certain evolutionary advantages. For instance, paralogous TFs could retain partial or complete functional redundancy⁸. This lends

robustness to biological systems, permitting organisms to better tolerate genic or environmental perturbations⁹. Alternatively, paralogous TFs could undergo functional divergence via subfunctionalization or neofunctionalization⁸. In the case of subfunctionalization, degeneration partitions the regulatory properties of a single, ancestral TF across its paralogs. During neofunctionalization, one paralog retains the ancestral function while the other takes on new roles. This functional divergence broadens the regulatory landscape of the TF family and is a

¹Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. ²Department of Molecular Genetics, The Ohio State University, Columbus, OH 43215, USA. ³Epigenetics Institute, University of Pennsylvania, Philadelphia, PA 19104, USA. ⁴Department of Microbiology, The Ohio State University, Columbus, OH 43215, USA. ⁵Molecular, Cellular, and Developmental Biology, The Ohio State University, Columbus, OH 43215, USA.

 e-mail: ayh@upenn.edu

central driver of cellular and organismal complexity^{8,10}. Identifying mechanisms of TF functional divergence is thus key to understanding how organisms execute specific decisions ranging from stress response to immunity to development. This is a particularly challenging problem in the plant lineage, whose TF families underwent a more dramatic expansion than other kingdoms^{11,12}.

One well-established mechanism of TF divergence is the occupation of different *in vivo* binding sites^{1–6,13–15}. The acquisition of new binding sites can be influenced by several factors such as chromatin accessibility¹⁶, distribution of co-factors^{14,17,18}, and preferences in DNA sequence and/or shape¹⁹. Another method by which paralogs diverge is ‘antifunctionalization’, wherein paralogous TFs act antagonistically on the same pathway²⁰. For instance, FLOWERING LOCUS T (FT) and TERMINAL FLOWER1 (TFL1) are paralogous TFs that promote or repress flowering, respectively²¹. Mechanistically, this is accomplished by mutually exclusive interactions between FT or TFL1 and their bZIP co-factor FD such that FD-FT complexes activate targets and FD-TFL1 complexes repress them²². In another example, the metazoan early gene 2 factor (E2F) family of activators and repressors similarly act antagonistically on a shared set of targets. Unlike FT and TFL1 however, their promoter binding events are temporally separated throughout the cell cycle²³. There are likely many more biological strategies to generate paralog-specific transcriptional outputs, and plant evolutionary history suggests this kingdom is a particularly fertile ground in which to discover them^{11,12}.

The CLASS III HOMEODOMAIN LEUCINE ZIPPER (HD-ZIPIII) family of plant TFs arose ~725 million years ago and proliferated over the course of evolution. The model plant *Arabidopsis thaliana* has five paralogs – *REVOLUTA* (*REV*), *PHABULOSA* (*PHB*), *PHAVOLUTA* (*PHV*), *CORONA* (*CNA*), and *ATHB-8* – that collectively impact nearly all aspects of development^{24–31} (Fig. 1a). HD-ZIPIII paralogs are divided into two sub-clades – the *REVOLUTA* clade and the *CORONA* clade – with members displaying both functional redundancy and functional divergence. Examples of redundancy include *PHB*, *PHV*, *REV*, and *CNA* redundantly promoting dorsal identity in lateral organs and xylem identity in the vasculature^{25,32,33}. An example of divergence is *CNA* antagonizing the stem cell niche, while *REV* promotes its formation or maintenance²⁶. This functional divergence is driven by substitutions in the coding sequences of *REV* and *CNA*, rather than modifications at the *cis*-regulatory level, however, the nature of the underlying mechanism remains unknown²⁶.

Consistent with their critical role in development, HD-ZIPIII TFs are regulated by numerous inputs including microRNAs and interacting proteins^{34–37}. Additional points of regulation are suggested by their characteristic architecture, which consists of a homeodomain (HD), a leucine zipper (LZ), a MEHKL/PAS-like domain, and a StAR-related transfer (START) domain. The latter is part of the StArkin domain superfamily, named for their kinship to steroidogenic acute regulatory protein (StAR). StArkin domains are present throughout the Tree of life and are defined by a conserved α/β helix grip fold structure^{38–40}. Dysregulation of StArkin proteins has profound effects on disease, stress responses, and development across Eukaryota^{39–45}. StArkin domains bind a variety of lipid ligands that trigger conformational changes to modulate the activity of StArkin proteins using context-dependent regulatory mechanisms ranging from protein turnover to subcellular localization to homomeric and heteromeric complex stoichiometry (reviewed in ref. 46). The diverse and varied architecture of HD-ZIPIII proteins facilitates the integration of multiple regulatory inputs and presents numerous opportunities to drive functional divergence, as amino acid substitutions underlying paralog-specific responses are not limited to DNA-binding domains^{13,47,48}. HD-ZIPIII TFs are thus an excellent model to identify new regulatory mechanisms driving TF paralog divergence at the protein level.

Using genetic and genomic analyses, we find that the functional divergence of CNA and PHB is not explained by binding to different loci or opposite regulation of shared targets. Rather, the primary mechanism by which these HD-ZIPIII paralogs generate distinct transcriptional outcomes is differential usage of shared binding sites. For instance, we show that the binding profiles of CNA and PHB are nearly overlapping. Despite this, CNA and PHB each have hundreds of uniquely regulated direct targets. These unique targets occasionally include genes not bound by the other paralog; however, this is the minority case. Paralog-specific regulons are thus primarily a function of whether a given bound site is considered responsive versus non-responsive. Using deletions, chimeras, and targeted amino acid substitutions, we further demonstrate that this differential usage of shared binding sites is driven, at least in part, by the HD-ZIPIII START domain. Taken together, our study identifies a mechanism of HD-ZIPIII TF paralog divergence and proposes the ubiquitously distributed START evolutionary module as a driver of functional divergence.

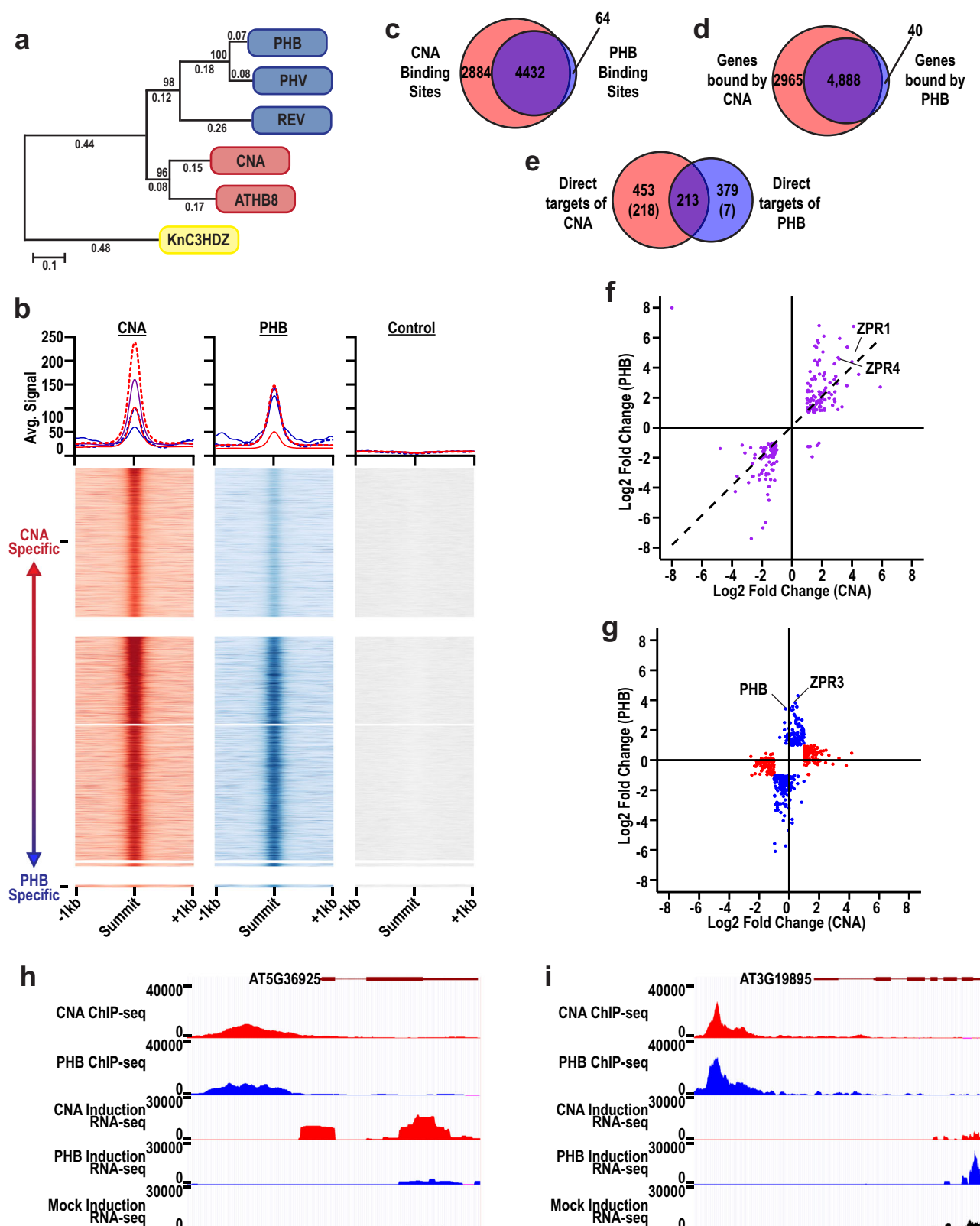
Results

CNA and PHB bind a largely overlapping set of genomic regions

Binding new sites in the genome is a well-known mechanism by which paralogs generate distinct outcomes^{1–6}. Occupation of different loci by CNA and PHB is thus a logical explanation for their functional divergence. To test this possibility, we identified genomic regions bound *in vivo* by each TF using an established short-term estradiol induction system and chromatin immunoprecipitation followed by sequencing⁴⁹ (ChIP-seq; Supplementary Fig. 1). CNA and PHB bound 7316 and 4496 sites in the genome, corresponding to 7853 and 4928 genes, respectively (Fig. 1b–d). Remarkably, 99% of genes bound by PHB (4888 out of 4928) are also bound by CNA, with CNA uniquely occupying an additional 2965 genes. Supporting the relevance of our findings, centrally localized partial HD-ZIPIII binding sites emerged as consensus motifs for CNA and PHB^{50,51} (VTAATNATTAB for CNA; TAATRATKATD for PHB; Supplementary Fig. 2). However, these motifs do not strictly predict CNA and PHB binding. For instance, VTAATNATTAB occurs 20,512 times in the *Arabidopsis* genome, yet only ~7% of these motifs are bound by CNA (1436 out of 20,512; Supplementary Data 1). Additional elements are thus likely to refine binding site selection. A potential candidate to provide this specificity is the three-dimensional shape of DNA which plays important roles in determining *in vivo* TF binding^{19,52,53}. Consistent with this prediction, DNA at bound versus unbound motifs showed clear differences in numerous shape features. For instance, comparisons of bound versus unbound motifs (and their flanking regions) uncovered strong deviations in minor groove width, helix twist, propeller twist, and roll for both CNA and PHB (Supplementary Fig. 3). Thus, CNA and PHB appear to bind a specific but largely overlapping set of sites, whose selection is guided by both shape and sequence of DNA.

Signal intensities at these mutually bound sites fall into three categories: higher signal for CNA (3512 out of 9571), higher signal for PHB (272 out of 9571), and equivalent signal for both TFs (5787 out of 9571; Fig. 1b left, middle). CNA thus appears to bind with higher affinity than PHB at many, but not all, regions of the genome. Interestingly, PHB ChIP signal intensities are slightly higher at the summits of CNA-specific binding sites than the flanking regions (Fig. 1b middle). Further, signal intensities at these regions are significantly higher than those from an immunoprecipitated non-transgenic control (Fig. 1b right; Supplementary Data 2). Thus, it is technically possible that PHB also binds to ‘CNA-specific’ sites, but its binding fails to meet our stringent statistical criteria. If so, CNA and PHB would occupy an essentially overlapping set of sites in the genome. Although we proceed under the assumption of CNA-specific binding, we note this alternate possible conclusion.

Our findings reveal a substantial overlap in the genomic occupancy of CNA and PHB. This observation is consistent with their partial



functional redundancy²⁶. Further, the unique occupancy at 2965 additional genes by CNA provides a potential explanation for its functional divergence from PHB. By contrast, the binding profile of PHB is essentially a fully contained subset of the binding profile of CNA (Fig. 1d). Functional divergence of PHB activity must therefore derive from an alternate mechanism.

CNA and PHB have hundreds of uniquely regulated direct targets despite substantial overlap in genomic occupancy

TF occupancy at a gene is not necessarily indicative of regulation (reviewed in ref. 54). We, therefore, tested whether binding by CNA and PHB leads to changes in gene expression using short-term estradiol induction followed by transcriptome profiling⁴⁹. After induction,

Fig. 1 | PHB and CNA have hundreds of uniquely regulated targets despite extensive overlap in genomic occupancy. **a** Phylogenetic tree of the HD-ZIPIII family in *A. thaliana* (blue = REV subclade, red = CNA subclade). Tree rooted to the *K. nitens* HD-ZIPIII protein (KnC3HDZ, yellow). **b** Histograms and heatmaps of ChIP-seq signal intensities compared to a non-transgenic control. Histogram lines and colors delineate five categories of binding sites: CNA specifically bound (red, solid); mutually bound – CNA higher affinity (red, dashed); mutually bound – equal affinities (purple); mutually bound – PHB higher affinity (blue, dashed), and PHB specifically bound (blue, solid). Heatmaps are separated into three major categories: CNA uniquely bound (top), mutually bound (center), and PHB uniquely bound (lower). Mutually bound heatmaps contain three further subcategories: mutually bound – CNA higher affinity (top); mutually bound – equal affinities

(center); mutually bound – PHB higher affinity (lower). **c** Venn diagram of sites bound by CNA and PHB. **d** Venn diagram of genes bound by CNA and PHB. **e** Venn diagram of direct targets of CNA and PHB, i.e. bound in ChIP-seq and differentially expressed in RNA-seq. Numbers in parentheses correspond to direct targets bound specifically by CNA or PHB. **f** Scatterplot showing differential expression of genes bound and regulated by both CNA and PHB. **g** Scatterplot showing differential expression of genes bound by both paralogs but uniquely regulated by CNA (red) or PHB (blue). **h, i** Representative genome browser shots illustrating differential usage of shared binding sites: AT2G22860 is bound by both paralogs but uniquely regulated by CNA (**h**), while AT3G19895 is bound by both paralogs but uniquely regulated by PHB (**i**).

