ARTICLES
https://doi.org/10.1038/s41477-021-00932-y

nature plants

Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters

Tobias Jores[1], Jackson Tonnies[1,2], Travis Wrightsman[3], Edward S. Buckler[3,4,5], Josh T. Cuperus[1]✉, Stanley Fields[1,6]✉ and Christine Queitsch[1]✉

Targeted engineering of plant gene expression holds great promise for ensuring food security and for producing biopharmaceuticals in plants. However, this engineering requires thorough knowledge of *cis*-regulatory elements to precisely control either endogenous or introduced genes. To generate this knowledge, we used a massively parallel reporter assay to measure the activity of nearly complete sets of promoters from *Arabidopsis*, maize and sorghum. We demonstrate that core promoter elements—notably the TATA box—as well as promoter GC content and promoter-proximal transcription factor binding sites influence promoter strength. By performing the experiments in two assay systems, leaves of the dicot tobacco and protoplasts of the monocot maize, we detect species-specific differences in the contributions of GC content and transcription factors to promoter strength. Using these observations, we built computational models to predict promoter strength in both assay systems, allowing us to design highly active promoters comparable in activity to the viral 35S minimal promoter. Our results establish a promising experimental approach to optimize native promoter elements and generate synthetic ones with desirable features.

Precise control of gene expression is necessary to generate transgenic plants with new properties, such as growth in formerly incompatible environments or production of medically or nutritionally important products[1,2]. Much of this control occurs at the initiation of transcription, the first committed step in gene expression. Transcription initiation involves the recruitment of the basal transcription machinery, comprised of general transcription factors (TFs) and RNA polymerase, to core promoters. Core promoters define the transcription start site (TSS) but their activity typically leads to only low levels of expression[3,4]. This basal level of transcription is increased by the interaction of core promoters with enhancers, which can reside upstream or downstream of the TSS and over a wide range of distances from the promoter[5–7].

The first core promoter element identified was the TATA box. This motif, with the consensus sequence TATA(A/T)A(A/T), is recognized by the TATA-binding protein, a subunit of TFIID, and plays an important role in recruiting the basal transcription machinery and in determining the TSS location[3,8,9]. Since then, several other core promoter elements have been discovered in viral and animal promoters[8,10–17]. In plants, short motifs composed of pyrimidine bases, termed the TC motif or Y patch, have been described as potential plant-specific core promoter elements[18–20].

Apart from these elements, promoters also contain binding sites for TFs close to the TSS. In contrast to the core promoter elements, which often occur at specific distances from, and in a fixed orientation to, the TSS, the TF-binding sites can be functional in either orientation and their activity is less constrained by their distance to the TSS. Promoter-proximal TF-binding sites can influence the transcriptional output from the nearby TSS and, in some cases, influence where transcription starts[21]. In this study, we refer to the region surrounding the TSS that harbours core promoter elements as the core promoter; the extended region that includes the core promoter and upstream TF-binding sites is referred to as the promoter.

To gain a better understanding of the regulatory principles governing promoter activity, several high-throughput studies have been performed in yeast, *Drosophila melanogaster* and human cells[22–29]. These studies validated the contribution of core promoter elements and promoter-proximal TF-binding sites to overall promoter activity and deduced rules governing the interaction among those elements. However, it is not clear whether these rules also apply to plant promoters. Although computational analyses have revealed that many of the core promoter elements identified in animals are enriched in plant promoters[18,19,30,31], only the TATA box and the Initiator (Inr) element have been functionally validated[32–35]. Some plant promoters do not harbour any of the known core promoter elements[30]. A recent study built synthetic plant promoters by combining TF-binding sites[36]. However, to date, large-scale functional studies have not been performed with plant core promoters.

A deeper understanding of the regulatory code of plant promoters and how it shapes transcription levels will further our knowledge of gene regulation, empower the controlled manipulation of gene expression for crop improvement and enable the rational design of promoters for use in genetic engineering. Here, we set out to comprehensively analyse the core promoters of the model plant *Arabidopsis thaliana* and the important crop maize (*Zea mays*) and its close relative sorghum (*Sorghum bicolor*). The genome of the crucifer *Arabidopsis* is compact (~135 megabases (Mb)) and AT-rich, while the genomes of the cereals maize and sorghum are GC-rich and many times larger (~2.7 gigabases and ~730 Mb, respectively). We sought to determine how these differences in genome content and architecture would be reflected in features of their promoter elements. Here, we identified key determinants of core promoter strength and characterized similarities and differences in the regulatory code of monocotyledonous and dicotyledonous plants. Using this knowledge, we designed synthetic core promoters with activities reaching levels comparable to that of the

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Graduate Program in Biology, University of Washington, Seattle, WA, USA. [3]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA. [4]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA. [5]Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA. [6]Department of Medicine, University of Washington, Seattle, WA, USA. ✉e-mail: cuperusj@uw.edu; fields@uw.edu; queitsch@uw.edu

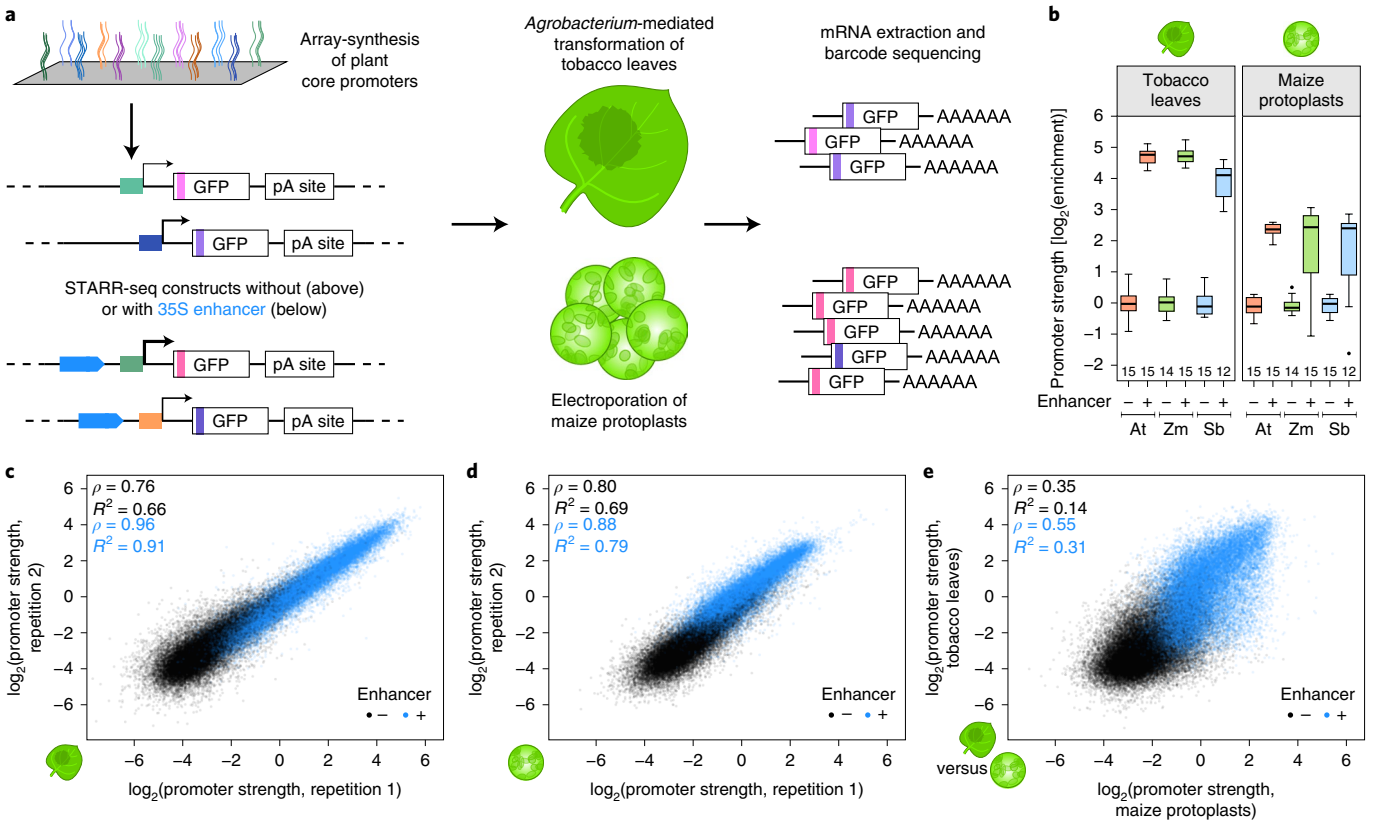


Fig. 1 | STARR-seq measures core promoter strength in tobacco leaves and maize protoplasts. a, Assay scheme. The core promoters (bases −165 to +5 relative to the TSS) of all genes of *Arabidopsis*, maize and sorghum were array-synthesized and cloned into STARR-seq constructs to drive the expression of a barcoded GFP reporter gene. For each species, two libraries, one without and one with a 35S enhancer upstream of the promoter, were created. The libraries were subjected to STARR-seq in transiently transformed tobacco leaves and maize protoplasts. **b**, Each promoter library (At, *Arabidopsis*; Zm, maize; Sb, sorghum) contained two internal control constructs driven by the 35S minimal promoter without (−) or with (+) an upstream 35S enhancer. The enrichment ($\log_2$) of recovered mRNA barcodes compared to DNA input was calculated with the enrichment of the enhancer-less control set to 0. In all following figures, this metric is indicated as promoter strength. Each boxplot (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) represents the enrichment of all barcodes linked to the corresponding internal control construct. The number of barcodes is indicated at the bottom of the plot. **c**,**d**, Correlation (Pearson's R^2 and Spearman's ρ) of two biological replicates of STARR-seq using the maize promoter libraries in tobacco leaves (**c**) or in maize protoplasts (**d**). **e**, Comparison of the strength of maize promoters in tobacco leaves and maize protoplasts. Pearson's R^2 and Spearman's ρ are indicated.

35S minimal promoter. Furthermore, we trained computational models that accurately predict promoter strength in our assays and help improve promoter activity.

Results

Use of the STARR-seq assay to study plant core promoters. We used the self-transcribing active regulatory region sequencing (STARR-seq) assay, which we had established in plants[35], to measure the strength of nearly complete sets of core promoters from *Arabidopsis*, maize and sorghum. Specifically, for each species, we interrogated the sequences from −165 to +5 relative to the annotated TSS for protein-coding and microRNA (miRNA) genes. These 170-bp regions were tested for promoter strength by using them to drive expression of a barcoded green fluorescent protein (GFP) reporter gene (Fig. 1a). We included the first five bases after the TSS to cover core promoter elements that span the TSS, like the Inr, while avoiding substantial parts of the 5′ untranslated region (UTR). The 5′ UTRs affect messenger RNA levels posttranscriptionally and hence their inclusion could confound assessment of promoter strength[37]. Instead, we used the 5′ UTR of a sorghum histone H3 gene (SORBI_3010G047100) for all sorghum promoters and the 5′ UTR of a maize histone H3.2 gene (Zm00001d041672) for all maize and *Arabidopsis* promoters (the 5′ UTR of the *Arabidopsis* histone H3.1 gene AT5G10390 had intrinsic promoter activity). We constructed three STARR-seq libraries that contained 18,329 *Arabidopsis*, 34,415 maize and 27,094 sorghum core promoters linked to ~400,000 unique barcodes per library (Supplementary Table 1). To test these promoters for their response to a strong enhancer, we also generated each library using a plasmid containing the cauliflower mosaic virus 35S enhancer[38,39] immediately upstream of the promoter insertion site[35]. The six libraries were assayed individually in transiently transformed tobacco leaves and maize protoplasts.