1464 and 1686 genes were differentially expressed in CNA- and PHB-expressing lines, respectively (Supplementary Data 3). Genes that were also bound in ChIP assays were called as direct targets. This corresponds to 666 and 592 high-confidence direct targets of CNA and PHB, respectively, which fall into three categories: CNA-specific, PHB-specific, and mutually regulated. Importantly, these datasets include known HD-ZIPIII direct targets such as the *LITTLE ZIPPER (ZPR)* genes, supporting the validity of our approach.

Of the 453 CNA-specific direct targets, 218 were from loci uniquely bound by CNA (Fig. 1e, g). Thus, one mechanism by which CNA generates its specific outcomes is the selection and regulation of new loci (or possibly the retention of ancestral binding sites no longer recognized by PHB). A second potential mechanism could be the opposite regulation of shared direct targets^{20,55}. However, this does not appear to be the case, as the 213 mutual targets of CNA and PHB are almost all regulated in the same direction at the same magnitude (Fig. 1f). A third potential mechanism is suggested by the fact that the remaining 235 CNA-specific direct targets are bound by both TFs but not differentially expressed in PHB-expressing lines. Similarly, PHB has 379 unique direct targets despite sharing a remarkable 99% of bound genes with CNA (Fig. 1d, e, g). This suggests differential usage of shared binding sites as a mechanism by which CNA and PHB generate specific transcriptional outcomes (illustrated in Fig. 1h, i). For each paralog, shared binding sites were termed ‘responsive’ if occupation promoted or inhibited transcription versus ‘non-responsive’ if occupation did not alter gene expression.

One hypothesis is that binding affinity, inferred by the intensity of ChIP signal^{54,56–58}, positively correlates with paralog-specific usage of a given binding site. To test this, we focused on two specific predictions of the hypothesis. The first states that CNA and PHB mutual direct targets should derive from genes containing shared binding sites with equivalent ChIP signal intensities (Fig. 1d). However, only 38% of mutual direct targets (81 out of 213) emerge from this set of genes. This set of genes also contributes 24% of CNA (107 out of 453) and 34% of PHB (129 out of 379) uniquely regulated targets, further arguing against binding affinity as a predictor of site identity (Supplementary Fig. 4a). A second prediction states that mutually bound, but uniquely regulated targets, should possess binding sites with ChIP signal intensities that are higher for the TF uniquely regulating the gene. We formally tested this prediction by focusing first on mutually bound genes whose shared binding sites have a higher ChIP signal for CNA (1,746 out of 4,888 bound genes). If affinity correlates with regulation, these sites should contribute most or all CNA unique targets. However, these sites contributed 12% of CNA uniquely regulated targets (54 out of 453), 27% of PHB uniquely regulated targets (101 out of 379), and 25% of mutually regulated targets (53 out of 213), arguing against this hypothesis (Supplementary Fig. 4b). In addition, a Spearman’s rank correlation test found no relationship between the magnitude of gene expression changes and the intensity of CNA ChIP signals ($p = 0.24$, $p = 1$; Supplementary Fig. 5a). Similar analyses found that PHB ChIP signals also failed to correlate with magnitude of gene expression changes ($p = 0.26$, $p = 0.99$; Supplementary Fig. 5b). These analyses

suggest binding affinity does not dictate responsive versus non-responsive site identity. To test this further, we used additional orthogonal tests to determine whether ChIP signals at responsive sites (i.e. in direct targets) are higher than those at non-responsive sites (i.e. in non-regulated genes). The empirical distribution of ChIP signals deviated for both CNA and PHB (Kruskal-Wallis p -values: 0.006 (CNA), 0.007 (PHB); Supplementary Fig. 6). However, differences in mean ChIP signal were small, showing a slight increase only in upregulated versus non-regulated genes (Mann-Whitney q -values: 0.007 (CNA), 0.031 (PHB); Supplementary Fig. 6). Taken together, these qualitative and quantitative analyses suggest that paralog-specific interpretation of shared binding sites is not driven primarily by their relative binding affinities.

Supporting the biological relevance of these findings, GO term analyses of CNA-specific, PHB-specific, and mutual direct targets revealed both similarities and differences in biological processes. For instance, the most overrepresented processes regulated specifically by CNA almost all relate to wax and fatty acid biosynthesis, as well as stress responses (Supplementary Fig. 7). By contrast, PHB specifically downregulates several players in the auxin and brassinosteroid hormone signaling pathways (Supplementary Fig. 7). Processes regulated by both paralogs include photosynthesis and light responses, as well as camalexin and phytoalexin biosynthesis (Supplementary Fig. 7). In sum, our analyses suggest functional divergence of HD-ZIPIII paralogs is driven primarily by differential usage of shared binding sites which appears to be independent of binding affinity.

The START domain is required for full CNA function

Our analyses propose paralog-specific interpretation of shared binding sites contributes to the functional divergence of HD-ZIPIII TFs. Further, coding sequence swaps between HD-ZIPIII subclade members indicate their functional divergence is due in part to amino acid substitutions; the identity and position of these substitutions are unknown²⁶. Notably, HD-ZIPIII TFs contain a START domain (Fig. 2a), known to control transcriptional regulators in plants and animals through diverse regulatory mechanisms^{46,49,59–63}. In addition, the START domain of PHB enhances its transcriptional potency at *ZPR3* and *ZPR4* targets⁴⁹. Thus, the START domain is an attractive candidate to drive HD-ZIPIII functional divergence through paralog-specific interpretation of shared binding sites.

A regulatory role for the PHB START domain has been established⁴⁹. We therefore tested whether the START domain of CNA is similarly required for its activity. To assess for START regulatory roles in development, we employed two genetic assays previously used to characterize the PHB START domain⁴⁹. Like *PHB*, loss of function single mutants of *CNA* appear wildtype²⁶. Thus, the first genetic assay involves complementing a *phb phv cna* loss-of-function mutant which has a strong pleiotropic phenotype²⁶. We began by replacing the START domain of the functional *pCNA:CNA* reporter⁶⁴ with the 21-nt microRNA166 (miR166) recognition site found within the START domain coding sequence²⁹ (*pCNA:CNA-SDΔ*). Impacts of START domain deletion can thus be assayed without confounding effects

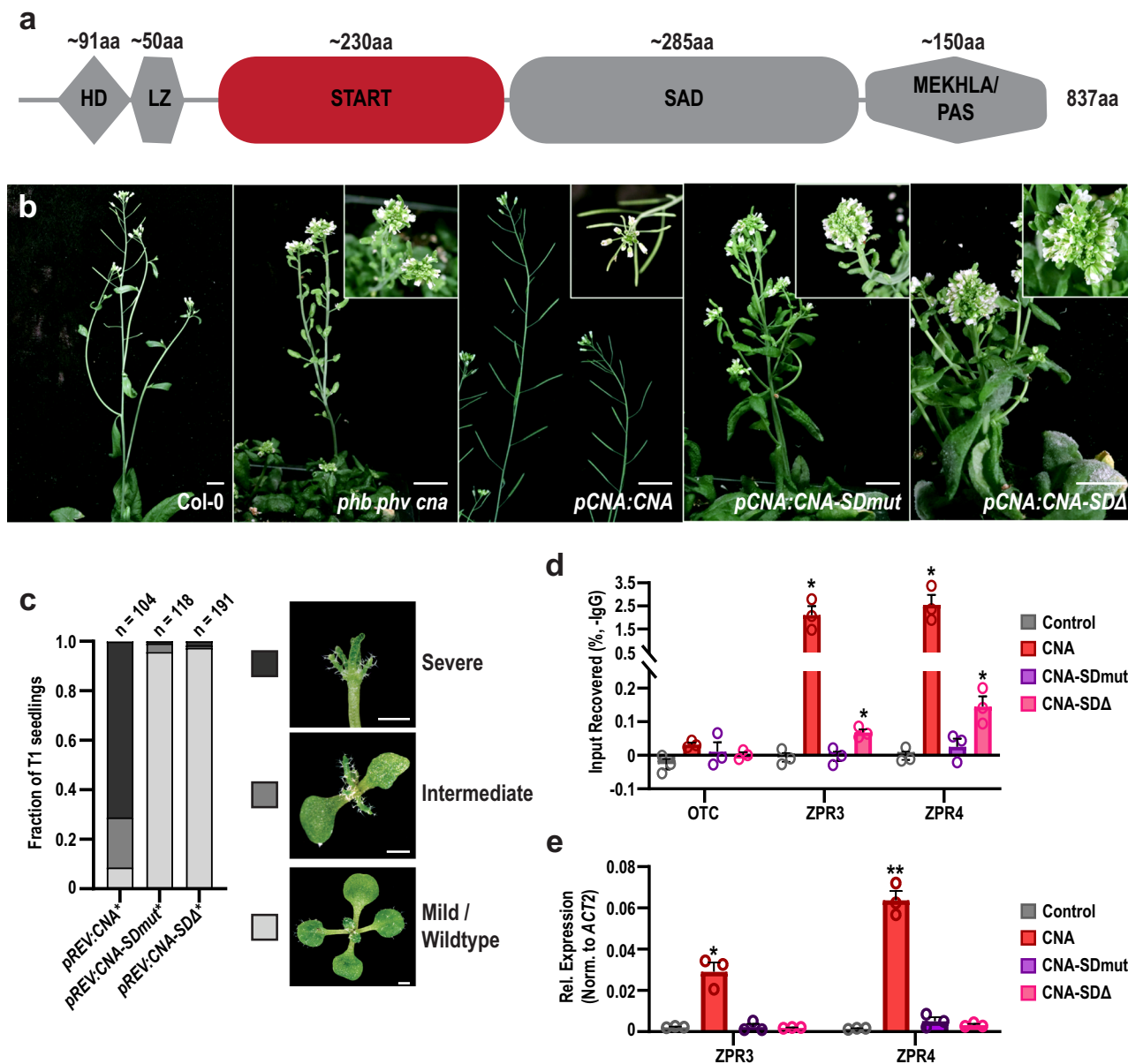


Fig. 2 | The START domain is required for the developmental activity of CNA.

a General structure of HD-ZIP III proteins (HD = homeodomain, LZ = leucine zipper, START = StAR-related lipid transfer, SAD = START adjacent domain, MEKHLA/PAS). **b** The *phb phv cna* triple mutant has a strong pleiotropic phenotype (Prigge et al.,²⁶). The *pCNA:CNA* transgene fully complements the mutant phenotype as plants appear wildtype, whereas the *pCNA:CNA-SDmut* and *pCNA:CNA-SDΔ* transgenes do not. **c** Phenotypic scoring of primary transformants (*n* above bar) expressing miR166 insensitive (*) *CNA-SDmut* and *CNA-SDΔ* under the *REV* promoter. Ectopic *CNA** expression leads to severe (black) or intermediate (dark gray) gain-of-function phenotypes, whereas plants mis-expressing *CNA-SDmut** or *CNA-SDΔ** appear

wildtype (light gray). **d** CNA and CNA-SDΔ bind to regulatory regions of ZPR3 and ZPR4 and are significantly enriched above the *ORNITHINE TRANSCARBAMYLASE* (*OTC*) negative control locus; note CNA-SDΔ signal is significantly lower than that of CNA. CNA-SDmut does not bind to ZPR3 or ZPR4 regulatory regions. **e** CNA robustly activates ZPR3 and ZPR4 targets in 24 h estradiol-induction experiments, whereas CNA-SDmut and CNA-SDΔ do not. Non-transgenic control (gray), *pOlexA:CNA** (red), *pOlexA:CNA-SDmut** (purple), and *pOlexA:CNA-SDΔ** (pink). Data are presented as mean values \pm SEM. Statistics are against the non-transgenic control and use three biological replicates per genotype. * = *p*-value \leq 0.05, ** = *p*-value \leq 0.01, two-way paired Student's *t*-test.

from ectopic CNA transcript accumulation⁶⁵. In a separate construct, we introduced a set of mutations in the ligand-binding pocket that do not perturb HD-ZIP III START secondary structure⁴⁹ (*pCNA:CNA-SDmut*). Unlike *pCNA:CNA*, neither *pCNA:CNA-SDmut* nor *pCNA:CNA-SDΔ* rescued the *phb phv cna* mutant phenotype (Fig. 2b). In an orthologous genetic approach, we repurposed the established *pREV:CNA* reporter into a gain-of-function, highly sensitive readout of CNA activity by introducing a silent mutation into its miR166 recognition site²⁶ (*pREV:CNA**). As expected, over 90% of *pREV:CNA** primary transformants show phenotypes characteristic of ectopic HD-ZIP III

activity including dorsalized leaves^{28,29,65} (Fig. 2c). By contrast, *pREV:CNA*SDmut* and *pREV:CNA*Delta* primary transformants are indistinguishable from wildtype plants (Fig. 2c). Our genetic assays demonstrate the START domain is required for CNA to fulfil its developmental function.