In each promoter library, we included two control constructs, one containing only the viral 35S minimal promoter (−46 to +5 relative to the TSS) and the other containing the 35S minimal promoter and enhancer (−199 to −47 relative to the TSS). The promoter strength for each tested plant promoter was normalized to the control construct containing only the 35S minimal promoter. The construct also containing the strong 35S enhancer upstream of the minimal promoter was used to test the dynamic range of the assay. Consistent with previous reports[35,40], the 35S enhancer was fourfold more active in the tobacco system than in maize protoplasts (Fig. 1b). We performed two biological replicates for each promoter library in each assay system. The replicates were highly correlated, especially for the libraries with the 35S enhancer, which reflected their generally higher promoter

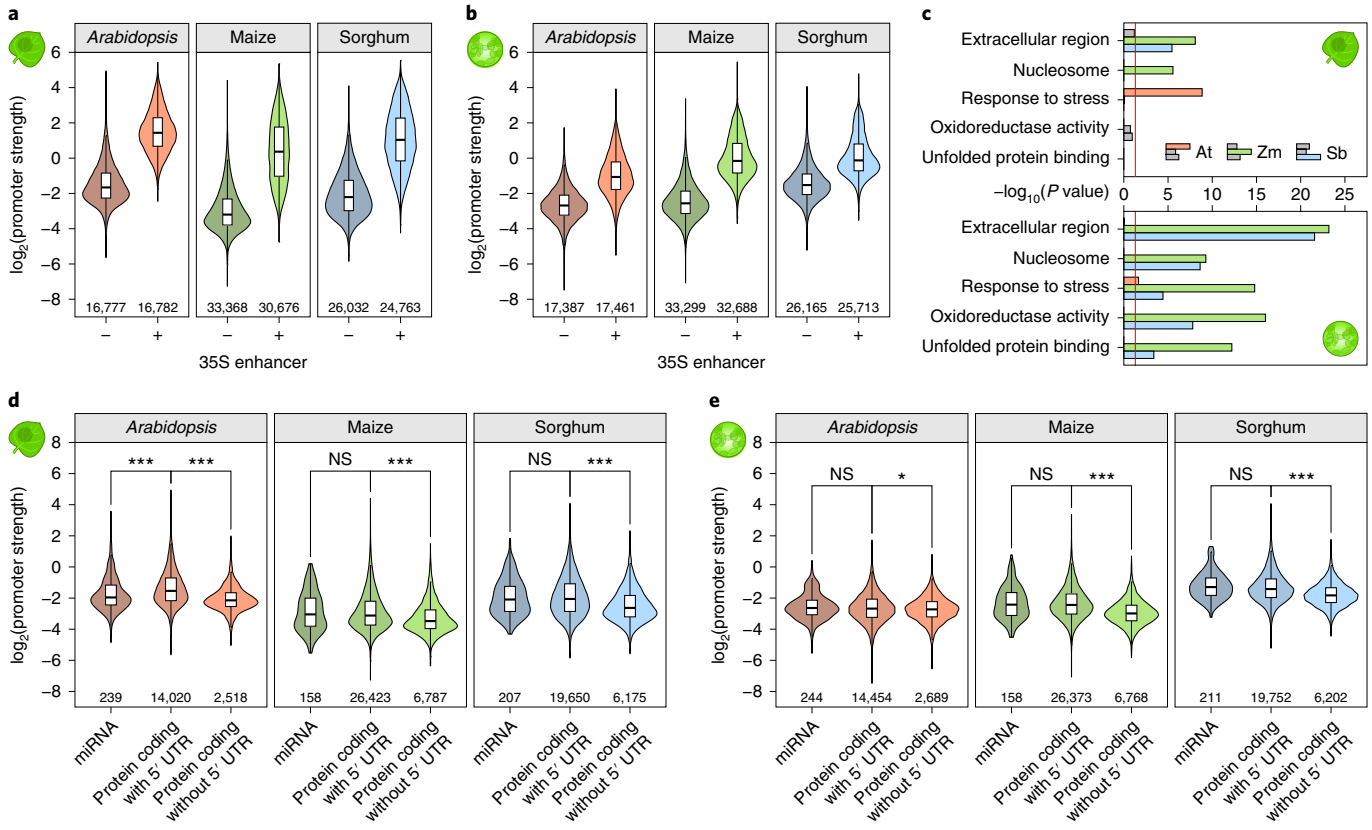


Fig. 2 | Plant core promoters span a wide range of activity. a,**b**, Violin plots of the strength of plant promoters from the indicated species as measured by STARR-seq in tobacco leaves (**a**) or maize protoplasts (**b**) for libraries without (−) or with (+) the 35S enhancer upstream of the promoter. **c**, Enrichment of selected GO terms for genes associated with the 1,000 strongest promoters in the *Arabidopsis* (At), maize (Zm) and sorghum (Sb) promoter libraries without enhancer in tobacco leaves (top panel) and maize protoplasts (bottom panel). The red line marks the significance threshold (adjusted $P \leq 0.05$). Non-significant bars are grey. The *P* values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. Exact *P* values are listed in Supplementary Table 11. **d**,**e**, Violin plots of promoter strength (libraries without 35S enhancer) in tobacco leaves (**d**) or maize protoplasts (**e**). Promoters were grouped by gene type. In **a**,**b**,**d** and **e**, violin plots represent the kernel density distribution and the boxplots within represent the median (centre line), upper and lower quartiles (box limits) and 1.5× the interquartile range (whiskers) for all corresponding promoters. Numbers at the bottom of the plot indicate the number of tested promoters. Significant differences between two samples were determined using the two-sided Wilcoxon rank-sum test and are indicated: $^{*}P \leq 0.01$; $^{**}P \leq 0.001$; $^{***}P \leq 0.0001$; NS, not significant. Exact *P* values are listed in Supplementary Table 11.

strength (Fig. 1c,d and Supplementary Fig. 1). Therefore, we used the average promoter strength from both replicates for all further analyses. We validated these results by retesting a subset of 166 and 173 promoters in two separate libraries, obtaining results that were highly correlated with the data from the comprehensive promoter libraries (Supplementary Fig. 2). Since the sorghum promoters were coupled to a sorghum 5′ UTR in the comprehensive library and to a maize 5′ UTR in the validation libraries, the high correlation between these datasets suggests that the two 5′ UTRs did not strongly affect promoter strength.

Promoter strengths as measured in the tobacco leaf system had a weak to intermediate (R^2 of 0.14–0.40) correlation with those obtained from maize protoplasts (Fig. 1e and Supplementary Fig. 1c,f), indicating that there are substantial differences in how the two systems interact with the core promoters. Irrespective of the assay system, the promoters spanned a wide range of activity, with >250-fold difference between the strongest and weakest promoters (Fig. 2a,b and Supplementary Table 2). Few promoters were stronger than the viral 35S minimal promoter, which is probably optimized for maximal activity. Overall, the promoters of the dicot *Arabidopsis* tended to perform better in the dicot tobacco system, while the promoters of the monocots maize and sorghum showed greater activity in protoplasts of the monocot maize (Fig. 2a,b).

Gene ontology (GO)-term enrichment analysis showed that the genes corresponding to the most active promoters in our assay were significantly (adjusted $P \leq 0.05$) enriched for components of nucleosomes, which are highly expressed housekeeping genes (Fig. 2c). In both systems, strong promoters often were also associated with genes annotated for response to stress and function in the extracellular region, including genes encoding defence and cell wall proteins. In the maize protoplast system, genes associated with strong promoters frequently encoded proteins with oxidoreductase activity or unfolded protein-binding functions. The latter is consistent with reports of wound-induced reactive oxygen species and a heatshock response in protoplasts[41]. Although these results show a qualitative agreement between core promoter strength and expression level for some genes, there was no substantial correlation overall between promoter strength and expression data[42–44] for the corresponding genes in planta (Extended Data Fig. 1). This lack of correlation is expected, as core promoters represent only a subset of all the regulatory elements that drive gene expression and other elements such as enhancers can drastically affect transcription rates in the genomic context.

Next, we asked if genes of different types use different promoters. The activity of miRNA promoters was indistinguishable from that of promoters of protein-coding genes (Fig. 2e,d). However, promoters

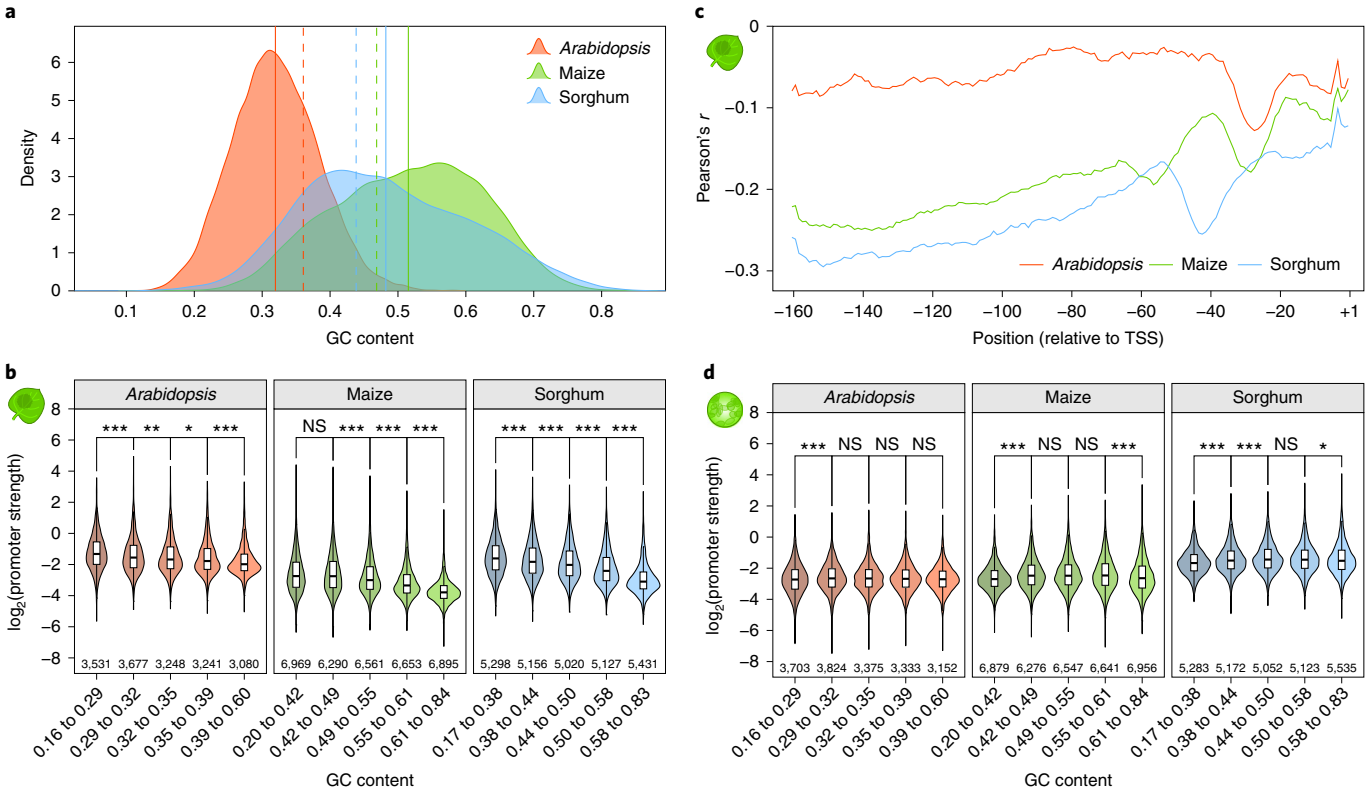


Fig. 3 | GC content affects promoter strength in tobacco leaves. a, Distribution of GC content for all promoters of the indicated species. Lines denote the mean GC content of promoters (solid line) and the whole genome (dashed line). **b**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves. Promoters are grouped by GC content to yield groups of approximately similar size. **c**, Correlation (Pearson's *r*) between promoter strength and the GC content of a ten-base window around the indicated position in the plant promoters. **d**, As **b** but for promoter strength in maize protoplasts.

from genes with an annotated 5′ UTR were generally stronger than those of genes without a 5′ UTR annotation. As the TSSs of the latter are probably not correctly annotated, these sequences are probably not true promoters, explaining their low activity.

Multiple sequence features influence promoter strength. Monocot genomes are more GC-rich than dicot genomes[30,45] and this bias holds true for their core promoter sequences (Fig. 3a). In the tobacco leaf system, GC content strongly affected promoter strength, with AT-rich promoters up to fourfold more active than GC-rich ones (Fig. 3b). A high GC content was especially detrimental close to the 5′ end of the promoters but was better tolerated towards the 3′ end (Fig. 3c). In contrast, in maize protoplasts, GC content was not predictive of promoter strength (Fig. 3d). Since the GC content of the *Arabidopsis* and tobacco genomes is similar[46], the transcriptional machinery in tobacco is probably tuned to AT-rich promoters and works less well with the GC-rich promoters of maize and sorghum. Conversely, the transcription machinery of maize commonly acts on GC-rich promoters and can effectively use them in protoplasts. The correlation between promoter strength and GC content is, therefore, a characteristic of the assay system and not an intrinsic feature of the promoters.

We next tested how known core promoter elements affect promoter strength. Considering first the location of TATA box motifs, we noticed marked differences among the promoters of *Arabidopsis*, maize and sorghum. In *Arabidopsis* promoters, the distribution of TATA boxes had a peak ~30 bp upstream of the TSS (Fig. 4a). Although this location also is common for maize promoters, the maize promoters showed two additional peaks for the TATA box at: ~55 and ~70 bp upstream of the TSS. In sorghum promoters, the TATA box distribution peaked at ~40 bp upstream of the TSS, with a shoulder ~30 bp upstream of the TSS.

Core promoters harbouring a TATA box were up to fourfold stronger than TATA-less ones, especially when the TATA box is located within the region from 23 to 59 bp upstream of the TSS, where most TATA boxes in the promoters of *Arabidopsis*, maize and sorghum reside (Fig. 4a–c). The location of the TATA box in maize promoters affected their strength only in maize protoplasts. In this assay system, maize promoters with a TATA box in one of the three peaks of the TATA box distribution were stronger than those with a TATA box elsewhere. Furthermore, maize promoters with a TATA box in the peak closest to the TSS were strongest and they became successively weaker in the other two peaks as the TATA box is located increasingly more TSS-distal (Extended Data Fig. 2). The effect of the TATA box on promoter strength was not a consequence of an increased AT-content in the promoters containing a TATA box. (Supplementary Fig. 3). To directly measure the effect of the TATA box, we mutated this motif in native promoters. Replacement of one or both T nucleotides in the core TATA motif with a G resulted in decreased transcriptional activity (Fig. 4d,e). Similarly, promoter strength was increased when a canonical TATA box was inserted into a TATA-less promoter; a mutated version of the TATA box did not have this effect (Fig. 4f,g).

In animal promoters, the TATA box is often surrounded by the upstream (BRE^u) and/or downstream (BRE^d) TFIIB recognition element. Mutational studies have demonstrated that these elements can modulate promoter strength[13,16]. In tobacco leaves, neither of the two elements had a strong effect on promoter activity; however, in maize protoplasts, BRE^u was associated with 25% increased, and BRE^d with 10% decreased, promoter strength (Extended Data Fig. 3a–d

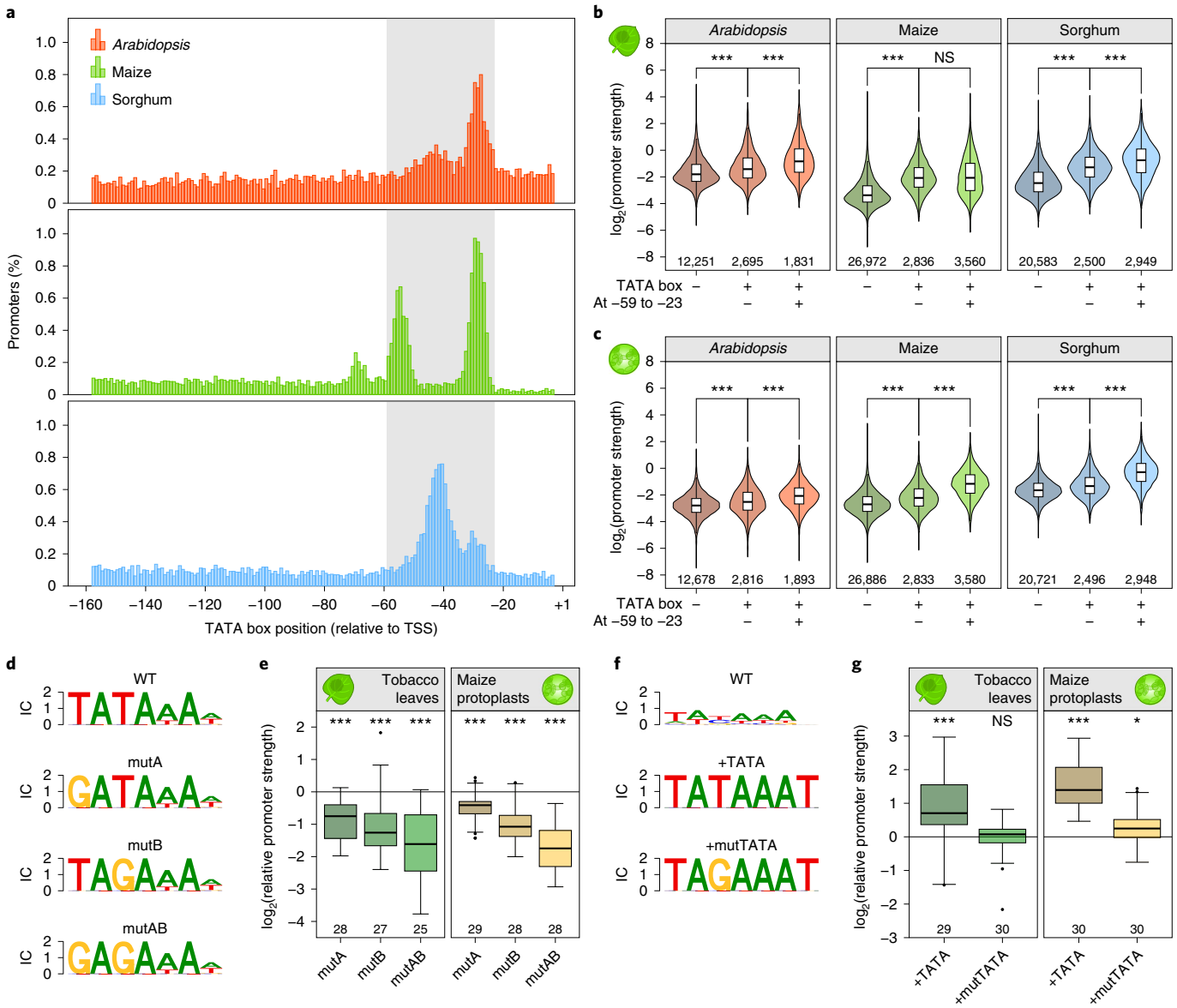


Fig. 4 | The TATA box is a key determinant of promoter strength. a, Histograms showing the percentage of promoters with a TATA box at the indicated position. The region between positions −59 and −23 in which most TATA boxes reside is highlighted in grey. **b**,**c**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves (**b**) or maize protoplasts (**c**). Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within (+/+) the −59 to −23 region. **d**–**g**, Thirty plant promoters with a strong (**d**,**e**) or weak (**f**,**g**) TATA box (wild type, WT) were tested. One (mutA and mutB) or two (mutAB) T > G mutations were inserted into promoters with a strong TATA box (**d**,**e**). A canonical TATA box (+TATA) or one with a T > G mutation (+mutTATA) was used to replace the weak TATA box (**f**,**g**). Logoplots (**f**,**d**) of the TATA box regions of these promoters and their strength (**g**,**e**) relative to the WT promoter (set to 0, horizontal black line) are shown. Boxplots (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) denote the strength of the indicated promoter variants. Numbers at the bottom of the plot indicate the number of tested promoter elements. Significant differences from a null distribution were determined using the two-sided Wilcoxon signed rank test and are indicated: $*P \leq 0.01$; $**P \leq 0.001$; $***P \leq 0.0001$; NS, not significant. Exact *P* values are listed in Supplementary Table 11. IC, information content.

and Supplementary Table 4). Consistent with these results, mutations that inactivate BRE^{u} decreased promoter strength in maize protoplasts but not in tobacco leaves. Inserting a canonical BRE^{u} led to increased promoter activity, especially in maize protoplasts. In contrast, mutating or inserting BRE^{d} had only modest effects on promoter activity in both assay systems (Extended Data Fig. 3e–h). A valine residue in the helix-turn-helix motif of the general transcription factor TFIIB is crucial for the recognition of BRE^{u} in animals[13,47]. Although this residue is not conserved in any plant TFIIB protein, the maize genome encodes an additional TFIIB-related protein with a valine at the corresponding position (Supplementary Fig. 4). The presence of this maize-specific TFIIB-related protein may explain the increased activity of BRE^{u} in the maize protoplast system.

Computational analyses of plant promoters[18–20] have detected an enrichment of short, pyrimidine-rich motifs upstream of the TSS (Extended Data Fig. 4a). Because such an enrichment was not detected in animal promoters, these motifs, termed Y patches, were

proposed to be plant-specific core promoter elements. Our data support this hypothesis, as Y patch-containing promoters showed 10–15% greater strength compared to those without the element (Extended Data Fig. 4b,c and Supplementary Table 4).

Consistent with previous studies[32,34], we observed that promoters with an Inr at the TSS were generally stronger than those without it. In contrast, the polypyrimidine initiator TCT, previously described in animals[17], was less effective (Extended Data Fig. 5).

Finally, we asked whether promoter-proximal TF-binding sites affect promoter strength. We first clustered TFs by similarity of their binding site motifs and created a consensus motif for each of the 72 clusters (Supplementary Table 3). We then compared the strength of promoters with a predicted binding site to that of promoters lacking it. About 67% of the TF clusters did not have a significant impact on promoter strength. However, 23 TF motifs were significantly ($P \leq 0.0005$) associated with altered promoter strength in at least one assay system (Supplementary Table 4). For example, the TCP TF motif tends to reside in promoters that were strong in tobacco leaves, while this effect was not observed in maize protoplasts (Extended Data Fig. 6a,b). On the other hand, promoters with a motif for heatshock factors (HSFs) were stronger than those without it in maize protoplasts but not in tobacco leaves (Extended Data Fig. 6c,d).

We asked whether core promoter elements and TF-binding sites are spatially constrained in relation to one another. In contrast to core promoter elements, most TF-binding sites did not show a preferential position relative to the TSS. However, we observed that TF-binding sites upstream of the TATA box were generally associated with a higher promoter strength compared to those downstream of the TATA box (Extended Data Fig. 7 and Supplementary Table 5). Since RNA polymerase is recruited to the region downstream of the TATA box, this enzyme may displace TFs bound here and thereby prevent them from activating transcription.

Promoters show varying degrees of enhancer responsiveness. In animals, promoters can interact differentially with enhancers[25,48]. Similarly, the 35S enhancer activated some plant core promoters more than others. However, the presence of the 35S enhancer resulted in increased transcription from almost all core promoters, up to 60-fold for the most responsive promoters in the tobacco leaf system and up to 15-fold in maize protoplasts; the 35S enhancer is less active in maize protoplasts[35,40]. Consistent with the notion that enhancers are the drivers of tissue- and condition-specific transcription[4,39], promoters of genes with high tissue specificity (top third of the genes as ranked by the tissue-specificity index τ; ref. [49]) showed on average 33% increased enhancer responsiveness compared to promoters of genes with low tissue specificity (bottom third of the τ distribution) (Fig. 5a,b). Similarly, promoters of miRNA genes, which are often differentially expressed in response to environmental or developmental cues, were 33% more responsive to the 35S enhancer than promoters of protein-coding genes (Supplementary Fig. 5).

To understand which promoter features influence enhancer responsiveness, we analysed the elements that affect promoter strength. Promoters with a TATA box were up to 67% more responsive to the 35S enhancer than TATA-less promoters; however, the location of the TATA box did not have a consistent impact on enhancer responsiveness (Fig. 5c,d). Furthermore, promoter GC content influenced enhancer responsiveness in the tobacco leaf system but not in maize protoplasts (Fig. 5e,f). While the GC content and TATA box had a similar effect on enhancer responsiveness as on promoter strength, the same was not true for TFs. Instead, TFs that increased promoter strength often reduced enhancer responsiveness (Extended Data Fig. 8a–d), potentially due to competition for a limited pool of TFs or because of incompatibilities between recruited downstream factors. In contrast, some TFs that did not influence promoter strength affected enhancer responsiveness (Extended Data Fig. 8e,f). The effects on enhancer responsiveness possibly reflect synergistic effects, whereby the core transcriptional machinery and the TFs at promoters and enhancers interact with one another.

Core promoter strength can be modulated by light. The plant STARR-seq assay can identify light-responsive enhancers[35]. To test whether core promoters that respond to light can also be identified, we subjected the promoter libraries to STARR-seq experiments in tobacco leaves that were kept in the light (16 h light, 8 h dark) for 2 d after transformation (Fig. 6a). We did not perform the same experiment with maize protoplasts, as known light-responsive enhancers were not active in this system (Supplementary Fig. 6). As expected, most promoters did not respond to the light. However, about 2,400 promoters were at least four times more active in the light or in the dark (Fig. 6b). The genes associated with the most highly light-dependent promoters were enriched for those encoding plastid proteins, especially for proteins in thylakoids, the membrane-bound chloroplast compartments that are the site of the light-dependent reactions of photosynthesis (Fig. 6c).

While promoters that are AT-rich were more light-dependent than GC-rich ones (Fig. 6d), the effects of GC content on light-dependency were much less pronounced than on promoter strength and enhancer responsiveness. Similarly, the presence of a TATA box showed weaker and even inconsistent effects on light-dependency compared to TATA box effects on promoter strength and enhancer responsiveness (Fig. 6d). We found that the light-dependency of a promoter was mainly determined by the TF-binding sites it contains. The presence of the TCP-binding site, for example, led to increased expression in the light (Fig. 6e) and, consistent with previous studies[50], the presence of the WRKY-binding site led to repressed expression in the light (Fig. 6f). These trends were confirmed by mutational analysis. Mutations that disrupt a binding site for WRKY TFs increased the light-dependency of the promoter, while mutations that disrupt a binding site for TCP TFs led to a noticeable, albeit not significant, decrease in light-dependency (Extended Data Fig. 9).