The START domain of PHB potentiates transcriptional activity using multiple distinct mechanisms prompting us to test whether the CNA START domain behaves similarly⁴⁹. First, we tested whether the START domain impacts the ability of CNA to bind to DNA. ChIP qPCR showed strong enrichment of CNA at two *ZPR* loci (Fig. 2d). This

binding was not observed with CNA-SDmut (Fig. 2d), matching the behavior of PHB variants with analogous mutations⁴⁹. Surprisingly, CNA-SDΔ also showed strongly reduced occupancy at *ZPR3* and *ZPR4* (Fig. 2d). This behavior contrasts with PHB-SDΔ whose occupancy at these loci was nearly identical to PHB⁴⁹. Thus, the impact of the START domain on DNA binding appears to differ between HD-ZIPIII paralogs. Next, we tested whether the CNA START domain affects the regulation of these targets using short-term estradiol inductions and RT-qPCR. As expected, transcript levels of *ZPR* targets were upregulated by CNA and unchanged after induction of CNA-SDmut (Fig. 2e). CNA-SDΔ also failed to upregulate these targets suggesting its sharp reduction in genomic occupancy at these loci has functional consequences (Fig. 2e). Taken together, these analyses demonstrate a role for the START domain in regulating CNA activity and suggest START-mediated effects on DNA binding differ across the HD-ZIPIII subclades.

CNA and PHB START domains promote DNA binding and may help distinguish responsive from non-responsive binding sites

Having established a regulatory role for START domains of both paralogs at their *ZPR* targets, we next tested how their loss affects binding and regulation of CNA and PHB targets genome wide. ChIP-seq identified 3276 and 4490 sites bound by CNA-SDΔ and PHB-SDΔ, corresponding to 4003 and 4300 bound genes, respectively (Fig. 3a–c, g–i). These binding profiles are essentially fully contained subsets of their respective wildtype counterparts. For instance, 99% of genes bound by CNA-SDΔ (3971 out of 4003) are also bound by CNA, with CNA recognizing an additional 3706 genes (Fig. 3c). In the case of PHB-SDΔ, 99% of bound genes (4260 out of 4300) are also bound by PHB, with PHB recognizing an additional 623 genes (Fig. 3i). Deletion of the CNA and PHB START domains thus leads to reduced genomic occupancy of these TFs, and this effect is more dramatic for CNA than PHB (Fig. 3b, h). Signal intensities at mutually bound sites fall into three categories: higher signal for the wildtype protein, higher signal for the START-deleted variant, and equivalent signal for both variants (Fig. 3a, g). Interestingly, nearly all binding sites fall in the first and last categories. This indicates occupancy at a given locus is either unaffected or reduced by START domain deletion, with reductions occurring much more often with CNA-SDΔ than PHB-SDΔ (Fig. 3a, g). Thus, one function of the START domain may be to increase HD-ZIPIII binding affinities in a paralog- and site-specific manner.

If shared binding site usage is controlled by HD-ZIPIII START domains, this should be reflected in gene expression changes at commonly bound genes. To begin to assess this, we tested how loss of their START domains impacts regulation of CNA and PHB targets. After induction, 642 and 1005 genes were differentially expressed in CNA-SDΔ- and PHB-SDΔ-expressing lines, respectively (Supplementary Data 3). This corresponds to 185 and 95 high-confidence direct targets of CNA-SDΔ and PHB-SDΔ, respectively (Fig. 3d, j). Comparing CNA-SDΔ and PHB-SDΔ to their wildtype counterparts generates three categories of direct targets: unique to the wildtype protein, unique to the START-deleted variant, and mutually regulated (Fig. 3d, j). We first asked whether the START domain controls the direction of target regulation and found that mutual targets of wildtype and START-deleted variants are almost all regulated in the same direction (Fig. 3e, k). Interestingly, the magnitudes of gene expression changes are often lower in START-deleted samples which could be due to reduced transcriptional potency of these variants⁴⁹. Alternatively, this may be a downstream consequence of their reduced genomic occupancy at these shared sites (Fig. 3a, g). We next asked whether direct targets specific to each wildtype protein are explained by their expanded genomic occupancy. Supporting this idea, 42% of CNA-specific targets are not bound by CNA-SDΔ (260 out of 614), and 24% of PHB-specific targets are not bound by PHB-SDΔ (78 out of 330). However, this is the minority case, with 58% of CNA-specific targets (354 out of 614) and 76% of PHB-specific targets (252 out of 330) also

bound by their START-deleted variant (Fig. 3d, f, j, l; Supplementary Fig. 8). In addition, despite the binding profiles of CNA-SDΔ and PHB-SDΔ being essentially fully contained subsets of their wildtype counterparts, these proteins have 185 and 95 unique direct targets, respectively (Fig. 3d, f, j, l; Supplementary Fig. 8). Thus, the START domain may help CNA and PHB discriminate between responsive and non-responsive binding sites. Taken together, our functional genomic assays support a role for START domains as potential drivers of HD-ZIPIII functional divergence.

Chimeric proteins have different effects on CNA developmental activity

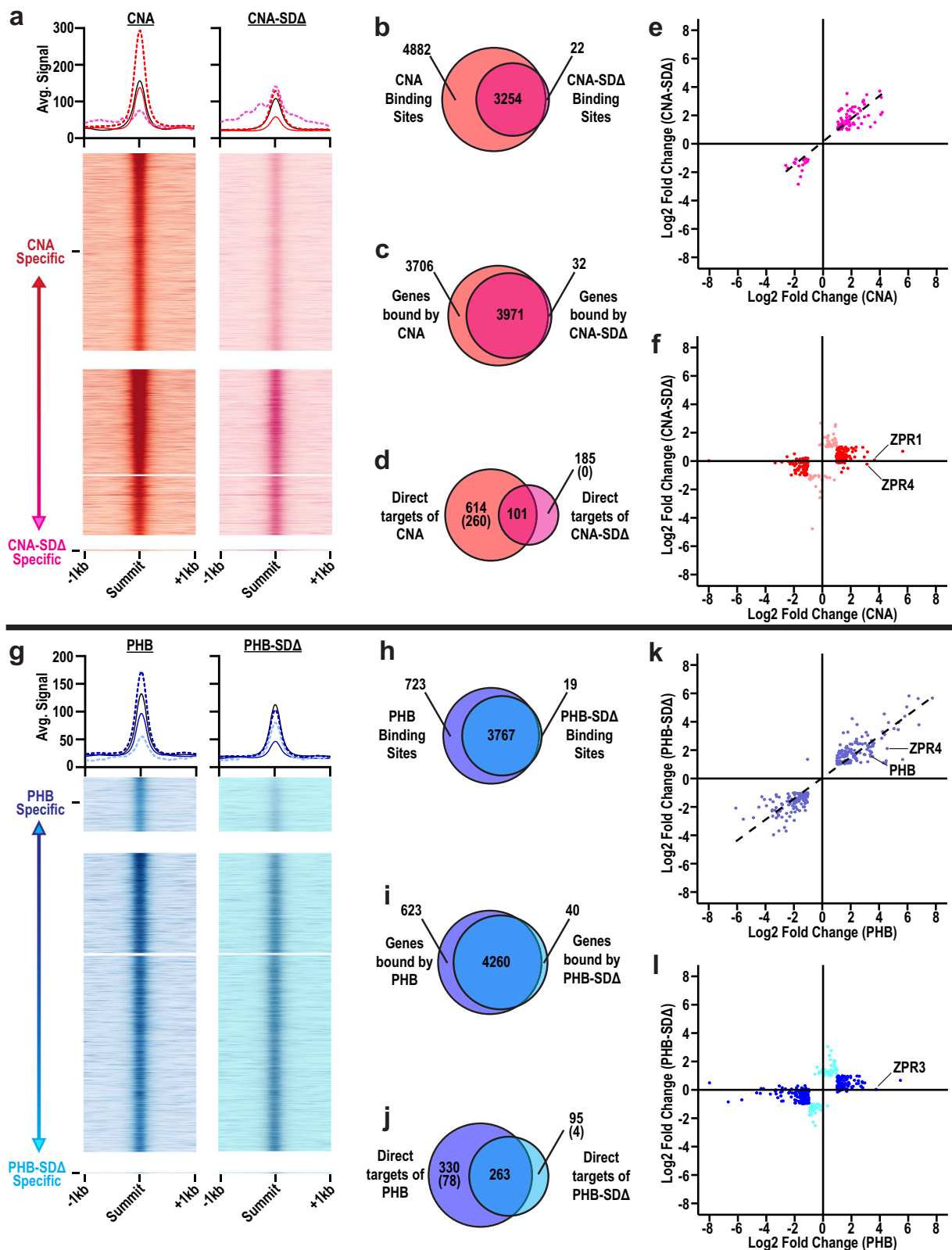
Deletion analyses suggest the START domain may contribute to both the selection and regulation of HD-ZIPIII targets (Fig. 3). One approach to generate additional support for this hypothesis is to leverage evolutionary history. HD-ZIPIII TFs are deeply conserved but contain numerous lineage-specific substitutions within their START domains. This pool of natural variation can be functionally interrogated by replacing the START domain of an *Arabidopsis* HD-ZIPIII protein with START domains from orthologs in other species. The resulting chimeras probe START-directed regulatory properties while minimizing potential confounding effects associated with full-domain deletions.

To guide the selection of this *Arabidopsis* HD-ZIPIII protein, we reconstructed the evolutionary history of the HD-ZIPIII family and discovered that the ancestral-most HD-ZIPIII protein most closely resembles CNA (Supplementary Fig. 9). CNA was thus selected as the transgenic platform for our chimeras. Orthologous START domain sequences were then chosen using horizontal and vertical approaches (Supplementary Figs. 9, 10). In the horizontal approach, the CNA START domain was replaced by START domains from orthologs in five extant species: *Zea mays* (ZmHOX29), *Selaginella moellendorffii* (SmHOX32), *Physcomitrella patens* (PpHOX32), *Klebsormidium nitens* (KnC3HDZ), and *Chlorokybus atmophyticus* (CaC3HDZ). In the vertical approach, the CNA START domain was replaced by START domains from ancestral reconstructions of HD-ZIPIII sequences at the base of the Angiosperm (AngioAnc) or Charophyte (CharoAnc) lineages.

To assess the effect of each START domain on CNA regulation, we first modified the *pREV:CNA** transgene from Fig. 2c to express miR166-resistant versions of chimeras. Using the frequency of HD-ZIPIII gain-of-function phenotypes as a readout, we found five sequences that can substitute for CNA START developmental function and three that cannot (Fig. 4a). Our chimeric constructs are thus likely to be useful tools to identify START-directed effects on HD-ZIPIII divergence. Chimeras were then cloned into the estradiol-inducible system to test for effects on DNA binding and target activation. As expected, the five chimeras that condition developmental phenotypes also activate *ZPR3* and *ZPR4* targets (Fig. 4b). Surprisingly, the three chimeras that could not affect development fell into two sub-classes. Members of one sub-class showed no activation of *ZPR3* or *ZPR4* (CNA-SmHOX32 and CNA-CaC3HDZ) while the other (CNA-KnC3HDZ) robustly activated both targets (Fig. 4b). We therefore selected three chimeras for further analyses: 1) CNA-SmHOX32, which does not appear to function in molecular or phenotypic assays, 2) CNA-PpHOX32, which appears to fully function in both assays, and 3) CNA-KnC3HDZ, which binds and activates *ZPR* targets yet fails to condition developmental phenotypes.

Chimeric proteins show differences in target selection and regulation

A simple explanation for the failure of CNA-SmHOX32 to activate *ZPR3* and *ZPR4* targets is loss of DNA binding. However, CNA-SmHOX32 ChIP qPCR enrichment values are indistinguishable from unmodified CNA at these loci (Fig. 4c; Supplementary Fig. 11). One possibility is this chimeric protein is non-functional and has lost the ability to regulate bound targets. Alternatively, replacing the CNA START domain with the SmHOX32 START domain may have shifted its binding profile



and/or repertoire of responsive versus non-responsive sites. The latter would be consistent with results from START deletion assays (Fig. 3). To differentiate between these possibilities, we first identified the genomic regions bound by CNA-SmHOX32 using ChIP-seq. CNA-SmHOX32 bound 9679 sites in the genome corresponding to 9816 genes (Fig. 5a–c). Over 99% of these bound genes (9809 out of 9816) are shared with CNA, with CNA uniquely occupying an additional 2

genes (Fig. 5c). Thus, replacing the CNA START domain with the SmHOX32 START domain has a negligible effect on genome occupancy as CNA and CNA-SmHOX32 have virtually identical binding profiles.

We next tested the impact of CNA-SmHOX32 on gene expression via transcriptome profiling of CNA-SmHOX32-induced lines. After induction, 2203 genes were differentially expressed in CNA-SmHOX32-

Fig. 3 | PHB and CNA START domains may help distinguish responsive from non-responsive shared binding sites. a–f CNA vs CNA-SDA analyses (red vs pink). **g–l** PHB vs PHB-SDA analyses (blue vs cyan). **a, g** Histograms and heatmaps of ChIP-seq signal intensities. Lines on histograms delineate categories as described in Fig. 1. **b, h** Heatmaps are separated into categories and subcategories as described in Fig. 1. **b, h** Venn diagrams of sites bound by wildtype and START-deleted variants. **c, i** Venn diagrams of genes bound by wildtype and START-deleted variants. **d, j** Venn diagrams of direct targets of wildtype and START-deleted variants, i.e. bound in

ChIP-seq and differentially expressed in RNA-seq. Numbers in parentheses correspond to direct targets bound specifically by wildtype or START-deleted variants. **e, k** Scatterplots showing differential expression of genes bound and regulated by both wildtype and START-deleted variants. **f** Scatterplot showing differential expression of genes bound by both variants but uniquely regulated by CNA (red) or CNA-Delta (pink). **l** Scatterplot showing differential expression of genes bound by both variants but uniquely regulated by PHB (blue) or PHB-SDA (cyan).

expressing lines corresponding to 991 high-confidence direct targets (Fig. 5d; Supplementary Data 3). CNA and CNA-SmHOX32 direct targets fall into three categories: CNA-specific, CNA-SmHOX32-specific, and mutually regulated (Fig. 5d). Despite binding to nearly identical sets of genes, CNA and CNA-SmHOX32 have remarkably few mutually regulated targets (Fig. 5d, e), with most genes showing specific regulation by either CNA (558 out of 805) or CNA-SmHOX32 (744 out of 991; Fig. 5d, f; Supplementary Fig. 8). Thus, replacing the CNA START domain with the SmHOX32 START domain profoundly affects target regulation, primarily through differential usage of shared binding sites.