Design of synthetic plant promoters. After identifying key features of native plant promoters, we sought to use these features in the design of synthetic promoters. We started by generating random sequences with nucleotide frequencies resembling either an average *Arabidopsis* or average maize promoter (Fig. 7a). We designed ten sequences each for the two nucleotide frequencies; however, due to their AT-rich nature, the synthesis of approximately half of the sequences with an *Arabidopsis* promoter-like base composition failed. Consistent with the findings for native promoters, the synthetic promoters with low GC content, similar to that of *Arabidopsis* promoters, were 30% more active in tobacco leaves than those with GC content similar to that of maize promoters (Fig. 7b,c). However, as expected, these random synthetic promoters were weak. To increase their activity, we modified them by adding an Inr, Y patch element or TATA box (Fig. 7a). Although all three of these core promoter elements, both alone and in combination, increased promoter strength, the TATA box showed the strongest effect and the Inr the weakest (Fig. 7b,c and Supplementary Table 6). The relative activity of these three elements was similar across synthetic promoters with initial nucleotide frequencies similar to either *Arabidopsis* or maize and across the two assay systems. However, in tobacco leaves, the absolute change in promoter strength was different for synthetic promoters of different GC content, indicating that the elements tested in this assay system require a favourable sequence environment to achieve full activity (Fig. 7b). Taken together, the results demonstrate that it is possible to rationally design synthetic core promoters of varying strength by choosing an appropriate

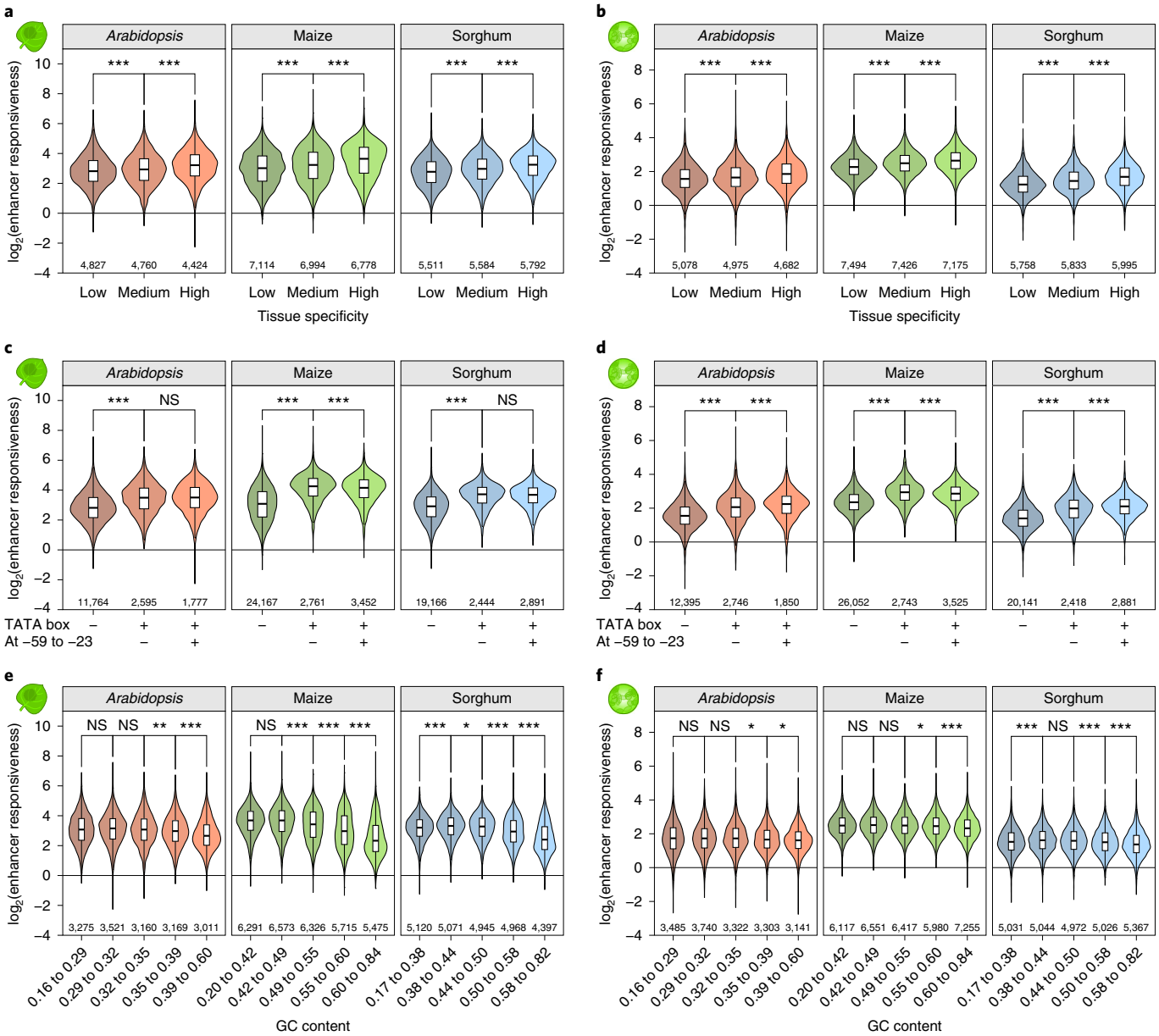


Fig. 5 | Enhancer responsiveness of promoters depends on the TATA box and GC content. a,**b**, Violin plots of enhancer responsiveness (promoter strength$^{\text{with enhancer}}$ divided by promoter strength$^{\text{without enhancer}}$) in tobacco leaves (**a**) or maize protoplasts (**b**). Promoters were grouped into three bins of approximately similar size according to the tissue-specificity τ of the expression of the associated gene. **c**,**d**, Violin plots of enhancer responsiveness in tobacco leaves (**c**) or maize protoplasts (**d**). Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within (+/+) the −59 to −23 region. **e**,**f**, Violin plots of enhancer responsiveness in tobacco leaves (**e**) or maize protoplasts (**f**) for promoters grouped by GC content. Violin plots, boxplots and significance levels in (**a**–**f**) are as defined in Fig. 2.

background nucleotide frequency and adding canonical core promoter elements. The strongest synthetic promoters reached activities comparable to the viral 35S minimal promoter.

We also used the synthetic promoters to further analyse the effect of promoter-proximal TF-binding sites. We focused on four different binding sites: two sites for TCP TFs and one each for HSF TFs and NAC TFs. The TF-binding sites were introduced at three positions in the synthetic promoters in which a TATA box had been added (Fig. 7d). Because we did not observe position-dependent differences for any of the three TF-binding sites, we grouped their respective data to perform the subsequent analyses. Consistent with our observations for native promoters, the TCP-binding sites had the strongest effect in tobacco leaves, the HSF sites were most active in maize protoplasts and the NAC sites had a weak but consistent effect across both assay systems (Fig. 7e). When more than one TF-binding site was introduced into the synthetic promoters, their activities were additive and the relative strengths of the promoters were conserved in combinations. The more binding sites that were present, the higher the promoter strength (Fig. 7f, Supplementary Fig. 7 and Supplementary Table 6).

Finally, to test whether the TFs show position-dependent activity with regard to the TATA box, the binding sites for TCP, HSF and NAC TFs were inserted at several positions upstream and downstream of the TATA box. While these TF-binding sites at all tested positions upstream of the TATA box led to similar increases in promoter strength, they did not increase promoter strength when

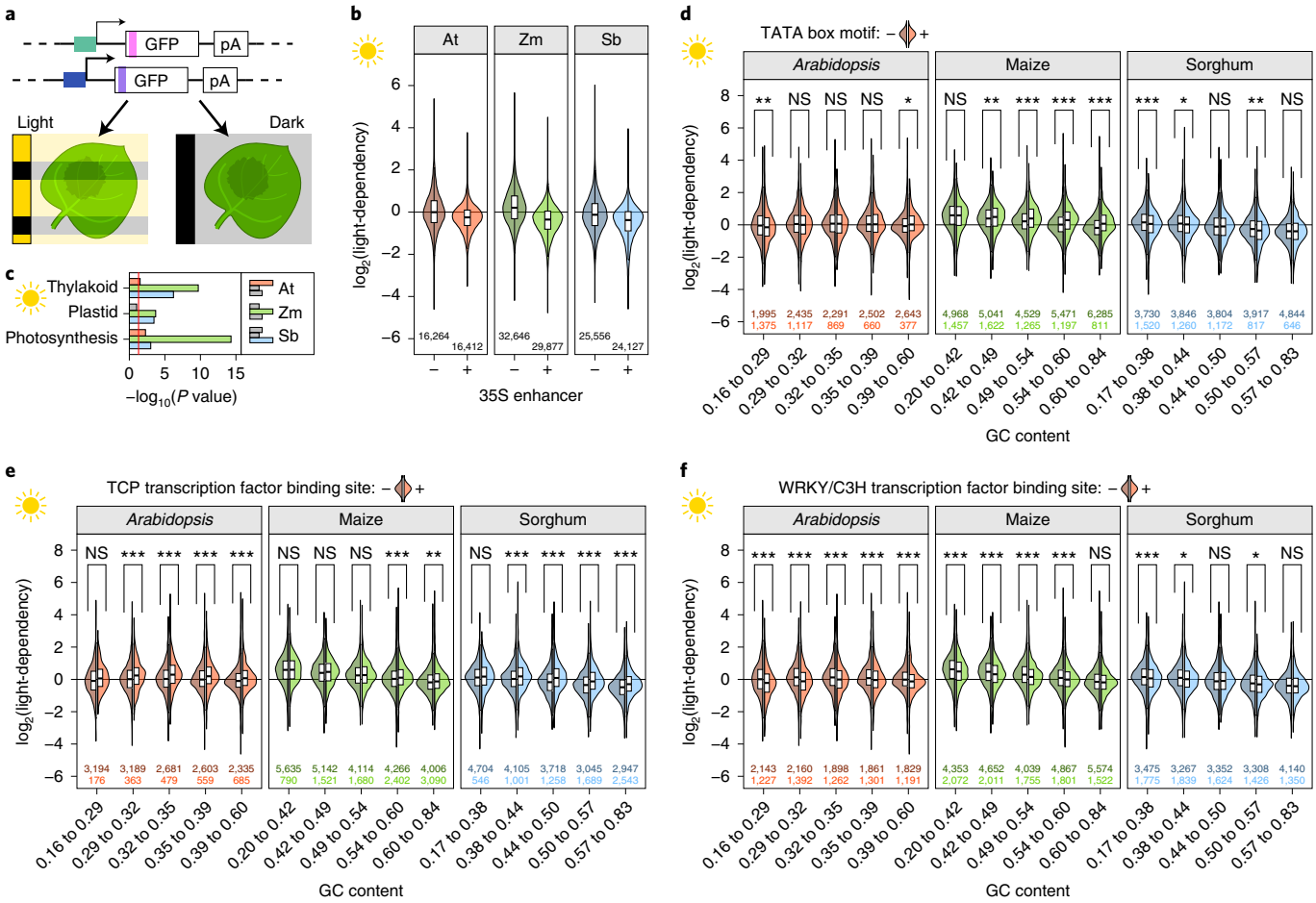


Fig. 6 | Promoter strength can be modulated by light. a, Tobacco leaves were transiently transformed with STARR-seq promoter libraries and the plants were kept for 2 d in 16 h light/8 h dark cycles (light) or completely in the dark (dark) before mRNA extraction. **b**, Violin plots of light-dependency (promoter strength^{light} divided by promoter strength^{dark}) for promoters in the libraries with (+) or without (−) the 35S enhancer. **c**, Enrichment of selected GO terms for genes associated with the 1,000 most light-dependent promoters. The red line marks the significance threshold (adjusted $P \leq 0.05$). Non-significant bars are grey. The *P* values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. Exact *P* values are listed in Supplementary Table 11. **d**–**f**, Violin plots of light-dependency. Promoters are grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a TATA box (**d**) or a binding site for TCP (**e**) or WRKY (**f**) TFs. Violin plots, boxplots and significance levels in **b** and **d**–**f** are as defined in Fig. 2. Only one half is shown for violin plots in **d**–**f**.

inserted downstream of the TATA box (Fig. 7g,h, Supplementary Fig. 8 and Supplementary Table 6). These results probably reflect competition with the core transcriptional machinery that binds to this region.

Computational models predict and improve promoter strength. Computational models have been used to optimize synthetic gene-regulatory sequences[29,51]. Therefore, we set out to develop predictive models for core promoter strength using the data from the libraries with the 35S enhancer to train the models, as they had a better replicate correlation. For each assay system, we trained a separate model using 90% of the promoters, with the remaining 10% used to validate the model. We initially used a linear regression model for this task. The GC content and the maximum score for a match to the position weight matrices for the core promoter elements and TF clusters of each sequence were used as input features. The linear models explained 51% and 45% of the variability in promoter strength in tobacco leaves and maize protoplasts, respectively (Fig. 8a). In both systems, the TATA box score was the most important feature for promoter strength, followed by GC content.

To obtain models with increased predictive power, we turned to a machine learning approach using a convolutional neural network (CNN). The models used the DNA sequence of the core promoters as input and predicted the strength of the promoters in the test set, resulting in an R^2 of 0.71 and 0.67 for the tobacco and the maize systems, respectively (Fig. 8b).