Similar analyses found the CNA-PpHOX32 chimera bound 8650 sites in the genome corresponding to 9163 genes (Fig. 5g–i). Over 99% of these bound genes (9148 out of 9163) are shared with CNA, with CNA uniquely occupying an additional 175 genes. Pairing ChIP-seq with transcriptome profiling again found widespread differential usage of shared binding sites. For instance, despite their extensive binding overlap, CNA-PpHOX32 has 354 direct targets that are not shared with CNA (Fig. 5j, l; Supplementary Fig. 8). Similarly, CNA has 579 unique targets, 98% of which derive from mutually bound genes (565 out of 579; Fig. 5j; Supplementary Fig. 8).

Taken together, these developmental and functional genomics assays support the idea that START domains modulate HD-ZIPIII TF activity primarily by distinguishing whether a site is considered responsive or non-responsive by a given paralog.

START point mutations mimicking HD-ZIPIII evolution drive differences in target selection and regulation

Domain swaps are consistent with START domains driving changes in HD-ZIPIII transcriptional outputs. However, TF evolution rarely involves wholesale exchanges of domains, instead working largely through gradual accrual of individual amino acid substitutions. If START domains are indeed drivers of HD-ZIPIII functional divergence, even minimal changes in their sequences should impact target selection and/or differential usage of shared binding sites. The CNA-KnC3HDZ chimera provides an opportunity to test this idea, as this protein retains some molecular properties of CNA but is unable to trigger its developmental program (Fig. 4).

We hypothesized that a small number of amino acid changes may be sufficient to convert the CNA-KnC3HDZ chimera into one with regulatory properties more closely resembling CNA. To identify these amino acids, we took advantage of the fact that CNA ortholog START domains displayed differential substitutability in our molecular and phenotypic assays (Fig. 4a, b). Using multiple sequence alignments, we found eight residues that are conserved only in fully-substituting chimeras (Supplementary Fig. 12). These residues were then introduced into the START domain of CNA-KnC3HDZ to create CNA-KnC3HDZ-8m. Supporting the functional relevance of these choices, a *pREV:CNA*-KnC3HDZ-8m* transgene conditioned partial HD-ZIPIII gain-of-function phenotypes in primary transformants whereas the *pREV:CNA*-KnC3HDZ* transgene did not (Supplementary Fig. 13).

To test the effects of the eight substitutions on target selection, we identified genomic regions bound by CNA-KnC3HDZ and CNA-KnC3HDZ-8m using ChIP-seq. CNA-KnC3HDZ bound 8480 sites in the

genome corresponding to 11,850 bound genes, and its binding profile is fully contained within that of CNA (Fig. 6a–c). By contrast, CNA-KnC3HDZ-8m showed a slightly more expanded binding profile, recognizing 8871 sites in the genome, corresponding to 12,228 bound genes (Fig. 6a–c). Interestingly, 26% of all sites newly recognized by CNA-KnC3HDZ-8m (102 out of 392) are shared with CNA (Fig. 6a–c). Thus, eight amino acid substitutions can shift the binding profile of CNA-KnC3HDZ towards that of CNA.

To test the effects of the eight substitutions on target regulation, we identified direct targets of CNA-KnC3HDZ and CNA-KnC3HDZ-8m by pairing ChIP-seq and transcriptome profiling. 378 and 1,151 genes were differentially expressed in KnC3HDZ- and KnC3HDZ-8m-expressing lines, corresponding to 256 and 780 high-confidence direct targets, respectively (Fig. 6d; Supplementary Data 3). Of the 248 direct targets shared only by CNA and KnC3HDZ-8m, 7 were from genes newly recognized by KnC3HDZ-8m. This supports the notion that selection and regulation of new loci is not the primary mechanism by which START domains generate TF-specific outcomes. The remaining 241 direct targets are derived from loci that are bound by all three proteins, but only regulated by CNA and KnC3HDZ-8m (Fig. 6e; Supplementary Fig. 8). These direct targets correspond to genes with sites considered non-responsive by CNA-KnC3HDZ which are now considered responsive by CNA-KnC3HDZ-8m. Finally, KnC3HDZ has 107 unique direct targets, despite its binding profile being fully contained within that of CNA and KnC3HDZ-8m (Fig. 6f; Supplementary Fig. 8). These direct targets correspond to genes with sites considered responsive by CNA-KnC3HDZ which are now considered non-responsive by CNA-KnC3HDZ-8m. The latter two findings support the notion that START domains mediate the interconversion of responsive and non-responsive binding sites. Taken together, these developmental and functional genomics assays demonstrate that a small number of amino acid substitutions can lead to large-scale changes in the selection and regulation of HD-ZIPIII targets.

Functional divergence of HD-ZIPIII proteins is partly explained by their START domains

Our findings indicate amino acid substitutions in HD-ZIPIII START domains lead to profound changes in target regulation. This supports the idea that START domains may contribute, at least in part, to HD-ZIPIII functional divergence. To formally test this notion, we repurposed the established complementation assay described in ref. 26. In this assay, coding sequences of HD-ZIPIII genes were placed downstream of *REV* regulatory elements, introduced into the *rev-6* mutant⁶⁶, and scored for their ability to complement the *rev-6* mutant phenotype²⁶. Approximately 40% of *rev-6* mutant flowers terminate before setting seed⁶⁶ (Fig. 7a, b), and this phenotype can be fully rescued by the introduction of a *pREV:REV* transgene⁶⁶ (Fig. 7a, b). By contrast, *pREV:CNA* primary transformants are indistinguishable from *rev-6* mutants consistent with functional divergence of CNA from *REV*²⁶ (Fig. 7a, b). We then replaced the CNA START domain with its counterpart from *REV* to create a CNA-AtREV chimera driven by *REV* regulatory elements (*pREV:CNA-AtREV*). Introduction of the *pREV:CNA-AtREV* transgene dramatically reduced the frequency of floral termination in *rev-6* mutants, with only 8% of siliques aborting during

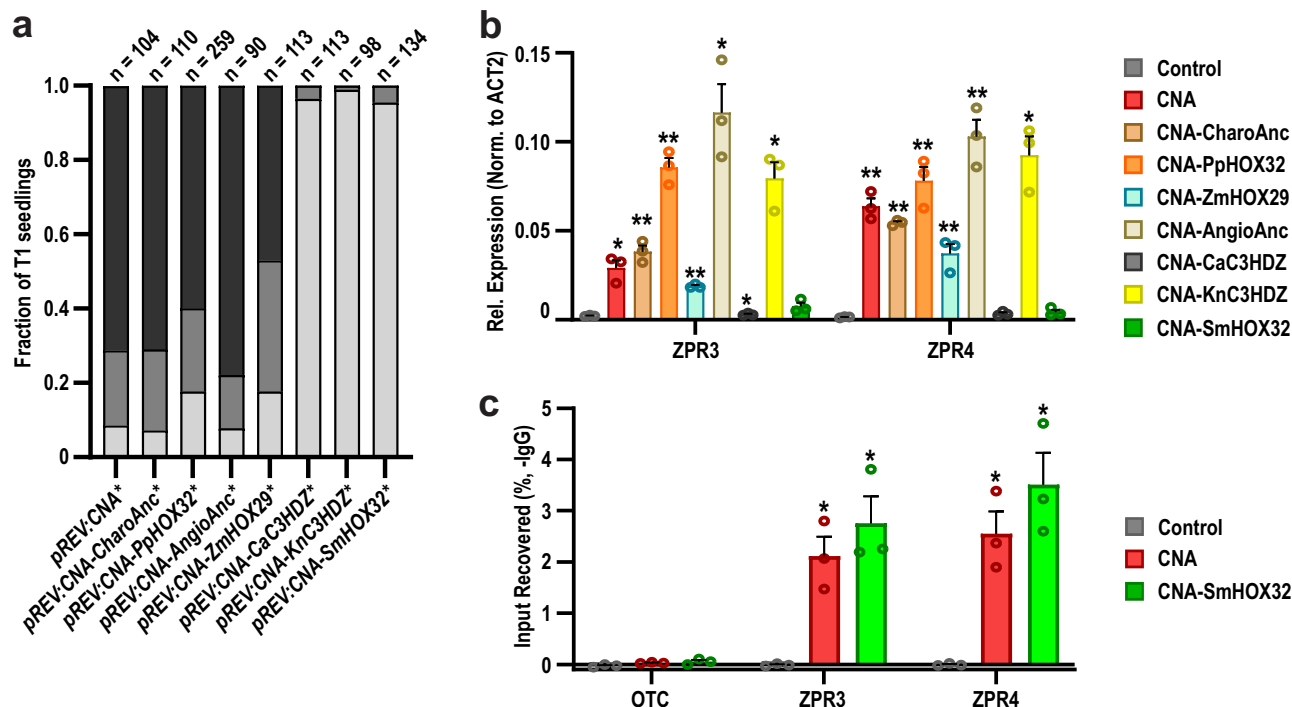


Fig. 4 | Chimeric CNA proteins have distinct effects on development and gene regulation. **a** Phenotypic scoring of primary transformants (n above bar) expressing miR-166 insensitive (*) chimeric variants of CNA driven by the REV promoter. Ectopic CNA* expression leads to severe (black) or intermediate (dark gray) gain-of-function phenotypes, and five chimeric constructs produce similar effects (left-most five bars). The remaining three chimeric constructs do not obviously perturb development as plants appear wildtype (light gray; rightmost three bars). Phenotypes resemble those shown in Fig. 2. **b** CNA robustly activates ZPR3 and ZPR4 in 24 h estradiol-induction experiments while chimeric variants of CNA show a

complex pattern of target induction. **c** CNA and CNA-SmHOX32 both bind to regulatory regions of ZPR3 and ZPR4 and are significantly enriched above the OTC negative control. Nomenclature: non-transgenic control (Control), Charophyte ancestrally reconstructed variant (CharoAnc), *Physcomitrella patens* (PpHOX32), Angiosperm ancestrally reconstructed variant (AngioAnc), *Zea mays* (ZmHOX29), *Chlorokybus atmophyticus* (CaC3HDZ), *Klebsormidium nitens* (KnC3HDZ), and *Selaginella moellendorffii* (SmHOX32). Data are presented as mean values \pm SEM. Statistics are against the non-transgenic control and use three biological replicates per genotype. * = p -value ≤ 0.05 , ** = p -value ≤ 0.01 , two-way paired Student's t -test.

development (Fig. 7a, b). Thus, functional divergence of CNA and REV subclade members is driven, at least in part, by substitutions within their START domains.

Discussion

TF functional divergence drives biological processes ranging from speciation to cellular decision making^{67,68}. A common mechanism underlying this divergence is binding to different sets of loci¹⁻⁶. Here we show that functional divergence of the HD-ZIP III paralogs CNA and PHB is instead driven primarily by differential usage of shared binding sites. CNA and PHB have largely overlapping binding profiles, yet each paralog has hundreds of uniquely regulated targets that affect distinct biological processes (Fig. 1e, g; Supplementary Fig. 7). CNA and PHB also have hundreds of shared, similarly regulated direct targets, consistent with the partial functional redundancy of these paralogs²⁶ (Fig. 1f). Thus, regulation of a given gene depends on whether its local binding site is considered responsive versus non-responsive. Interestingly, the interpretation of binding site identity by HD-ZIP III proteins appears to be controlled, at least in part, by their START domain. For instance, as few as eight amino acid substitutions in the START domain of a chimeric CNA protein were sufficient to drive both gain and loss of regulation across hundreds of bound loci (Fig. 6). In addition, replacing the CNA START domain with that of REV enabled CNA to partly fulfil the developmental functions of its divergent paralog (Fig. 7b). Our findings support a model in which HD-ZIP III TFs use information integrated via their START domain to generate paralog-specific transcriptional outcomes across a network of commonly bound genes (Fig. 7c).

Binding of shared sets of genes by functionally divergent paralogous TFs has been noted previously. For instance, E2F TFs oppositely regulate a common set of cell-cycle related genes²³, while MyoD and Myf5 induce histone modifications and recruit transcriptional machinery, respectively, at shared skeletal muscle specification and differentiation genes⁶⁹. However, unlike the HD-ZIP III family, the activities of these metazoan TFs are spatiotemporally separated by sequential expression dynamics^{23,69}. The plant paralogs FT and TFL1, on the other hand, do spatiotemporally overlap and share a network of commonly bound genes, much like HD-ZIP III TFs²². However, these TFs are thought to function by changing the valence of gene expression rather than making paralog-specific decisions on whether a given gene is to be regulated or not. The HD-ZIP III family is a particularly clear example of this regulatory paradigm given their dramatic overlap in genomic occupancy. However, this mechanism is likely at play in other TF families as well, including those whose functional divergence was previously attributed solely to binding of new targets. Reanalysis of shared binding sites from the perspective of 'responsive versus non-responsive' is likely to yield additional, biologically relevant insights. Related to this, genomes are characterized by prevalent 'non-functional DNA binding events', prompting animal transcription networks to be described as "Continuous Networks" (reviewed in ref. 54). In this model, TFs occupy broad swathes of the genome, and occupancy levels positively correlate with functional consequences for gene expression. However, our analyses indicate occupancy levels are largely uncoupled from target regulation and magnitude of gene expression changes (Fig. 1; Supplementary Figs. 4a, 4b, 5a, 6). Thus, broad occupancy, followed by functional discrimination of responsive