We used these models for in silico evolution of 150 native promoters with weak, intermediate or strong activity in our assay. Additionally, we subjected the synthetic promoters with or without various core promoter elements to evolution. For each promoter, we generated every possible single nucleotide substitution variant and scored these variants with the CNN models. The best variant was retained and subjected to another round of evolution. We synthesized the starting sequences and those obtained after three and ten rounds of evolution and experimentally determined their activity. As predicted, we observed a large increase in promoter strength after three rounds of evolution and another, albeit less pronounced, increase after ten rounds (Fig. 8c,d and Extended Data Fig. 10). We obtained the best results when the evolution was performed with the CNN model trained on data from the same assay system. However, when we used a combination of both models to score the promoter variants, we could generate promoters with high activities in both systems that were on par with those evolved with the CNN model that was trained

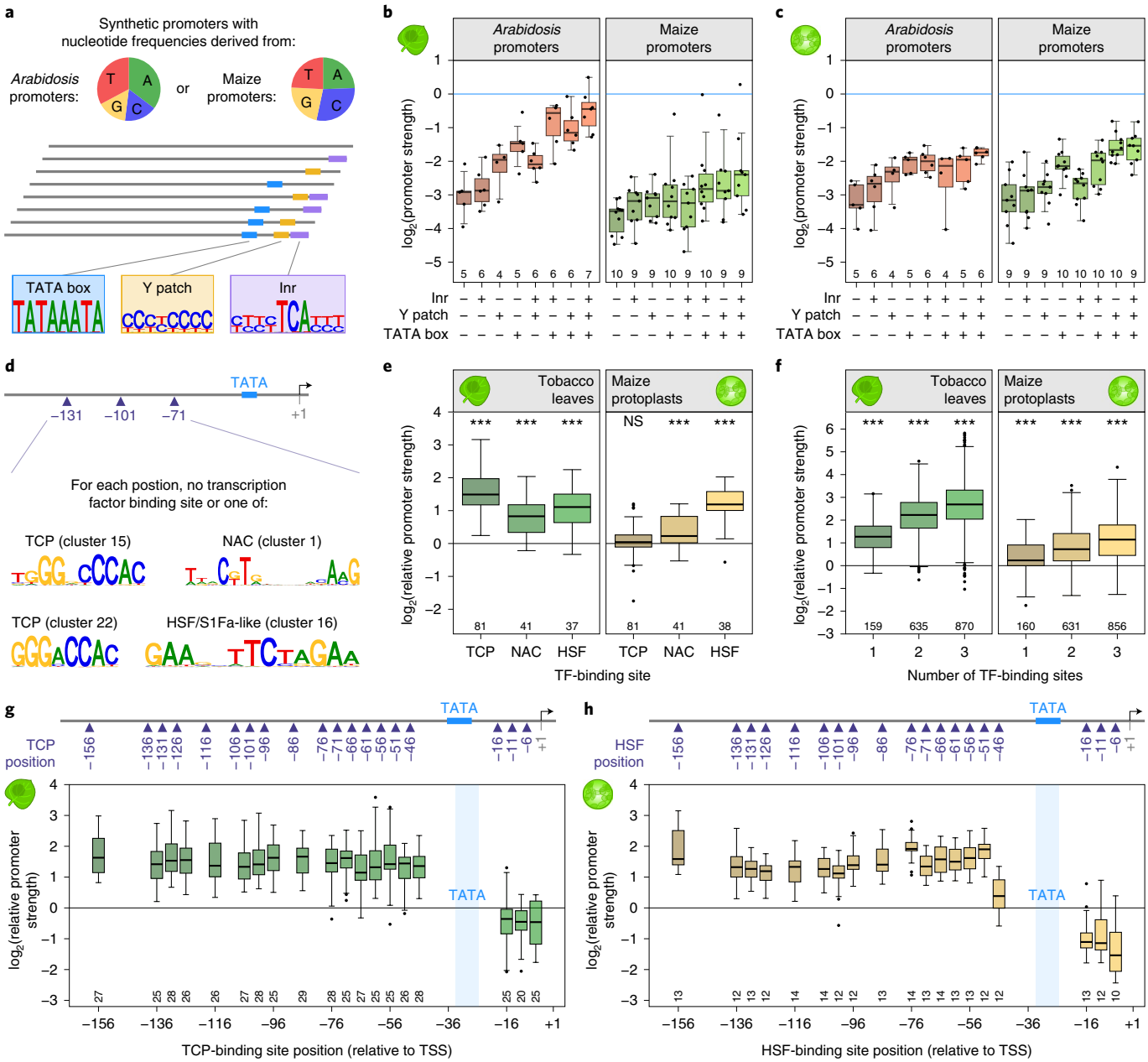


Fig. 7 | Design and validation of synthetic promoters. a–c, Synthetic promoters with nucleotide frequencies similar to an average *Arabidopsis* (35.2% A, 16.6% C, 15.3% G and 32.8% T) or maize (24.5% A, 29.0% C, 22.5% G and 23.9% T) promoter were created and modified by adding a TATA box, Y patch and/or Inr element (**a**); promoter strength was determined by STARR-seq in tobacco leaves (**b**) and maize protoplasts (**c**). Promoters with an *Arabidopsis*-like nucleotide composition are shown on the left, those with maize-like base frequencies on the right. The strength of the 35S minimal promoter is indicated by a horizontal blue line. Individual data points are shown. **d–f**, TF-binding sites for TCP, NAC and HSF transcription factors were inserted at positions 35, 65 and/or 95 of the synthetic promoters with a TATA box (**d**) and the activity of promoters with a single binding site for the indicated TF (**e**) or multiple binding sites (**f**) was determined in tobacco leaves (left panel) or maize protoplasts (right panel). **g**,**h**, A single TCP (**g**) or HSF (**h**) TF-binding site was inserted at the indicated position in the synthetic promoters containing a TATA box. The strength of these promoters was measured in tobacco leaves (**g**) or maize protoplasts (**h**). Boxplots and significance levels in **b**,**c** and **e**–**h** are as defined in Fig. 4. In **e**–**h**, the corresponding promoter without any TF-binding site was set to 0 (horizontal black line).

on data from the system in which the evolved sequences were tested (Fig. 8c–f and Supplementary Table 7). The models used for the in silico evolution were trained on data from libraries with an upstream 35S enhancer; however, when we tested the evolved promoters without the 35S enhancer, their activities followed the same trend, with a large increase in activity after three rounds and an additional increase after ten (Fig. 8e,f). These results suggest that the increased promoter strength generated by the evolution process was not enhancer-dependent and that these promoters might similarly work well with other enhancers.

Discussion

The use of plants to synthesize medical and nutritional products requires precise control of foreign genes; similarly, precise control of endogenous genes is required to generate plants that can better withstand stresses. This precision can be realized through the design of

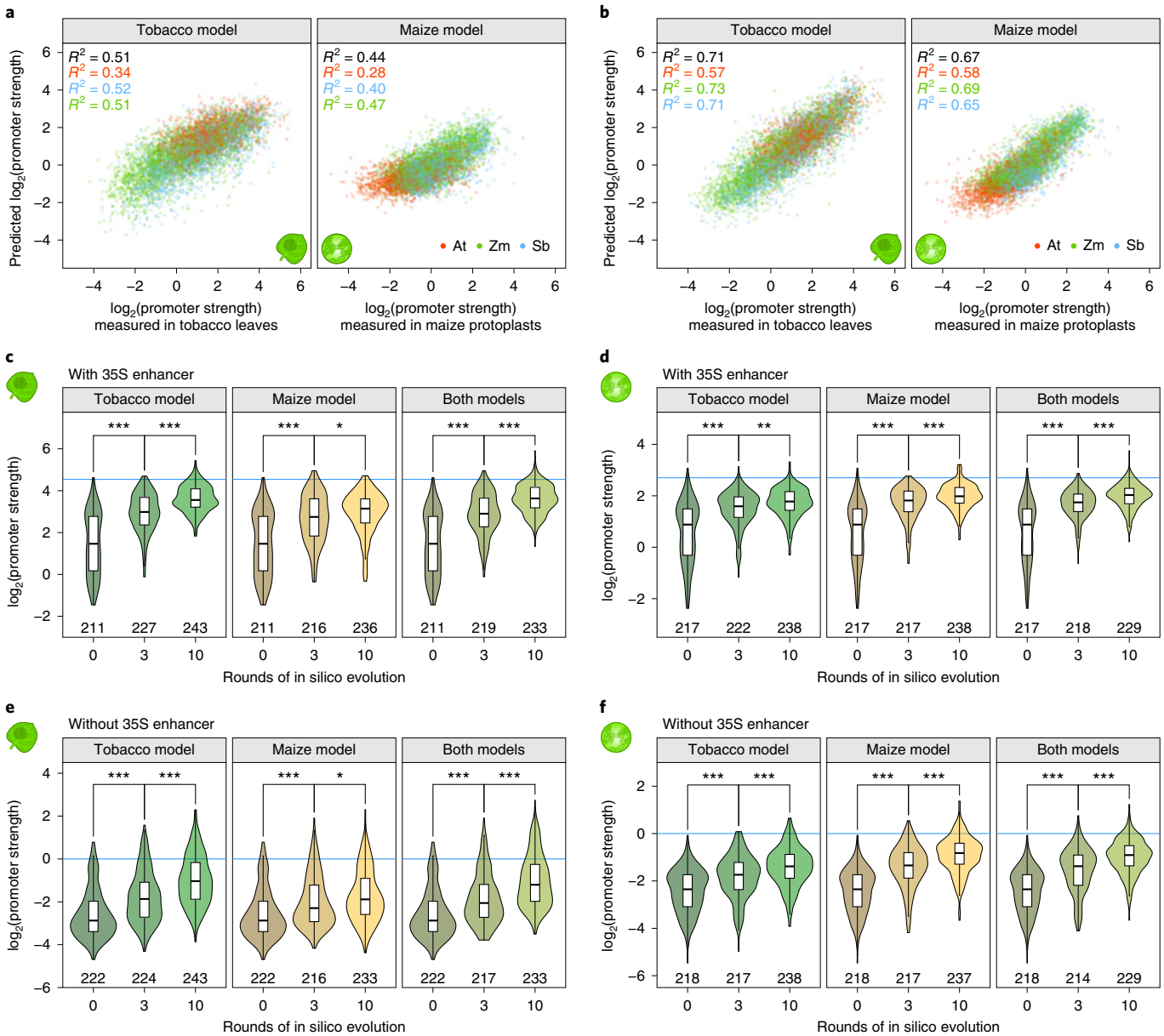


Fig. 8 | Computational models can predict promoter strength and enable in silico evolution of plant promoters. a, Correlation between the promoter strength as determined by STARR-seq using promoter libraries with the 35S enhancer and predictions from a linear model based on the GC content and motif scores for core promoter elements and TFs. The models were trained on data from the tobacco leaf system (tobacco model) or the maize protoplasts (maize model). The overall correlation is indicated in black and correlations for each species are coloured as indicated (inset). Correlations (Pearson's R^2) are shown for a test set of 10% of all promoters. **b**, Similar to **a** but the prediction is based on a CNN trained on promoter sequences. **c**–**f**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength of the unmodified promoters (0 rounds of evolution) or after they were subjected to three or ten rounds of in silico evolution as determined in tobacco leaves (**c**,**e**) or maize protoplasts (**d**,**f**). The promoters were tested in a library with (**c**,**d**) or without (**e**,**f**) an upstream 35S enhancer. The models used for the in silico evolution are indicated on each plot. The promoter strength of the 35S promoter is indicated by a horizontal blue line.

synthetic promoters with optimal sequences, spacings and orientations of regulatory elements. Here, we used the STARR-seq assay to characterize plant core promoters in depth. We demonstrate that the most critical element of a strong plant core promoter is the presence of a TATA box ~30–40 bp upstream of the TSS. The next most critical element is a nucleotide composition appropriate for the plant that is being engineered. A promoter can further be improved with an Inr motif at the TSS and a pyrimidine-rich region between the TATA box and the Inr. Such rationally designed promoters can reach activities comparable to the highly active viral 35S minimal promoter.

While it might be optimal to conduct these experiments within the genomic context in planta, current technologies make such large-scale studies feasible only with transient expression of reporter constructs. However, the lack of genomic context may be less important for promoter strength than is commonly assumed. Studies in human and *Drosophila* cells found that results from plasmid-based regulatory elements are highly correlated with those from genome-integrated ones in massively parallel reporter assays[52,53]. Moreover, human core promoters retain their relative strength regardless of where they are inserted in the genome or if

they drive expression of a plasmid-encoded reporter; the genomic context appears merely to scale their activity but does so independently of promoter identity[54]. Furthermore, we and others have previously demonstrated that transient STARR-seq assays in plants recapitulate the relative strength and the condition-specificity of known regulatory elements[5,35]. Our findings about the relative strength of promoters should, therefore, apply to promoters integrated in the genome, with the caveat that nearby enhancers may modulate the absolute expression level in addition to tissue- and condition-specificity.

Promoter activity and conditional response can be further modified by the addition of TF-binding sites upstream of the TATA box. Such binding sites affected promoter strength in an additive manner. The choice of binding site, however, will depend on the assay system and on the TFs that are present and active in it. TF presence and activity cannot simply be inferred from TF motifs because plant TF families are large and often encode both activating and repressing factors with highly similar binding preferences. However, single-cell genomics can determine which TFs are expressed in specific cell types and associated with chromatin accessibility of regulatory elements[55,56,57]. This knowledge offers a promising avenue to explore the activity of cell type-specific regulatory elements. In the absence of an assay system derived from a cognate cell type, cell type-specific TFs can be co-expressed in the assay systems used here. Alternatively, a large array of promoters can be designed with an assortment of TF-binding sites, followed by an assay like the one described here to identify the most active ones.

Nevertheless, the design of strong core promoters appears feasible without such cell type-specific or even species-specific data. Our CNN models accurately predicted promoter strength and could be used for in silico evolution to yield native and synthetic promoters with increased activity. Moreover, a combination of CNN models trained on data from the tobacco and maize assay systems yielded promoters active in both systems. Such promoters are robust candidates to use across a broad range of tissues and species and in conjunction with multiple enhancers.

In animals, enhancer–promoter interactions are fine-tuned to execute distinct regulatory programmes, like expression of housekeeping or developmental genes[25,48]. Here, we studied the effect of only the viral 35S enhancer on plant promoters. However, this assay could be applied to study interactions between promoters and native plant enhancers; such experiments might reveal specific interactions between distinct types of promoters and enhancers. Combining the potent core promoters characterized here with equally well-characterized enhancers will add the desired condition-specific and cell type-specific regulation needed for applications in plant engineering and biotechnology.

Methods

Library design and construction. For this study, we used the sequence from −165 to +5 relative to the annotated TSS as core promoters. We used the Araport11 annotation[58] for *A. thaliana* Col-0 and the NCBI_v3.43 annotation[59] for *S. bicolor* BTx623. For *Z. mays* L. cultivar B73 promoters, we used experimentally determined TSSs[60] and supplemented this set with the B73_RefGen_v4.42 annotation[61] for genes without an experimentally confirmed TSS. The core promoter sequences were ordered as an oligo pool from Twist Biosciences.

The STARR-seq plasmids used herein are based on the plasmid pPSup (https://www.addgene.org/149416/; ref. [35]). It harbours a phosphinothricin resistance gene (BlpR) and a GFP reporter construct terminated by the polyA site of the *A. thaliana* ribulose bisphosphate carboxylase small-chain 1A gene in the transfer DNA region. The plant core promoters followed by a 5′ UTR from maize (Zm00001d041672; used for the *Arabidopsis*, maize and validation promoter libraries) or sorghum (SORBI_3010G047100; used the sorghum promoter library) histone H3 gene, an ATG start codon and a 12-bp random barcode (VNNVNNVNNVNN; V = A, C or G) was cloned in front of the second codon of GFP by Golden Gate cloning[62]. For control constructs, the 35S minimal promoter was used instead of the plant core promoters. Each library was bottlenecked to contain, on average, 10–20 barcodes per promoter. The 35S core was inserted upstream of the core promoters by Golden Gate cloning. The sequences of the 5′ UTRs and the 35S enhancer and minimal promoter are listed in Supplementary Table 8. All primers are listed in Supplementary Table 9. The STARR-seq plasmid libraries were introduced into *Agrobacterium tumefaciens* GV3101 strain harbouring the helper plasmid pSoup[63] by electroporation.

Tobacco cultivation and transformation. Tobacco (*Nicotiana benthamiana*) was grown in soil (Sunshine Mix no. 4) at 25 °C in a long-day photoperiod (16 h light and 8 h dark; cool-white fluorescent lights (Philips TL-D 58 W/840; intensity 300 µmol m^{-2} s^{-1}). Plants were transformed 3–4 weeks after germination. For transformation, an overnight culture of *A. tumefaciens* was diluted into 100 ml of YEP medium (1% (w/v) yeast extract, 2% (w/v) peptone) and grown at 28 °C to an optical density (OD) of ~1. A 5-ml input sample of the cells was taken and plasmids were isolated from it. The remaining cells were harvested and resuspended in 100 ml of induction medium (M9 medium supplemented with 1% (w/v) glucose, 10 mM MES, pH 5.2, 100 µM $CaCl_2$, 2 mM $MgSO_4$ and 100 µM acetosyringone). After overnight growth, the bacteria were harvested, resuspended in infiltration solution (10 mM MES, pH 5.2, 10 mM $MgCl_2$, 150 µM acetosyringone and 5 µM lipoic acid) to an OD of 1 and infiltrated into the first two mature leaves of three to six tobacco plants. The plants were further grown for 48 h under normal conditions or in the dark before mRNA extraction.

Maize protoplast generation and transformation. We used a slightly modified version of a published protoplasting and electroporation protocol[64]. Maize (*Z. mays* L. cultivar B73) seeds were germinated for 4 d in the light and the seedlings were grown in soil at 25 °C in the dark for 9 d. The centre 8–10 cm of the second leaf from ten to 12 plants were cut into thin strips perpendicular to the veins and immediately submerged in 10 ml of protoplasting solution (0.6 M mannitol, 10 mM MES, 15 mg ml^{-1} cellulase R-10 (GoldBio), 3 mg ml^{-1} Macerozyme R-10 (GoldBio), 1 mM $CaCl_2$, 5 mM β-mercaptoethanol, 0.1% (w/v) BSA, pH 5.7). The mixture was covered to keep out light, vacuum infiltrated for 30 min and incubated with 40 r.p.m. shaking for 2 h. Protoplasts were released with 80 r.p.m. shaking for 5 min and filtered through a 40 µm filter. The protoplasts were harvested by centrifugation (3 min at 200*g*, room temperature) in a round-bottom glass tube and washed with 3 ml of ice-cold electroporation solution (0.6 M mannitol, 4 mM MES, 20 mM KCl, pH 5.7). After centrifugation (2 min at 200*g*, room temperature), the cells were resuspended in 3 ml of ice-cold electroporation solution and counted. Approximately one million cells were mixed with 25 µg of plasmid DNA in a total volume of 300 µl, transferred to a 4-mm electroporation cuvette and incubated for 5 min on ice. The cells were electroporated (300 V, 25 µFD, 400 Ω) and 900 µl of ice-cold incubation buffer (0.6 M mannitol, 4 mM MES, 4 mM KCL, pH 5.7) was added. After 10 min of incubation on ice, the cells were further diluted with 1.2 ml of incubation buffer and kept at 25 °C in the dark for 16 h before mRNA collection. To cover each library, four electroporation reactions were performed, except for the smaller validation libraries in which two electroporation reactions were performed. For the maize protoplast STARR-seq, the plasmid library used for electroporation was sequenced as the input sample.

STARR-seq assay. For each STARR-seq experiment, two independent biological replicates were performed. Different plants and fresh *Agrobacterium* cultures were used for each biological replicate and the replicates were performed on different days. For experiments in tobacco, 12 transformed leaves were collected from six plants. They were frozen in liquid nitrogen, ground in a mortar and immediately resuspended in 25 ml of TRIzol (Thermo Fisher Scientific). The suspension was cleared by centrifugation (5 min at 4,000*g*, 4 °C) and the supernatant was thoroughly mixed with 5 ml of chloroform. After centrifugation (15 min at 4,000*g*, 4 °C), the upper, aqueous phase was transferred to a new tube, mixed with 5 ml of chloroform and centrifuged again (15 min at 4,000*g*, 4 °C). Then 13 ml of the upper, aqueous phase was transferred to new tubes and RNA was precipitated with 1.3 ml of 8 M LiCl and 32.5 ml of 100% (v/v) ethanol by incubation at −80 °C for 15 min. The RNA was pelleted (30 min at 4,000*g*, 4 °C), washed with 10 ml of 70% (v/v) ethanol, centrifuged again (5 min at 4,000*g*, 4 °C) and resuspended in 1.5 ml of nuclease-free water. The solution was split into two halves and mRNAs were isolated from each using 150 µl of magnetic Oligo$(dT)_{25}$ beads (NEB) according to the manufacturer's protocol. The mRNAs were eluted in 40 µl. The two samples per library were pooled and supplemented with 10 µl of DNase I buffer, 10 µl of 100 mM $MnCl_2$, 2 µl of DNase I (Thermo Fisher Scientific) and 1 µl of RNaseOUT (Thermo Fisher Scientific). After 1 h incubation at 37 °C, 2 µl of 20 mg ml^{-1} glycogen (Thermo Fisher Scientific), 10 µl of 8 M LiCl and 250 µl of 100% (v/v) ethanol were added to the samples. Following precipitation at −80 °C, centrifugation (30 min at 20,000*g*, 4 °C) and washing with 200 µl of 70% (v/v) ethanol (5 min at 20,000*g*, 4 °C), the pellet was resuspended in 100 µl of nuclease-free water. Eight reactions with 5 µl of mRNA each and a GFP construct-specific primer were prepared for complementary DNA synthesis using SuperScript IV reverse transcriptase (Thermo Fisher Scientific) according to the manufacturer's instructions. Half of the reactions were used as no reverse transcription control, in which the enzyme was replaced with water. After cDNA synthesis, the reactions were pooled and purified with DNA Clean & Concentrator-5 columns (Zymo Research). The barcode region was amplified with 10–20 cycles of polymerase chain reaction (PCR) and read out by next generation

sequencing. For the smaller validation libraries, only six leaves were used and all volumes except the reverse transcription were halved.

For the STARR-seq assay in maize protoplasts, transformed protoplasts were harvested by centrifugation (3 min at 200*g*, 4 °C) 16 h after electroporation. The protoplasts were washed three times with 1 ml of incubation buffer and centrifuged for 2 min at 200*g* and 4 °C. The cells were resuspended in 600 µl of TRIzol (Thermo Fisher Scientific) and incubated for 5 min at room temperature. The suspension was thoroughly mixed with 120 µl of chloroform and centrifuged (15 min at 20,000*g*, 4 °C). The upper, aqueous phase was transferred to a new tube, mixed with 120 µl of chloroform and centrifuged again (15 min at 20,000*g*, 4 °C). RNA was precipitated from 400 µl of the supernatant with 1 µl of 20 mg ml^{-1} glycogen (Thermo Fisher Scientific), 40 µl of 8 M LiCl and 1 ml of 100% (v/v) ethanol by incubation at −80 °C for 15 min. After centrifugation (30 min at 20,000*g*, 4 °C), the pellet was washed with 200 µl of 70% (v/v) ethanol, centrifuged again (5 min at 20,000*g*, 4 °C) and resuspended in 200 µl of nuclease-free water. The mRNAs were isolated from this solution using 50 µl of magnetic Oligo(dT)$_{25}$ beads (NEB) according to the manufacturer's protocol and the mRNAs were eluted in 40 µl of water. DNase I treatment and precipitation were performed as for the mRNAs obtained from tobacco plants but with half the volume. Reverse transcription, purification, PCR amplification and sequencing were performed as for the tobacco samples.

Subassembly and barcode sequencing. Paired-end sequencing on an Illumnia NextSeq 550 system was used for the subassembly of promoters with their corresponding barcodes. The promoter region was sequenced using partially overlapping, paired 144-bp reads and two 15-bp indexing reads were used to sequence the barcodes. The promoter and barcode reads were assembled using PANDAseq[65] and the promoters were aligned to the designed core promoter sequences. Promoter-barcode pairs with less than five reads and promoters with a mutation or truncation were discarded. Barcode sequencing was performed using paired-end reads on a Illumnia NextSeq 550 platform. The reads were trimmed to only the barcode portion assembled with PANDAseq. All sequencing results were deposited in the NCBI Sequence Read Archive under the BioProject accession PRJNA714258. The scripts used for processing the raw reads are available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters.

Computational methods. For analysis of the STARR-seq experiments, the reads for each barcode were counted in the input and cDNA samples. Barcode counts below five were discarded. Barcode enrichment was calculated by dividing the barcode frequency (barcode counts divided by all counts) in the cDNA sample by that in the input sample. The enrichment of the promoters was calculated as the median enrichment of all barcodes linked to them. We calculated the promoter strength as the $\log_2$ of the promoter enrichment normalized to the enrichment of 35S minimal promoter. We used the average promoter strength from both replicates for all analyses. Spearman and Pearson correlations were calculated using the base R function. Significance was determined using the two-sided Wilcoxon rank-sum test as implemented in base R. GO-term enrichment analysis was performed using the ggprofiler2 (v.0.1.9; ref. [66]) library for R and a custom gmt file with GOslim terms. Gene expression data was obtained from the EMBL-EBI Expression Atlas (https://www.ebi.ac.uk/gxa/about.html) using experiments E-MTAB-7978 (ref. [44]), E-GEOD-50191 (ref. [42]) and E-MTAB-5956 (ref. [43]) for *Arabidopsis*, maize and sorghum, respectively. The tissue-specificity index τ was calculated as previously published[49]. Sequences for TFIIB proteins were obtained from Uniprot (https://www.uniprot.org/; see Supplementary Table 10 for accession numbers) and aligned using Clustal Omega[67]. The code used for analyses is available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters.

Prediction of core promoter elements and TF-binding sites. The TATA box and Inr motifs were obtained from the plant promoter database[68] and for each a consensus motif was created by merging the motifs from dicot and monocot promoters using the universalmotif (v.1.6.3) library for R. Motifs for BREu and BREd were obtained from JASPAR[69]. The motifs for the polypyrimidine initiator TCT and the Y patch were created from published sequences of these elements[17,19]. Binding site motifs for *Arabidopsis* TFs were obtained from the PlantTFDB[70]. TF motifs were clustered by similarity using the compare_motifs() function from the R library universalmotif. The original clusters were improved by manual inspection and reannotation. Consensus motifs for the final TF motifs were created using the merge_motifs() function from universalmotif. Meme files with the motifs used in this study are available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters. Promoter sequences were analysed with the universalmotif library assuming a neutral background nucleotide frequency. For the initiator elements, only the last ten (Inr) or the last six (TCT) bases were scanned. For BREu and BREd, the sequences immediately upstream and downstream of the highest scoring TATA box were analysed. For each sequence, the maximum motif score was calculated and normalized to the minimum (set to 0) and maximum (set to 1) scores possible. Sequences with a score of at least 0.85 were considered positive. For testing the effect of the BREu and BREd motifs (Extended Data Fig. 3), only sequences with a TATA box score of at least 0.7 were considered.