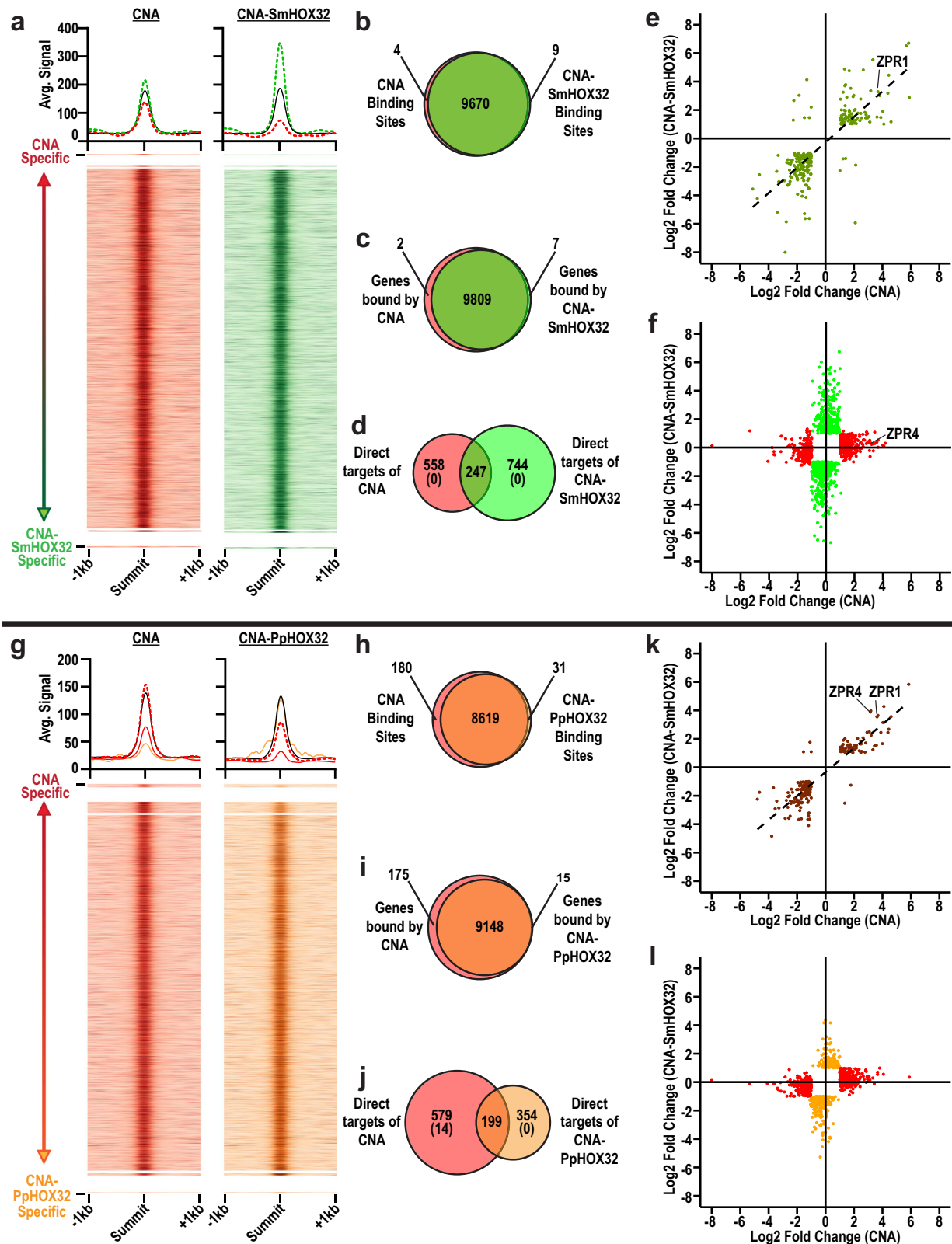


Fig. 5 | Chimeric CNA proteins show widespread differential usage of shared binding sites. a–f CNA vs CNA-SmHOX32 analyses (red vs green). **g–l** CNA vs CNA-PpHOX32 analyses (red vs orange). **a, g** Histograms and heatmaps of ChIP-seq signal intensities. Lines on histograms delineate categories as described in Fig. 1. Heatmaps are separated into categories and subcategories as described in Fig. 1. **b, h** Venn diagrams of sites bound by CNA and its chimeras. **c, i** Venn diagrams of genes bound by CNA and its chimeras. **d, j** Venn diagrams of direct targets of CNA

and its chimeras, i.e. bound in ChIP-seq and differentially expressed in RNA-seq. Numbers in parentheses correspond to direct targets bound specifically by CNA or its chimeras. **e, k** Scatterplots showing differential expression of genes bound and regulated by both CNA and its chimeras. **f** Scatterplot showing differential expression of genes bound by CNA but uniquely regulated by CNA (red) or CNA-SmHOX32 (green). **l** Scatterplot showing differential expression of genes bound by both variants but uniquely regulated by CNA (red) or CNA-PpHOX32 (orange).

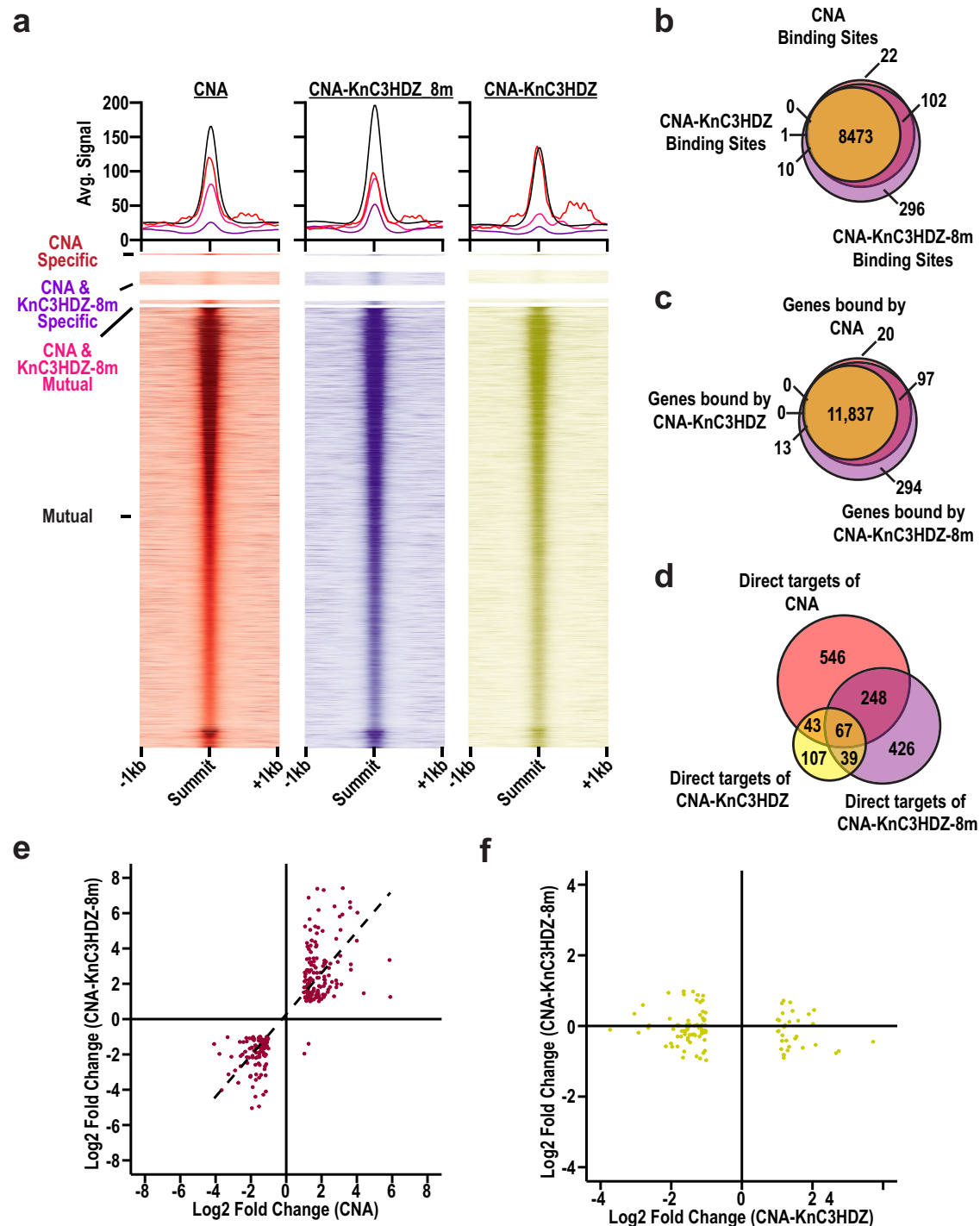


Fig. 6 | Evolutionarily relevant START point mutations alter target selection and regulation. **a** Histograms and heatmaps of ChIP-seq signal intensities. Histogram lines and colors delineate four categories of binding sites: CNA specifically bound (red, solid); CNA-KnC3HDZ_8m specifically bound (purple, solid); bound by CNA and CNA-KnC3HDZ_8m (pink, solid); and bound by all three variants (black, solid). Heatmaps are separated into the same four categories. **b** Venn diagram of

sites bound by CNA and its chimeras. **c** Venn diagram of genes bound by CNA and its chimeras. **d** Venn diagram of direct targets of CNA and its chimeras, i.e. bound in ChIP-seq and differentially expressed in RNA-seq. **e** Scatterplot showing genes bound by all three variants whose expression is regulated only by CNA and CNA-KnC3HDZ_8m (purple). **f** Scatterplot showing genes bound by all three variants whose expression is regulated only by CNA-KnC3HDZ (yellow).

versus non-responsive sites, may be a more complete description of how TFs engage with genomes.

One advantage of paralogs binding to non-overlapping sets of loci is that it minimizes inappropriate cross-regulation. What advantages might come from an alternate strategy in which a family of TFs occupies the same genes but regulates transcriptional outcomes on a

paralog-specific basis? One possibility is suggested by the opposite phenotypes of the *phb phv rev* and *phb phv cna* triple mutants²⁶. The former has a meristem-termination defect while the latter has enlarged fasciated meristems^{26,70}. These phenotypes only emerge when both *PHB* and *PHV* are mutated, suggesting *PHB* and *PHV* buffer the meristem-promotive effects of *REV* and the meristem-repressive

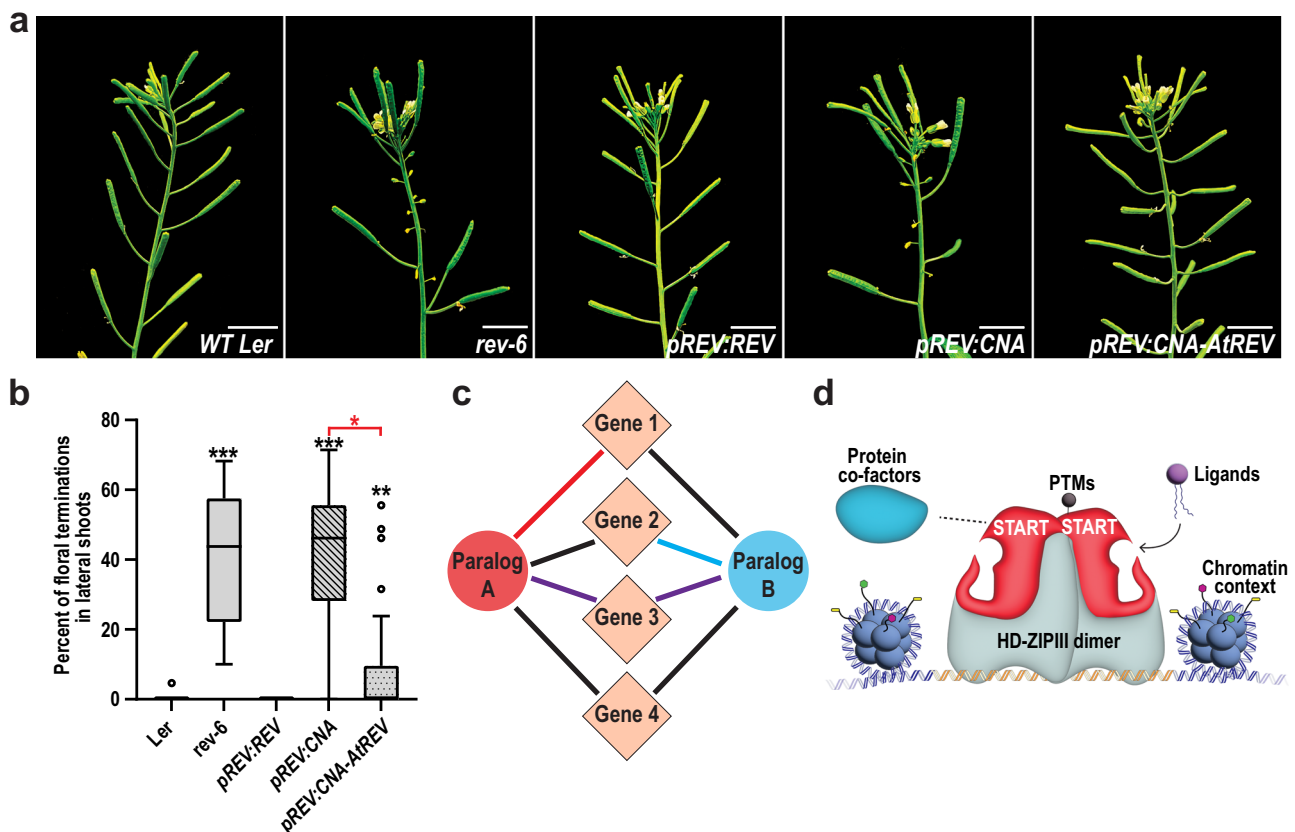


Fig. 7 | Functional divergence of HD-ZIPIII proteins is driven by their START domains which convert a single network architecture into paralog-specific regulons. **a** The *rev-6* mutant has a strong floral termination phenotype⁶⁶. The *pREV:REV* transgene fully complements the *rev-6* mutant phenotype as plants appear wildtype (*Ler*), whereas the *pREV:CNA* transgene does not. The majority of primary transformants carrying a *pREV:CNA-AtREV* transgene appear wildtype indicating near-complete rescue of the *rev-6* mutant phenotype. **b** Quantification of *rev-6* complementation (measured as percentage of floral termination in lateral shoots). Data is presented as a box and whiskers plot using Tukey whiskers. Statistics are against the *Ler* wildtype in all cases except the red asterisks which are compared to *pREV:CNA*. * - p-value $= 7.7 \times 10^{-4}$, ** - 9.3×10^{-7} , *** - p-value $\leq 2.78 \times 10^{-11}$, two-way paired Student's t-test. $n = 24, 25, 16, 30$, and 36 primary transformants scored per genotype (left to right). **c** HD-ZIPIII TF paralog divergence through differential regulation of mutually bound genes. In this simplified model, two members of a TF

family bind a common set of genes. A given gene could be regulated by both TFs (i.e. Gene 3) or unaffected by both TFs (i.e. Gene 4). Alternatively, a gene could be uniquely regulated by one of the two TFs (Gene 1 vs Gene 2) generating paralog-specific regulatory outcomes from a commonly bound genetic network. Line legend: purple (regulated by both paralogs), black (regulated by neither paralog), red (regulated only by paralog A), blue (regulated only by paralog B). **d** Speculative, non-mutually exclusive regulatory mechanisms occurring at shared binding sites (orange DNA). Inputs integrated via START domains could include ligands (purple), post-translational modifications (PTMs; brown), and protein co-factors such as other TFs, chromatin remodelers, or subunits of the general transcriptional machinery (blue). Functional consequences for gene expression may also depend on histone modifications, DNA methylation, degree of chromatin compaction, or other aspects of chromatin context.

effects of CNA. One mechanism by which PHB and PHV could accomplish this buffering is occupation – but not necessarily regulation – of the genes *REV* and *CNA* use to affect meristem size. In this scenario, PHB and PHV would prevent meristem termination in a *rev* single mutant by limiting the effects of the meristem-repressive CNA. Similarly, in a *cna* single mutant, PHB and PHV would prevent meristem enlargement by blocking runaway effects of the meristem-promotive *REV*. This strategy would lend robustness to the transcriptional outcomes of each paralog and therefore the developmental contributions of the family.