Design of validation sequences. To directly validate the importance of the TATA box, BRE^u and BRE^d elements, we picked 30 promoters (ten each from *Arabidopsis*, maize and sorghum if possible) according to the following criteria: for mutations of a canonical TATA box, we selected promoters with a TATA box motif score >0.9 in the −59 to −23 region. The two conserved T nucleotides in the core TATA motif were replaced individually or together with Gs. We also selected 30 promoters with a maximum TATA box motif score of 0.7 to 0.75. This weak TATA box was replaced with either a canonical TATA box motif (TATAAAT) or a mutated version of it (TAGAAAT). For the BRE elements, we first filtered for promoters with a TATA box motif score of at least 0.85 in the −59 to −23 region. From these, we picked promoters with a BRE motif score >0.85. For the BRE^u element, we mutated bases 3, 6 and 7 to T, A and A respectively. For the BRE^d element, we mutated bases 2,4 and 6 to A. We also selected promoters where both the BRE^u and the BRE^d motif scores are <0.5 to insert either a canonical BRE^u (AGCGCGCC) or BRE^d (GTTTGTT) element.

Synthetic promoter design. Synthetic promoters were designed by generating 170-bp long random sequences with a nucleotide composition similar to an average *Arabidopsis* (35.2% A, 16.6% C, 15.3% G, 32.8% T) or maize (24.5% A, 29.0% C, 22.5% G, 23.9% T) promoter. We filtered out any random sequence with motif scores higher than 0.75 for a TATA box, Inr or Y patch element or for TF-binding site of clusters 1, 15, 16 or 22. Promoters containing recognition sites for the restriction enzymes used for cloning (BsaI and BbsI) were also removed. From each set of promoters (*Arabidopsis* or maize nucleotide composition) that passed the filters, we randomly selected ten variants for further modification. The promoters were kept as is or modified with a TATA box (TATAAATA) at positions 133–140, a Y patch (A and G nucleotides of the promoter were changed to C) at positions 147–154 and/or an Inr element (yyyyTCAyyy, where y indicates a change of A to T or G to C) at positions 160–169. To study the effect of TFs, the synthetic promoters with the TATA box were chosen as backgrounds. Binding sites for NAC (cluster 1, TTACGTGnnnnACAAG, where n represents bases of the promoter background), TCP (cluster 15, TGGGGCCCAC and cluster 22, GGGACCAC) or HSF/S1Fa-like (cluster 16, GAAGCTTCTAGAA) TFs were inserted at various positions of these promoters.

Computational modelling of promoter strength. To predict promoter strength, we built separate models for the tobacco leaf and the maize protoplast system. We used the results from the libraries with the 35S enhancer in the dark for training and validation. The models were trained on a set of 90% of all measured promoters and tested against the held-out set of the remaining 10% of the promoters.

We used the base R function lm() to build a linear model for predicting promoter strength on the basis of the promoter's GC content and its maximum motif score for six core promoter elements (TATA box, Inr, TCT, BRE^u, BRE^d and Y patch) and 72 consensus TF-binding motifs.

To build a direct sequence to promoter strength model we built a CNN using the tensorflow (v.2.2) package in python. The model consists of two forward- and reverse-sequence scan layers adapted from DeepGMAP[71] with 128 filters and a kernel width of 13 that feed into a regular convolutional layer (128 filters, kernel width 13, ReLU activation). Each convolutional layer is followed by a dropout layer with a 0.15 dropout rate. The output of the convolutional layers is fed into a dense layer with 64 filters with batch-normalization and ReLU activation that is followed by a final dense layer generating the single output. We initialized the first convolutional layer kernel with the clustered TF motifs. The source code and the models are available on GITHUB.

In silico evolution of promoter sequences. We used the CNNs to improve promoter performance in an iterative fashion. In each round, we generated all possible single nucleotide variants of a given promoter, scored them with the CNN models and kept the variant with the highest predicted activity for the next round. The sequences were scored with either just one of the models trained on the tobacco leaf or the maize protoplast data or with both models in which case the mean of both predictions was used to select the best-performing variant. We experimentally tested these sequences after three and ten rounds of this process. For the evolution, we selected native promoters showing either weak, intermediate or strong activity in both assay systems or were strong in one system and weak in the other one. Additionally, we also performed the in silico evolution with the synthetic promoters described above.

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

Data availability

All sequencing results are deposited in the NCBI Sequence Read Archive under the BioProject accession PRJNA714258.

Code availability

The code used in this study is available on Github (https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters).

Received: 13 January 2021; Accepted: 27 April 2021;
Published online: 3 June 2021

References

1. Liu, W. & Stewart, C. N. Plant synthetic biology. *Trends Plant Sci.* **20**, 309–317 (2015).
2. Lomonossoff, G. P. & D'Aoust, M.-A. Plant-produced biopharmaceuticals: a case of technical developments driving clinical deployment. *Science* **353**, 1237–1240 (2016).
3. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
4. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
5. Ricci, W. A. et al. Widespread long-range *cis*-regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
6. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
7. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).
8. Grosschedl, R. & Birnstiel, M. L. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc. Natl Acad. Sci. USA* **77**, 1432–1436 (1980).
9. Wasylyk, B. et al. Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in T-A-T-A box homology sequence. *Proc. Natl Acad. Sci. USA* **77**, 7024–7028 (1980).
10. Smale, S. T. & Baltimore, D. The "initiator" as a transcription control element. *Cell* **57**, 103–113 (1989).
11. Ince, T. A. & Scotto, K. W. A conserved downstream element defines a new class of RNA polymerase II promoters. *J. Biol. Chem.* **270**, 30249–30252 (1995).
12. Burke, T. W. & Kadonaga, J. T. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA box-deficient promoters. *Genes Dev.* **10**, 711–724 (1996).
13. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **12**, 34–44 (1998).
14. Lewis, B. A., Kim, T.-K. & Orkin, S. H. A downstream element in the human β-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl Acad. Sci. USA* **97**, 7172–7177 (2000).
15. Lim, C. Y. et al. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* **18**, 1606–1617 (2004).
16. Deng, W. & Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **19**, 2418–2423 (2005).
17. Parry, T. J. et al. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.* **24**, 2013–2018 (2010).
18. Molina, C. & Grotewold, E. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genom.* **6**, 25 (2005).
19. Yamamoto, Y. Y. et al. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res.* **35**, 6219–6226 (2007).
20. Bernard, V., Brunaud, V. & Lecharny, A. TC-motifs at the TATA box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genom.* **11**, 166 (2010).
21. Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W. & Azizkhan, J. C. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol. Cell. Biol.* **10**, 6632–6641 (1990).
22. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
23. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
24. Lubliner, S. et al. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* **25**, 1008–1017 (2015).
25. Arnold, C. D. et al. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* **35**, 136–144 (2017).
26. van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).
27. Weingarten-Gabbay, S. et al. Systematic interrogation of human promoters. *Genome Res.* **29**, 171–183 (2019).
28. de Boer, C. G. et al. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
29. Kotopka, B. J. & Smolke, C. D. Model-driven generation of artificial yeast promoters. *Nat. Commun.* **11**, 2113 (2020).
30. Kumari, S. & Ware, D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS ONE* **8**, e79011 (2013).
31. Morton, T. et al. Paired-end analysis of transcription start sites in *Arabidopsis* reveals plant-specific promoter signatures. *Plant Cell* **26**, 2746–2760 (2014).
32. Zhu, Q., Dabi, T. & Lamb, C. TATA box and initiator functions in the accurate transcription of a plant minimal promoter in vitro. *Plant Cell* **7**, 1681–1689 (1995).
33. Kiran, K. et al. The TATA box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. *Plant Physiol.* **142**, 364–376 (2006).
34. Srivastava, R. et al. Distinct role of core promoter architecture in regulation of light-mediated responses in plant genes. *Mol. Plant* **7**, 626–641 (2014).
35. Jores, T. et al. Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell* **32**, 2120–2131 (2020).
36. Cai, Y.-M. et al. Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res.* **48**, 11845–11856 (2020).
37. Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z. & Zhu, J.-K. UTR-dependent control of gene expression in plants. *Trends Plant Sci.* **23**, 248–259 (2018).
38. Fang, R. X., Nagy, F., Sivasubramaniam, S. & Chua, N. H. Multiple *cis* regulatory elements for maximal expression of the cauliflower mosaic virus 35S promoter in transgenic plants. *Plant Cell* **1**, 141–150 (1989).
39. Benfey, P. N., Ren, L. & Chua, N. H. Tissue-specific expression from CaMV 35S enhancer subdomains in early stages of plant development. *EMBO J.* **9**, 1677–1684 (1990).
40. Bruce, W. B., Christensen, A. H., Klein, T., Fromm, M. & Quail, P. H. Photoregulation of a phytochrome gene promoter from oat transferred into rice by particle bombardment. *Proc. Natl Acad. Sci. USA* **86**, 9692–9696 (1989).
41. Yahraus, T., Chandra, S., Legendre, L. & Low, P. S. Evidence for a mechanically induced oxidative burst. *Plant Physiol.* **109**, 1259–1266 (1995).
42. Walley, J. W. et al. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).
43. Wang, B. et al. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Res.* **28**, 921–932 (2018).
44. Mergner, J. et al. Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* **579**, 409–414 (2020).
45. Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant genomes. *Trop. Plant Biol.* **9**, 136–149 (2016).
46. Rensink, W. A. et al. Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genom.* **6**, 124 (2005).
47. Tsai, F. T. F. & Sigler, P. B. Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO J.* **19**, 25–36 (2000).
48. Gehrig, J. et al. Automated high-throughput mapping of promoter–enhancer interactions in zebrafish embryos. *Nat. Methods* **6**, 911–916 (2009).
49. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
50. Heerah, S., Katari, M., Penjor, R., Coruzzi, G. & Marshall-Colon, A. WRKY1 mediates transcriptional regulation of light and nitrogen signaling pathways. *Plant Physiol.* **181**, 1371–1388 (2019).
51. Cuperus, J. T. et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).
52. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
53. Klein, J. C. et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
54. Hong, C. K. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.08.434469 (2021).
55. Dorrity, M. W. et al. The regulatory landscape of *Arabidopsis thaliana* roots at single-cell resolution. *Nat. Commun.* https://doi.org/10.1038/s41467-021-23675-y (2021).
56. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A *cis*-regulatory atlas in maize at single-cell resolution. *Cell* https://doi.org/10.1016/j.cell.2021.04.014 (2021).
57. Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H. & Wang, J.-W. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nat. Commun.* **12**, 2053 (2021).

58. Cheng, C.-Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
59. McCormick, R. F. et al. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **93**, 338–354 (2018).
60. Mejía-Guerra, M. K. et al. Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *Plant Cell* **27**, 3309–3320 (2015).
61. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
62. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
63. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **42**, 819–832 (2000).
64. Sheen, J. Metabolic repression of transcription in higher plants. *Plant Cell* **2**, 1027–1038 (1990).
65. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinform.* **13**, 31 (2012).
66. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
67. Madeira, F. et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
68. Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M. & Solovyev, V. V. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.* **31**, 114–117 (2003).
69. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
70. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* **48**, D1104–D1113 (2020).
71. Onimaru, K., Nishimura, O. & Kuraku, S. Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. *PLoS ONE* **15**, e0235748 (2020).

Acknowledgements

We thank A. Gutierrez Diaz and E. Grotewold for providing maize TSS data, and A. Gallavotti for providing maize B73 seeds. This work was supported by the National Science Foundation (RESEARCH-PGR grant no. 1748843 to E.S.B., S.F. and C.Q.), the German Research Foundation (DFG; fellowship no. 441540116 to T.J.) and the National Institutes of Health (T32 training grant no. HG000035 to J.T. and R01-GM079712 to C.Q. and J.T.C.).

Author contributions

All authors conceived and interpreted experiments and wrote the article. T.J. and J.T. performed experiments. T.J. analysed the data and prepared the figures. T.J. and T.W. did the in silico modelling.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-021-00932-y.

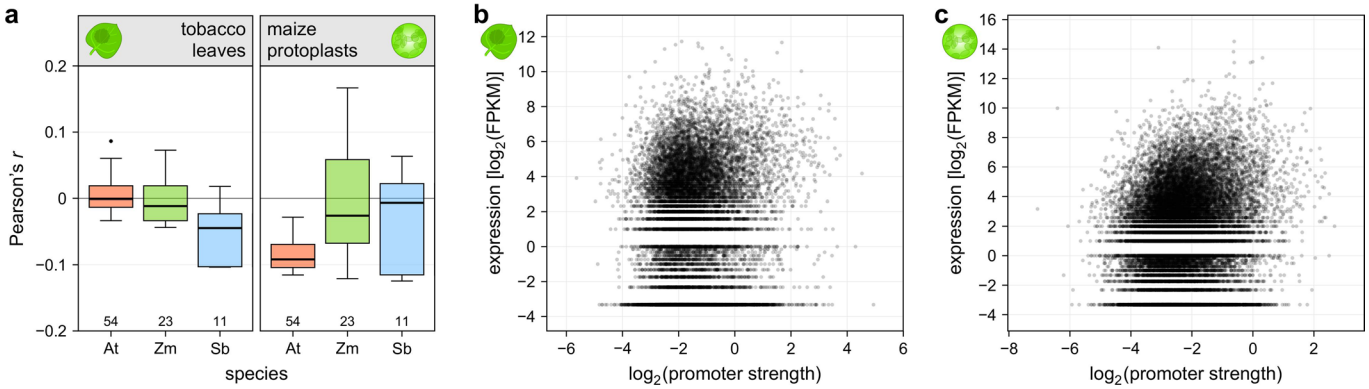
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-021-00932-y.