Another possible advantage could be ease of network rewiring. For instance, it may be useful for a TF paralog to gain or lose control of entire regulatory modules, or for a diverged paralog to regain certain targets given environmental or evolutionary pressures. This would be mechanistically challenging for paralogs which diverged via changes to genome occupancy. By contrast, a strategy of occupation – but not necessarily regulation – of target genes preserves connections between members of a TF family and their downstream targets (Fig. 7c). Inputs from respective paralogs can then be turned on or off in a coordinated fashion. This could be accomplished by stimuli

operating at organismal or population levels, increasing speed and flexibility of rewiring. Paralog-specific stimuli could then be used to further expand regulatory potential within a given TF family.

The acquisition of additional protein domains by homeobox TFs is thought to have provided new targets and increased specificity of binding⁷¹. Consistent with this, our findings indicate a weak relationship between START domains and the frequency of HD-ZIPIII genomic occupancy, perhaps accomplished by tuning of DNA binding affinities (Figs. 1, 3, 5, 6). More importantly, our analyses suggest that capture of START domains by HD-ZIPIII precursors added an additional layer of regulation at the level of binding site usage. How might differential usage of shared binding sites occur? CNA and PHB have nearly identical endogenous expression patterns⁷², and the estradiol-inducible system generates uniform TF accumulation in all cell types⁴⁹ (Supplementary Fig. 1). This argues against spatiotemporal separation of requisite co-factor(s) as a mechanistic explanation. However, formally ruling this out will require single-cell resolution assays that control for heterogeneity inherent to whole-tissue experiments. Instead, one potential explanation for distinct START-directed licensing of CNA and PHB could be paralog-specific ligands. For instance, the PHB START

domain binds to phosphatidylcholine, and this binding is required for PHB to occupy *ZPR* regulatory regions and strongly activate their transcription⁴⁹. The binding of other ligands could conceivably uncouple these two features, creating situations in which a paralog binds to a given gene but cannot affect its transcription. A second non-mutually exclusive mechanism to facilitate the interpretation of binding site identity is paralog-specific interacting partners. These partners could directly interact with the START domain, as observed for other StArkin-containing proteins^{42,45,63,73–78}. Co-factor binding could also occur at other regions of the HD-ZIPIII protein and be modulated by START-directed allosteric effects. In either scenario, the accrual of amino acid substitutions throughout evolution would lead to distinct interactomes for each paralog. Interactions could then be used to convert non-responsive binding sites into responsive ones and vice versa. Numerous mechanisms could mediate this interconversion including opening or closing of chromatin, formation of higher-order chromatin structures, and recruitment and/or stabilization of basal transcriptional machinery, among others.

We note that StArkin domains use a remarkably diverse array of regulatory mechanisms and are present in gene families across the tree of life^{46,59–63,79–81}. Their ubiquity and flexible regulatory nature may make StArkin domains particularly amenable tools for evolution to generate functional divergence. Studies into other StArkin-containing multigene families will be instrumental in testing this intriguing hypothesis.

Finally, we mention some caveats of the work presented here. First, ectopic inducible systems have numerous advantages. For instance, they can circumvent confounding morphological defects of mutants or stable overexpression lines, standardize TF dosage, and eliminate spatiotemporal variance of co-factors. However, supplementary experiments using endogenous promoters are critical for findings to be properly integrated back into a developmental context. Second, pairing ChIP-seq with transcriptome profiling is an efficient approach to identify direct targets (e.g.⁸²). However, these assays can suffer from false positives (Type I errors) and false negatives (Type II errors), influenced in part by induction time prior to transcriptome profiling. Shorter inductions capture strongly activated targets but are more likely to miss repressed targets that require additional time to be detected. By contrast, longer inductions are better positioned to capture both activated and repressed targets but will miss genes whose regulation has been dampened by network feedback. Combining both time frames is a good strategy to capture the full range of likely TF direct targets.

Methods

Arabidopsis Growth Conditions

Arabidopsis thaliana (Col-0 and Ler-0 ecotype) seedlings were grown at 22 °C under long-day conditions on soil or 1% agarose Murashige and Skoog medium plates (pH 5.7). Inductions were performed by spraying 9-day-old T2 seedlings on agarose plates 10 times with a 50 μM β-Estradiol, 1% DMSO, and 0.005% Silwet L-77 solution.

Molecular cloning and plant transformation

PHB constructs (*pOlexA:PHB** & *pOlexA:PHB*SDA*) used in this study had been previously generated⁴⁹. *CNA-mCitrine-3xFLAG* was constructed via Gibson assembly (NEB) in a pCR8/GW/TOPO cloning plasmid backbone. The *CNA-SDA* CDS was constructed via Gibson assembly with primers that remove the START domain (451 – 1149 bp from start codon) but retain the 21-nt regulatory miR166 binding site, GGAATGAAGCCTGGACCGGAT (553 – 573 bp from start codon) through overlapping primers. To generate the *CNA-SDmut* CDS, the mutated variant of the START domain was synthesized (GeneArt), replacing the amino acid sequences RDFWLLR and RAEML to GAVVGVG and VAAGV, respectively⁴⁹ using the following nucleotide substitutions: CGCGATTCTGGCTGTTACGT (802 – 822 bp from start

codon) to GGTGCCGTCGTAGGAGCAGGC and AGAGCAGAGATGCTT (922 – 936 bp from start codon) to GTGGCGGCCGGCGTC. miR166 insensitivity (*CNA**, *CNA-SDmut**, and *CNA-SDA**) was introduced by PCR amplification of the START domain in two fragments using mismatched primers to introduce a silent SNP within the 21nt miR166 recognition site (GGAATGAAGCCTGGTCCGGAT to GGAATGAAGCC TGGACCGGAT, 553 – 573 bp from start codon) followed by reassembly via Gibson assembly.

CNA START domain substitutions were generated via Gibson assembly by amplifying the CNA CDS and pCR8 backbone to exclude the START domain (451 – 1149 bp from start codon) and insertion of orthologous CNA START domains made miR166 insensitive, if necessary, as previously described from cDNA (KnC3HDZ, PpHOX32, SmHOX32, ZmHOX29, AtREV) or via gene synthesis (CaC3HDZ, CharoAnc, AngioAnc). Inducible pOlexA constructs were generated by subcloning into the two component system (pUBQ:XVE pOlexA:CDS; modified pMDC7) binary vector⁴⁹ via LR Gateway reaction. Constructs with the regulatory elements of CNA (*pCNA*, 3728 bp upstream⁶⁴) or REV (*pREV*, 4975 bp upstream of the start codon upstream, 1026 bp downstream of stop codon) were assembled via Gibson assembly and subcloned into a pB7GW binary vector. *A. thaliana* was transformed with these constructs via *Agrobacterium* mediated floral dip transformation. Cloning primers and synthesized sequences can be found in Supplementary Data 4.

Confocal imaging and western blotting

Induced seedlings were incubated in a 0.5 μg/ml propidium iodide solution for 5 minutes to stain root cell plasma membranes. Seedlings were washed 3 times for 5 minutes each with dH₂O. The hypocotyl was excised, and roots were wet mounted to a glass slide with coverslip. Roots were examined under a Nikon Eclipse Ni-E microscope with C2si confocal system using a 60x Nikon Plan APO VC 60x / 1.40 oil immersion objective lens with Cargille Type A immersion oil. The mCitrine YFP tag was excited using a wavelength of 526 nm, and emission was captured between 490 – 550 nm. Propidium iodide was excited with the same wavelength as mCitrine, 526 nm, and emission was captured between 570 – 620 nm.

Induced seedlings were ground in 2x Laemmli buffer, separated on an SDS-PAGE gel, and blotted to nitrocellulose membrane. Membranes were probed with either FLAG primary antibody (Millipore Sigma, Monoclonal ANTI-FLAG® M2 antibody produced in mouse, Cat No. F1804-1MG, Source No. SLCQ9256, PCode: 1003602111, 1:1000 dilution) or Actin primary antibody (Millipore Sigma, Anti-Actin (plant) antibody Mouse monoclonal, Cat No. A0480-25UL, Source No. 0000240291, Clone 10-B3, Lot No. 0000269837, 1:5000 dilution). The same secondary was used for both blots (Jackson ImmunoResearch, Peroxidase AffiniPure™ Donkey Anti-Mouse IgG (H + L), Cat No. 715-035-150, Polyclonal, RRID: AB_2340770, 1:5000 dilution).

Chromatin Immunoprecipitation and qPCR

Induced seedlings (24 hrs post-induction) were collected off the plates and moved to a conical tube with 30 ml of a 2% formaldehyde (ThermoFisher catalog# 28906) crosslinking solution. Uncapped tubes were placed in a vacuum chamber for 30 minutes at 20–25 mmHg vacuum pressure, with a brief shake after the first 15 minutes. The crosslinking reaction was inactivated by adding 2 ml of 2 M glycine to the formaldehyde solution, briefly shaking, and placing the uncapped tubes into the vacuum for 5 additional minutes. The inactivated formaldehyde solution was poured out and the seedlings were rinsed three times with 1xPBS. Paper towels were used to dry the seedling, removing as much excess water as possible. The dried seedlings were separated into individual 500 mg biological replicates and flash-frozen with liquid nitrogen.

Seedlings were ground using a mortar and pestle on dry ice, and the powdered tissue was transferred to a conical tube containing 8 ml

of nuclear isolation buffer (10 mM HEPES pH 8.0, 1 M Sucrose, 5 mM KCl, 5 mM MgCl₂, 0.6% Triton, 0.4 mM PMSF, 1x cOmplete™ EDTA free protease inhibitor) and rotated for 15 minutes at 4 °C. The solution was filtered through two layers of Miracloth into a separate conical tube and centrifuged at 3000 g for 15 minutes at 4 °C. The supernatant was discarded, and the pelleted nuclei were resuspended in 1 ml of nuclei wash buffer (10 mM Tris pH 8.0, 250 mM Sucrose, 10 mM MgCl₂, 1 mM EDTA, 1% Triton, 1 mM PMSF, 1x cOmplete™ EDTA free protease inhibitor), transferred to a microcentrifuge tube, and centrifuged at 12,000 g for 10 minutes at 4 °C. The supernatant was discarded, and the nuclei were resuspended in nuclear lysis buffer (20 mM Tris pH 8.0, 2 mM EDTA, 0.01% SDS, 1 mM PMSF, 1x cOmplete™ EDTA free protease inhibitor) and transferred to a 1 ml Covaris milliTUBE with AFA Fiber. Samples were sonicated using a Covaris E220 at 150 W peak power, 20% duty factor, 200 cycles/burst for 6 minutes. The sonicated chromatin was transferred to a new microcentrifuge tube and centrifuged at 12,000 g for 10 minutes at 4 °C to pellet insoluble debris. 920 µl of the supernatant was moved to a new microcentrifuge tube, and 30 µl of 5 M NaCl and 20 µl of 30% Triton were added to the supernatant. The sonicated chromatin was divided into two microcentrifuge tubes (420 µl each), and 42 µl of the remaining chromatin was retained for a 10% input control. 2 µl of M2 mouse monoclonal anti-FLAG (Sigma, F3165) or mouse IgG was added to the tubes and rotated overnight at 4 °C. Pierce Protein A/G magnetic beads (ThermoFisher Scientific, 88802), were equilibrated with the nuclear lysis buffer with added NaCl and Triton as described previously and resuspended to their initial volume 40 µl of equilibrated magnetic beads were added to each tube, and incubated for 2 hrs rotating at 4 °C. Tubes were then placed on a magnetic rack and the supernatant was removed. Beads were washed twice for 5 minutes each rotating at 4 °C with the subsequent buffers, Low Salt Wash Buffer (150 mM NaCl, 0.1% SDS, 1% Triton x-100, 2 mM EDTA, 20 mM Tris pH 8.0), High Salt Wash Buffer (500 mM NaCl, 0.1% SDS, 1% Triton x-100, 2 mM EDTA, 20 mM Tris pH 8.0), LiCl Wash Buffer (250 mM LiCl, 1% IGEAL, 0.1% SDS, 1 mM EDTA, 10 mM Tris pH 8.0) and TE Buffer (10 mM Tris pH 8.0, 1 mM EDTA, 0.1% IGEAL). 250 µl of preheated (65 °C) Elution Buffer (1% SDS, 0.1 M NaHCO₃) was added to the beads and 10% input and incubated at 65 °C for 15 minutes. The supernatant was transferred to a new microcentrifuge tube, 10 µl of 5 M NaCl was added to each tube, and placed at 65 °C overnight. The heat block was reduced to 45 °C and 10 µl of 0.5 M EDTA, 20 µl 1 M Tris pH 7.0, 1 µl of 20 mg/ml Protease K, and 1 µl Rnase (50 ng/µl) were added to each and incubated for 1 h at 45 °C. ChIP DNA was recovered using a Zymo DNA Clean and Concentrator-25 kit.

Recovered DNA was quantified using qPCR with 10 µl of Sybr-Green, 1 µl of combined 10 µM forward and reverse primer sets, 5 µl dH₂O, and 4 µl of DNA with two technical replicates for each tested target (*OTC*, *ZPR3*, and *ZPR4*). Data was plotted using GraphPad.

ChIP-seq and bioinformatics

Libraries were constructed using the NEBNext Ultra II DNA Library kit and single-end sequenced for 100 cycles using the Illumina NextSeq500/550 located on-site at the Penn Epigenetics Institute. Read files were merged using command prompt in Windows OS, using the command: "Type file_L001.fastq.gz file_L002.fastq.gz file_L003.fastq.gz file_L004.fastq.gz > merged-file.fastq.gz"; if necessary and uploaded to the Galaxy web-based platform⁸³. Reads were trimmed using Trimmomatic⁸⁴ and performing the initial ILLUMINACLIP using adaptor sequences for TruSeq3 (single-ended, for MiSeq and HiSeq) and the trimmomatic operation "Sliding window trimming" averaging across 4 bases with an average required score of 20. Trimmed sequencing files were aligned to the TAIR10 Arabidopsis thaliana genome using Bowtie2⁸⁵ under the default settings. Number of mapped read statistics can be found in Supplementary Data 5.

Peaks were called using MACS2⁸⁶ running the individual FLAG IP samples individually against their respective 10% input control, using

the effective genome size of *D. melanogaster* equal to 1.2e8 (equivalent to the genome size of *A. thaliana*) and building a shifting model with a lower mfold bound = 5, upper mfold bound = 50, and a band width for picking regions to compute fragment size = 300, and the minimum FDR (q-value) cutoff for peak detection set to 0.01. Window size was set to 150 bp then BED files were analyzed using a publicly available Irreproducibility Discovery Rate (IDR) python package (<https://github.com/nboley/idr>). An IDR of 0.01 was used (following ChIP-Hub guidelines for plant genomic analyses⁸⁷), and we kept only the peaks that were called in all replicates of a given genotype. Diffbind⁸⁸ was then used to analyze differential peak signals between samples, and further analyzed with custom R scripts to perform a 3-way comparisons of DiffBind results and additional outputs to include bound genes using the TAIR10 gff3 genome annotation with further integration of DEseq2 RNA-seq results (Supplementary Data 2, 3) using R packages profileplyr⁸⁹, dplyr⁹⁰, reshape⁹¹, stringr⁹², and rtracklayer⁹³. A peak was assigned to a gene if it was <2 kb upstream of its transcription start site, within the gene body, or <1 kb downstream of its transcription termination site. Histograms were plotted using GraphPad (Prism) and heatmaps were generated using ggplot2⁹⁴.

MEME-ChIP motif identification

BED coordinates of the binding sites centralized on the summits for each transcription factor were uploaded as custom tracks to the UCSC genome browser⁹⁵ and the TAIR10 genome. Using Table Browser feature⁹⁶, DNA sequences in fasta format of these 400 bp summits, and an additional 50 bp upstream and downstream (total 500 bp centralized on the summit) were extracted. These sequences were then uploaded to the MEME-ChIP discovery tool⁹⁷ of the MEME suite 5.5.5⁹⁸. The "Input the motifs" setting was set to known motifs of "ARABIDOPSIS (*Arabidopsis thaliana*) DNA" and "DAP motifs⁵⁰". The 2nd order model of sequences was used as the background model, looking for motifs with widths between 6 and 11 in MEME and STREME. Zero or one occurrence per sequence was selected for the expected motif site distribution and MEME was restricted to search for palindromes only.

FIMO identification of bound motifs and DNA shape prediction

Individual instances of the top motif for CNA and PHB mutual sites (VTAATNATTAB for CNA; TAATRATKATD for PHB) were identified using FIMO⁹⁹ in the MEME suite 5.5.5⁹⁸, with a p-value cutoff of 5e-4 and input sequences of Ensemble Plants Genomes and Proteins: Arabidopsis thaliana (version 57). The output bed file of genomic coordinates with this motif were uploaded as a custom track to UCSC genome browser⁹⁵ and the table browser tool⁹⁶ was used to find any overlap of these motif genome coordinates with regions 75 bp upstream or downstream from the summits of mutual CNA and PHB peaks.

Table browser was used to output sequences of the bound and unbound motifs in the genome with an addition 8 bp upstream and downstream. Fasta sequences were then converted into DNA shapes using the DNashapeR package¹⁰⁰ in R. Two-tailed Z-tests were then performed for predicted minor groove width (MGW), helical twist (HelT), propeller twist (ProT) and roll at individual position along the sequences between the shapes of bound and unbound motifs and flanking sequences, and q-value was calculated using Bonferroni correction¹⁰¹.

RNA-seq and bioinformatics

All seedlings were induced (or mock induced) two hours after the beginning of the light cycle to minimize circadian differences between genotypes. Seedlings were then collected 24 hrs post treatment and flash frozen in liquid nitrogen. Total RNA was extracted using TRIzol Reagent (ThermoFisher Scientific, 15596026) and treated with DNase. Sequencing was performed by Lexogen (Austria) using the QuantSeq 3' mRNA protocol which prioritizes the 3' end of the mRNA transcripts.

Sequencing files were demultiplexed by Lexogen, and reads were uploaded to the Galaxy web-based platform⁸³. Adaptor sequences were trimmed using Trimmomatic⁸⁴, performing the initial ILLUMINACLIP using adaptor sequences for TruSeq3 (single-ended, for MiSeq and HiSeq) and the trimmomatic operation “Sliding window trimming” averaging across 4 bases with an average required score of 20. Remaining polyA tail sequences were removed with a subsequent rerun of trimmomatic, performing an initial ILLUMINACLIP with custom adaptor (PolyA20) sequences of “AAAAAAAAAAAAAAAAAAAA” and the trimmomatic operation “Sliding window trimming” averaging across 4 bases with an average required score of 20.

Reads were aligned to the TAIR10 *A. thaliana* reference genome using RNA STAR¹⁰² with the following settings: Length of the SA pre-indexing string = 12; Maximum number of alignments to output a read's alignment results, plus 1 = 200; Minimum overhang for spliced alignments = 8; Minimum overhang for annotated spliced alignments = 1; Maximum number of mismatches to output an alignment, plus 1 = 999; Maximum ratio of mismatches to mapped length = 0.6; Minimum intron size = 20; Maximum intron size = 1000; Maximum gap between two mates = 1000, Maximum number of collapsed junctions = 5000000. Number of mapped read statistics can be found in Supplementary Data 5.

Gene expression was then measured using featureCounts¹⁰³ using the TAIR10 gene annotation gff3 and TAIR10 transcriptome using the following settings: Specify stand information = Forward; GFF Feature Type filter = gene; GFF gene identifier = ID; Exon-exon junctions = Count reads supporting each exon-exon junction (-J).

Results of featureCounts were combined into a single csv file and analyzed with DESeq2¹⁰⁴ in R under the default settings. Comparison of respective induced and mock samples of a genotype were run through the lfcShrink function to account for the log fold change standard error across the samples using type = “ashr”¹⁰⁵, shrinking the Log2 fold change values. Differentially expressed genes (DEGs) were determined using DESeq2, comparing induced genotypes to their respective mock control. Significant DEGs were identified as having ≥ 2 or ≤ -2 fold change in expression levels, and a q -value ≤ 0.1 . Significant DEGs were then referenced to bound genes identified as part of the ChIP-seq analysis to identify true targets. Data was plotted using the ggplot2 package⁹⁴ in R.

Gene ontology analysis

Gene ontology analysis was performed as in¹⁰⁶. In brief, a custom wrapper around the topGO package version 2.54.0¹⁰⁷ was run on upregulated or downregulated direct targets for PHB specific, CNA specific and mutual targets. The gene universe was all detected transcripts in the RNA-Seq dataset. Packages were obtained from CRAN or Bioconductor version 3.7¹⁰⁸. The complete output of gene ontology terms can be found in Supplementary Data 6.

qRT-PCR

cDNA was generated from 500 ng total RNA from 24 hr post-induced seedlings SuperScriptIV Reverse Transcriptase kit (ThermoFisher). cDNA was diluted 1:1 with dH2O and prepared in a 20ul volume (10ul of SsoAdvanced Universal SYBR Green Supermix (Bio-Rad), 8ul of dH2O, 1ul of cDNA, and 1ul of a 10uM mixture of the appropriate forward and reverse primer). Ct values were measured using a Eppendorf Mastercycler RealPlex² with 40 cycles. Averaged Ct values of the technical replicates were normalized ACTIN 2 (ACT2) by subtracting the averaged Ct value of ACT2. Relative expression levels normalized to ACT2 were calculated with the formula: $2^{-(1 \times \text{normalized value})}$. Data was plotted with Graphpad (Prism).

Ancestral sequence reconstruction

Thirty CNA orthologs were identified using reciprocal best hit with BLASTp of the National Center for Biotechnology Information (NCBI)

toolkits¹⁰⁹. These sequences consist of five eudicots, five monocots, four gymnosperms, four ferns, five bryophytes, three lycophytes, and four charophytes (Supplementary Fig. 10). Full length amino acid sequences were aligned using the ClustalW algorithm in MEGA X^{110,111}. Phylogeny was inferred by Bayesian Markov Chain Monte Carlo (BMC) analysis using MrBayes (Version 3)^{112,113} and the maximum likelihood method with Jones-Taylor-Thornton (JTT, gamma distribution = 4) model¹¹⁴. Predictions were made with 100,000 generations, sampling every 100 and the first 250 samples discarded as burn-in, achieving a standard deviation of 0.0017. Bayesian posterior probability scores of 1.0 suggested high probability of inferring sequences at the nodes where ancestral sequences of interest were located. Ancestral sequences were predicted using the maximum likelihood method with JTT + 4 gamma model in MEGAX. Coding sequences were assigned using the *A. thaliana* CNA coding sequences as a template, retaining the codons of conserved residues, and assigning codons for non-conserved residues. These sequences were ordered through GeneArt as synthesized DNA in vectors.

Multiple sequence alignment to generate KfC3HDZ(8 m)

CNA ortholog amino acid sequences were aligned using the ClustalW algorithm¹¹⁰ in the MEGA X software¹¹¹. Conserved aligned residues were identified using R and “msa” package¹¹⁵ from Bioconductor. A 70% consensus threshold score was used to identify conservation in 5 of the 8 aligned START domains, identifying residues conserved in the START domains of CNA, AngioAnc, ZmHOX29, PpHOX32, and CharoAnc, but not conserved in SmHOX32, KnC3HDZ, or CaC3HDZ.

Reporting summary

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

Data availability

The raw ChIP-seq and RNA-seq data generated in this study have been deposited in the NCBI Gene Expression Omnibus database under accession code [GSE253676](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE253676). The ChIP-seq, RNA-seq, gene ontology, and genetic data generated in this study are provided in the Supplementary Information/Source Data file. Source data are provided with this paper. Supplemental Dataset files and the Source Data file are available at [10.5281/zenodo.13900601](https://zenodo.org/record/13900601). Accession codes for *Arabidopsis* genes: PHB [<https://www.arabidopsis.org/gene?key=27627>], CNA [<https://www.arabidopsis.org/locus?key=30526>], REV [<https://www.arabidopsis.org/locus?key=134901>], ZPR1 [<https://www.arabidopsis.org/locus?key=33116>], ZPR3 [<https://www.arabidopsis.org/locus?key=37164>], ZPR4 [<https://www.arabidopsis.org/locus?key=1501129729>]. Accession codes for CNA orthologs in other species: *Zea mays* (ZmHOX29, <https://www.uniprot.org/uniprotkb/Q147T1/entry>), *Selaginella moellendorffii* (SmHOX32, <https://www.uniprot.org/uniprotkb/D8QNI3/entry>), *Physcomitrella patens* (PpHOX32, https://www.ncbi.nlm.nih.gov/protein/XP_024399219.1), *Klebsormidium nitens* (KnC3HDZ, <https://www.uniprot.org/uniprotkb/A0A1Y1IAZ9/entry>), and *Chlorokybus atmophyticus* (CaC3HDZ, AZZW_21737, ref. 116). Source data are provided with this paper.

Code availability

Relevant R codes are available at github.com/HusbandsLab and <https://doi.org/10.5281/zenodo.13900094>.

References

- Berger, M. F. et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **27**, 1266–1276 (2018).
- Shen, N. et al. Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding. *Cell Syst.* **6**, 470–483 (2018).

3. Wei, G. H. et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **29**, 2147–2160 (2010).
4. Siggers, T., Reddy, J., Barron, B. & Bulyk, M. L. Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding. *Mol. Cell* **55**, 640–648 (2014).
5. Rogers, J. M. & Bulyk, M. L. Diversification of transcription factor-DNA interactions and the evolution of gene regulatory networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **10**, e1423 (2018).
6. Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L. & Bulyk, M. L. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc. Natl Acad. Sci. USA* **110**, 12349–12354 (2013).
7. de Mendoza, A. et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl Acad. Sci. USA* **110**, E4858–66 (2013).
8. Ohno, S. *Evolution by Gene Duplication*. Springer-Verlag, New York (1970).
9. De Kegel, B. & Ryan, C. J. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS Genet* **15**, e1008466 (2019).
10. Li, W. H., Yang, J. & Gu, X. Expression divergence between duplicate genes. *Trends Genet* **21**, 602–607 (2005).
11. Shiu, S. H., Shih, M. C. & Li, W. H. Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol.* **139**, 18–26 (2005).
12. Edger, P. P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**, 699–717 (2009).
13. Brodsky, S. et al. Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Mol. Cell* **79**, 459–471.e4 (2020).
14. Feng, S. et al. Transcription factor paralogs orchestrate alternative gene regulatory networks by context-dependent cooperation with multiple cofactors. *Nat. Commun.* **13**, 3808 (2022).
15. Gera, T., Jonas, F., More, R. & Barkai, N. Evolution of binding preferences among whole-genome duplicated transcription factors. *Elife* **11**, e73225 (2022).
16. Porcelli, D., Fischer, B., Russell, S. & White, R. Chromatin accessibility plays a key role in selective targeting of Hox proteins. *Genome Biol.* **20**, 115 (2019).
17. Baker, C. R., Hanson-Smith, V. & Johnson, A. D. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**, 104–108 (2013).
18. Slattery, M. et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270–82 (2011).
19. Sielemann, J., Wulf, D., Schmidt, R. & Bräutigam, A. Local DNA shape is a general principle of transcription factor binding specificity in *Arabidopsis thaliana*. *Nat. Commun.* **12**, 6549 (2021).
20. Thirulogachandar, V. et al. Dosage of duplicated and anti-functionalized homeobox proteins influences spikelet development in barley. *bioRxiv*; <https://doi.org/10.1101/2021.11.08.467769> (2021).
21. Wickland, D. P. & Hanzawa, Y. The flowering locus t/terminal flower 1 gene family: functional evolution and molecular mechanisms. *Mol. Plant* **8**, 983–997 (2015).
22. Zhu, Y. et al. TERMINAL FLOWER 1-FD complex target genes and competition with FLOWERING LOCUS T. *Nat. Commun.* **11**, 5118 (2020).
23. Stevaux, O. & Dyson, N. J. A revised picture of the E2F transcriptional network and RB function. *Curr. Opin. Cell Biol.* **14**, 684–691 (2002).
24. Carlsbecker, A. & Helariutta, Y. Phloem and xylem specification: pieces of the puzzle emerge. *Curr. Opin. Plant Biol.* **8**, 512–517 (2005).
25. Ramachandran, P., Carlsbecker, A. & Etchells, J. P. Class III HD-ZIPs govern vascular cell fate: an HD view on patterning and differentiation. *J. Exp. Bot.* **68**, 55–69 (2017).
26. Prigge, M. J. et al. Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *Plant Cell* **17**, 61–76 (2005).
27. Hawker, N. P. & Bowman, J. L. Roles for Class III HD-Zip and KANADI genes in *Arabidopsis* root development. *Plant Physiol.* **135**, 2261–2270 (2004).
28. McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).
29. McConnell, J. R. et al. Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature* **411**, 709–713 (2001).
30. Kelley, D. R., Skinner, D. J. & Gasser, C. S. Roles of polarity determinants in ovule development. *Plant J.* **57**, 1054–1064 (2009).
31. Sebastian, J. et al. PHABULOSA controls the quiescent center-independent root meristem activities in *Arabidopsis thaliana*. *PLoS Genet* **11**, e1004973 (2015).
32. Floyd, S. K., Zalewski, C. S. & Bowman, J. L. Evolution of class III homeodomain-leucine zipper genes in streptophytes. *Genetics* **173**, 373–388 (2006).
33. Byrne, M. E. Shoot meristem function and leaf polarity: the role of class III HD-ZIP genes. *PLoS Genet* **2**, e89 (2006).
34. Zhong, R. & Ye, Z. H. Regulation of HD-ZIP III Genes by MicroRNA 165. *Plant Signal Behav.* **2**, 351–353 (2007).
35. Sakaguchi, J. & Watanabe, Y. miR165/166 and the development of land plants. *Dev. Growth Differ.* **54**, 93–99 (2012).
36. Du, Q. & Wang, H. The role of HD-ZIP III transcription factors and miR165/166 in vascular development and secondary cell wall formation. *Plant Signal Behav.* **10**, e1078955 (2015).
37. Kim, Y. S. et al. HD-ZIP III activity is modulated by competitive inhibitors via a feedback loop in *Arabidopsis* shoot apical meristem development. *Plant Cell* **20**, 920–933 (2008).
38. Iyer, L. M., Koonin, E. V. & Aravind, L. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. *Proteins* **43**, 134–144 (2001).
39. Clark, B. J. The mammalian START domain protein family in lipid transport in health and disease. *J. Endocrinol.* **212**, 257–275 (2012).
40. Soccio, R. E. & Breslow, J. L. StAR-related lipid transfer (START) proteins: mediators of intracellular lipid metabolism. *J. Biol. Chem.* **278**, 22183–22186 (2003).
41. Stocco, D. M. Tracking the role of a star in the sky of the new millennium. *Mol. Endocrinol.* **15**, 1245–1254 (2001).
42. Du, X. et al. Functional interaction of tumor suppressor DLC1 and caveolin-1 in cancer cells. *Cancer Res.* **72**, 4405–4416 (2012).
43. Dittrich, M. et al. The role of *Arabidopsis* ABA receptors from the PYR/PYL/RCAR family in stomatal acclimation and closure signal integration. *Nat. Plants* **5**, 1002–1011 (2019).
44. Zhao, Y. et al. *Arabidopsis* Duodecuple Mutant of PYL ABA Receptors Reveals PYL Repression of ABA-Independent SnRK2 Activity. *Cell Rep.* **23**, 3340–3351.e5 (2018).
45. Park, S. Y. et al. Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* **324**, 1068–1071 (2009).
46. Dresden, C. E., Ashraf, Q. & Husbands, A. Y. Diverse regulatory mechanisms of StARKin domains in land plants and mammals. *Curr. Opin. Plant Biol.* **64**, 102148 (2021).
47. Shively, C. A., Liu, J., Chen, X., Loell, K. & Mitra, R. D. Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc. Natl Acad. Sci. USA* **116**, 16143–16152 (2019).
48. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1–8 (2017).

49. Husbands, A. Y. et al. The START domain potentiates HD-ZIPIII transcriptional activity. *Plant Cell* **35**, 2332–2348 (2023).
50. O'Malley, R. C. et al. Cistrome and epicistrome features shape the regulatory DNA Landscape. *Cell* **165**, 1280–1292 (2016).
51. Sessa, G., Steindler, C., Morelli, G. & Ruberti, I. The Arabidopsis Athb-8, -9 and -14 genes are members of a small gene family coding for highly related HD-ZIP proteins. *Plant Mol. Biol.* **38**, 609–622 (1998).
52. Rohs, R. et al. The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
53. Li, J. et al. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.* **45**, 12877–12887 (2017).
54. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21**, 611–26 (2011).
55. Reinhardt, B. J. et al. Establishing a framework for the Ad/abaxial regulatory network of Arabidopsis: ascertaining targets of class III homeodomain leucine zipper and KANADI regulation. *Plant Cell* **25**, 3228–3249 (2013).
56. Fernandez, P. C. et al. Genomic targets of the human c-Myc protein. *Genes Dev.* **17**, 1115–1129 (2003).
57. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* **10**, 605–616 (2009).
58. Jaini, S. et al. Transcription Factor Binding Site Mapping Using ChIP-Seq. *Microbiol. Spectr.* **2**, <https://doi.org/10.1128/microbiolspec.MGM2-0035-2013> (2014).
59. Schrick, K., Nguyen, D., Karłowski, W. M. & Mayer, K. F. START lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors. *Genome Biol.* **5**, R41 (2004).
60. Mukherjee, T. et al. The START domain mediates Arabidopsis GLABRA2 dimerization and turnover independently of homeodomain DNA binding. *Plant Physiol.* **190**, 2315–2334 (2022).
61. Iida, H., Yoshida, A. & Takada, S. ATML1 activity is restricted to the outermost cells of the embryo through post-transcriptional repressions. *Development* **146**, dev169300 (2019).
62. Nagata, K., Ishikawa, T., Kawai-Yamada, M., Takahashi, T. & Abe, M. Ceramides mediate positional signals in Arabidopsis thaliana protoderm differentiation. *Development* **148**, dev194969 (2021).
63. Kanno, K. et al. Interacting proteins dictate function of the minimal START domain phosphatidylcholine transfer protein/StarD2. *J. Biol. Chem.* **282**, 30728–30736 (2007).
64. Yamada, T., Sasaki, Y., Hashimoto, K., Nakajima, K. & Gasser, C. S. CORONA, PHABULOSA and PHAVOLUTA collaborate with BELL1 to confine WUSCHEL expression to the nucellus in Arabidopsis ovules. *Development* **143**, 422–426 (2016).
65. Ochoa, I. et al. JL. Mutations in the microRNA complementarity site of the INCURVATA4 gene perturb meristem function and adaxialize lateral organs in arabidopsis. *Plant Physiol.* **141**, 607–619 (2006).
66. Zhong, R. & Ye, Z. H. IFL1, a gene regulating interfascicular fiber differentiation in Arabidopsis, encodes a homeodomain-leucine zipper protein. *Plant Cell* **11**, 2139–2152 (1999).
67. Pougach, K. et al. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Erratum : Nat. Commun.* **6**, 6543 (2015).
68. Teichmann, S. A. & Babu, M. M. Gene regulatory network growth by duplication. *Nat. Genet.* **36**, 492–496 (2004).
69. Conerly, M. L., Yao, Z., Zhong, J. W., Groudine, M. & Tapscott, S. J. Distinct Activities of Myf5 and MyoD Indicate Separate Roles in Skeletal Muscle Lineage Specification and Differentiation. *Dev. Cell* **36**, 375–385 (2016).
70. Lee, C. & Clark, S. E. A WUSCHEL-Independent Stem Cell Specification Pathway Is Repressed by PHB, PHV and CNA in Arabidopsis. *PLoS One* **10**, e0126006 (2015).
71. Bürglin, T. R. Homeodomain subtypes and functional diversity. *Subcell. Biochem.* **52**, 95–122 (2011).
72. Smith, Z. R. & Long, J. A. Control of Arabidopsis apical-basal embryo polarity by antagonistic transcription factors. *Nature* **464**, 423–426 (2010).
73. Miyakawa, T., Fujita, Y., Yamaguchi-Shinozaki, K. & Tanokura, M. Structure and function of abscisic acid receptors. *Trends Plant Sci.* **18**, 259–266 (2013).
74. Ma, Y. et al. Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* **324**, 1064–1068 (2009).
75. Alpy, F. et al. Functional characterization of the MENTAL domain. *J. Biol. Chem.* **280**, 17945–52 (2005).
76. Prashek, J. et al. Interaction between the PH and START domains of ceramide transfer protein competes with phosphatidylinositol 4-phosphate binding by the PH domain. *J. Biol. Chem.* **292**, 14217–14228 (2017).
77. Tugaeva, K. V. et al. Molecular basis for the recognition of steroidogenic acute regulatory protein by the 14-3-3 protein family. *FEBS J.* **287**, 3944–3966 (2020).
78. Carrat, G. R. et al. The type 2 diabetes gene product STARD10 is a phosphoinositide-binding protein that controls insulin secretory granule biogenesis. *Mol. Metab.* **40**, 101015 (2020).
79. Wong, L. H. & Levine, T. P. Lipid transfer proteins do their thing anchored at membrane contact sites but what is their thing? *Biochem Soc. Trans.* **44**, 517–527 (2016).
80. Schrick, K. et al. Shared functions of plant and mammalian StAR-related lipid transfer (START) domains in modulating transcription factor activity. *BMC Biol.* **12**, 70 (2014).
81. Ponting, C. P. & Aravind, L. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends Biochem Sci.* **24**, 130–132 (1999).
82. Simonini, S., Bencivenga, S., Trick, M. & Østergaard, L. Auxin-Induced Modulation of ETTIN Activity Orchestrates Gene Expression in Arabidopsis. *Plant cell* **29**, 1864–1882 (2017).
83. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
84. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
85. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
86. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
87. Fu, L. Y. et al. ChIP-Hub provides an integrative platform for exploring plant regulome. *Nat. Commun.* **13**, 3413 (2022).
88. Stark, R. & Brown, G. DiffBind: differential binding analysis of ChIP-Seq peak data. Available at: <http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf> (2011).
89. Carroll, T. & Barrows, D. profileplyr: Visualization and annotation of read signal over genomic ranges with profileplyr. R package. <https://doi.org/10.18129/B9.bioc.profileplyr> (2023).
90. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4, <https://github.com/tidyverse/dplyr> (2023).
91. Wickham, H. Reshaping Data with the reshape Package. *J. Stat. Softw.* **21**, 1–20 (2007).
92. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1, <https://github.com/tidyverse/stringr> (2023).
93. Lawrence, M., Gentleman, R. & Carey, V. tracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

94. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org> (2016).
95. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
96. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
97. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
98. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
99. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
100. Chiu, T. P. et al. DNashapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211–1213 (2016).
101. Bonferroni, C. E. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*. 1935 Rome: Italy. 13–60. <https://www.semanticscholar.org/paper/Il-calcolo-delle-assicurazioni-su-gruppi-di-teste-Bonferroni-Bonferroni/98da9d46e4c442945bfd88db72be177e7a198fd3>.
102. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
103. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
104. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
105. Stephens, S. False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).
106. Dasgupta, A. et al. A phosphorylation-deficient ribosomal protein eS6 is largely functional in *Arabidopsis thaliana*, rescuing mutant defects from global translation and gene expression to photosynthesis and growth. *Plant Direct* **8**, e566 (2024).
107. Alexa, A. & Rahnenführer, J. topGO: Enrichment analysis for gene ontology. R package version 2.26.0 (2016).
108. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
109. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
110. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
111. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
112. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
113. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
114. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282 (1992).
115. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).
116. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

Acknowledgements

We are grateful to Dr. Ken Zaret, Dr. Doris Wagner, and Nicole Callery for thoughtful comments that have improved the manuscript. Work in the Husbands Lab is funded by grants from the National Science Foundation: #2039489 and #2310356 to A.Y.H.

Author contributions

Conceptualization: A.S.H. and A.Y.H.; performing experiments and analyzing results: A.S.H., S.G.C., E.P.A., C.E.D., R.U.C.; writing manuscript: A.S.H. and A.Y.H.; editing manuscript: A.S.H., S.G.C., C.E.D., R.U.C., and A.Y.H.; funding acquisition: I.B.Z. and A.Y.H.; project supervision: I.B.Z. and A.Y.H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-54269-z>.

Correspondence and requests for materials should be addressed to Aman Y. Husbands.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024