Correspondence and requests for materials should be addressed to J.T.C., S.F. or C.Q.

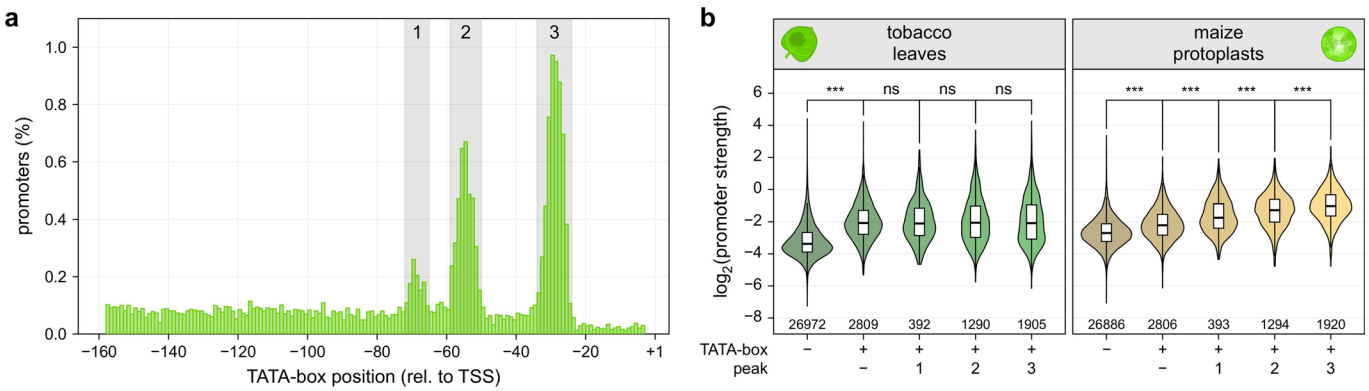
Peer review information *Nature Plants* thanks Philip Benfey, Shira Weingarten-Gabbay and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

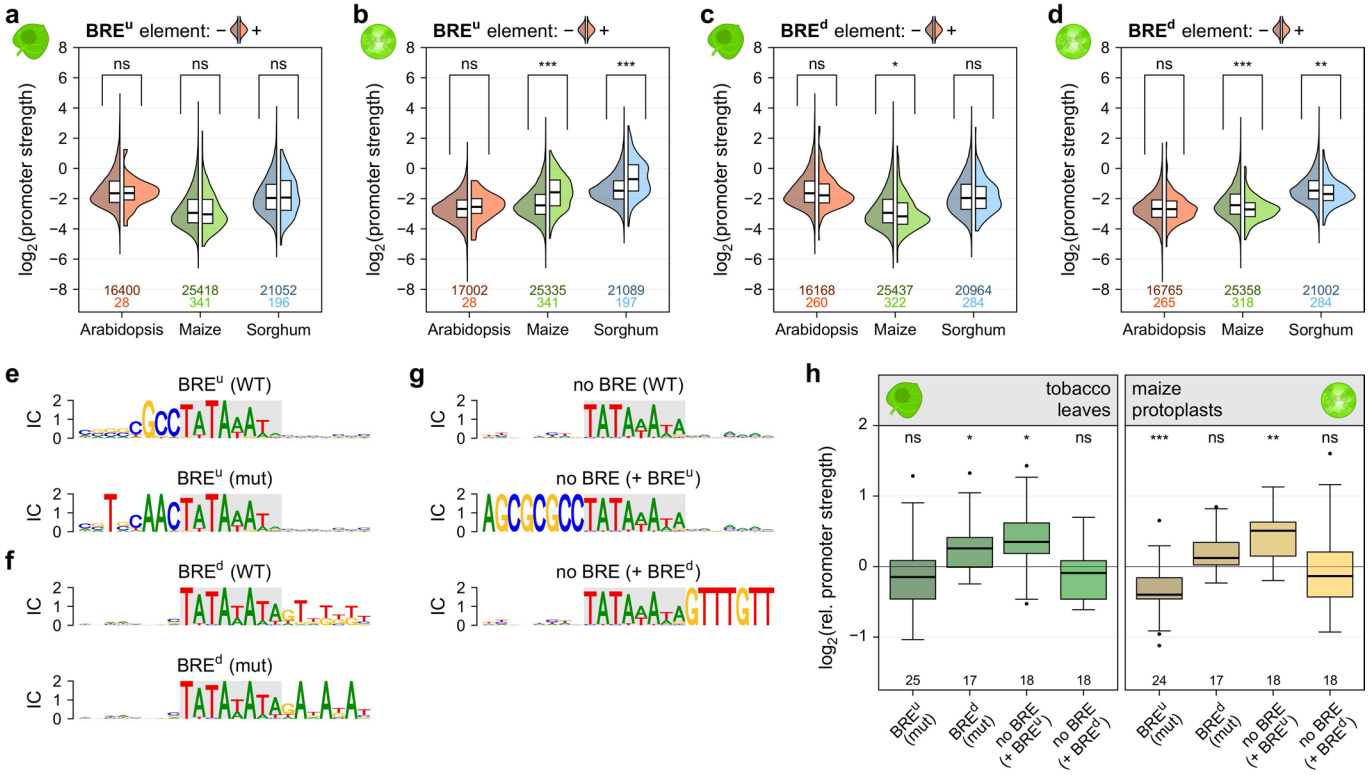
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



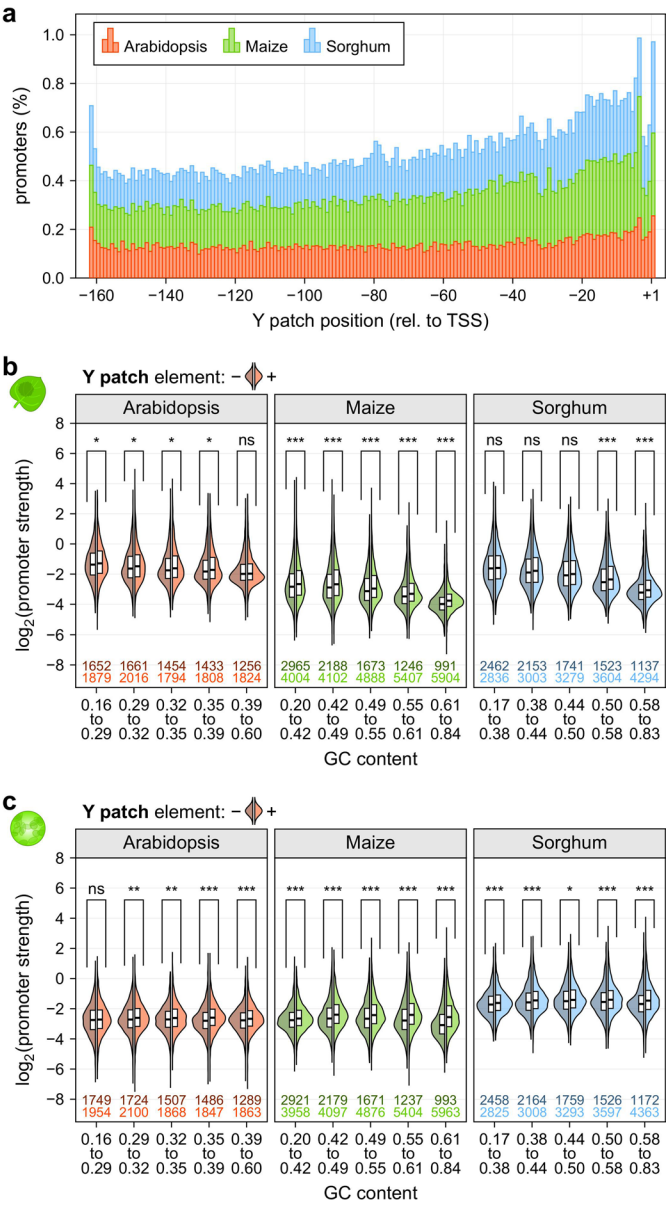
Extended Data Fig. 1 | Promoter strength and in vivo expression levels of corresponding genes are not correlated. a, Correlation (Pearson's *r*) between the promoter strength and expression levels of the corresponding genes in the indicated species. Each boxplot (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; points, outliers) represents the correlation for all individual tissue samples in the RNA-seq dataset (see Methods). The number of samples in the RNA-seq dataset is indicated at the bottom of the plot. **b**,**c**, Examples of the correlation between gene expression (Arabidopsis adult cotyledon (**b**) or maize root cortex (**c**) samples) and promoter strength as determined in tobacco leaves (**b**) or maize protoplasts (**c**). These examples correspond to the highest correlations in (**a**).



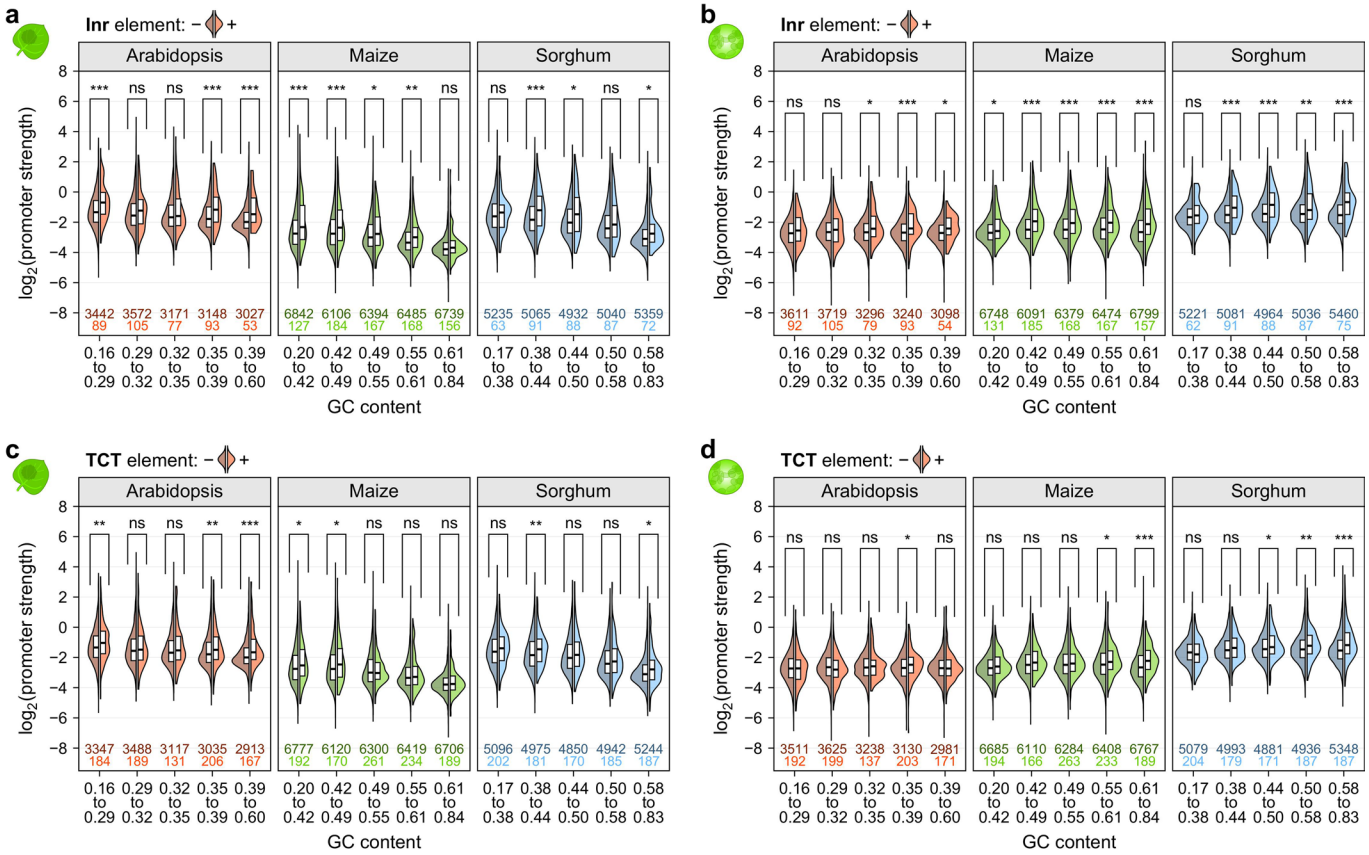
Extended Data Fig. 2 | Strength of maize promoters depends on the TATA box location in maize protoplasts. a, Histogram showing the percentage of maize promoters with a TATA box at the indicated position (reproduced from Fig. 4). Three peaks in the distribution of TATA boxes are highlighted in grey. Peak 1 spans bases −72 to −65, peak 2 spans bases −59 to −50, and peak 3 spans bases −34 to −24. **b**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for maize promoters without enhancer in the indicated assay system. Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within one of the three peaks highlighted in (**a**).



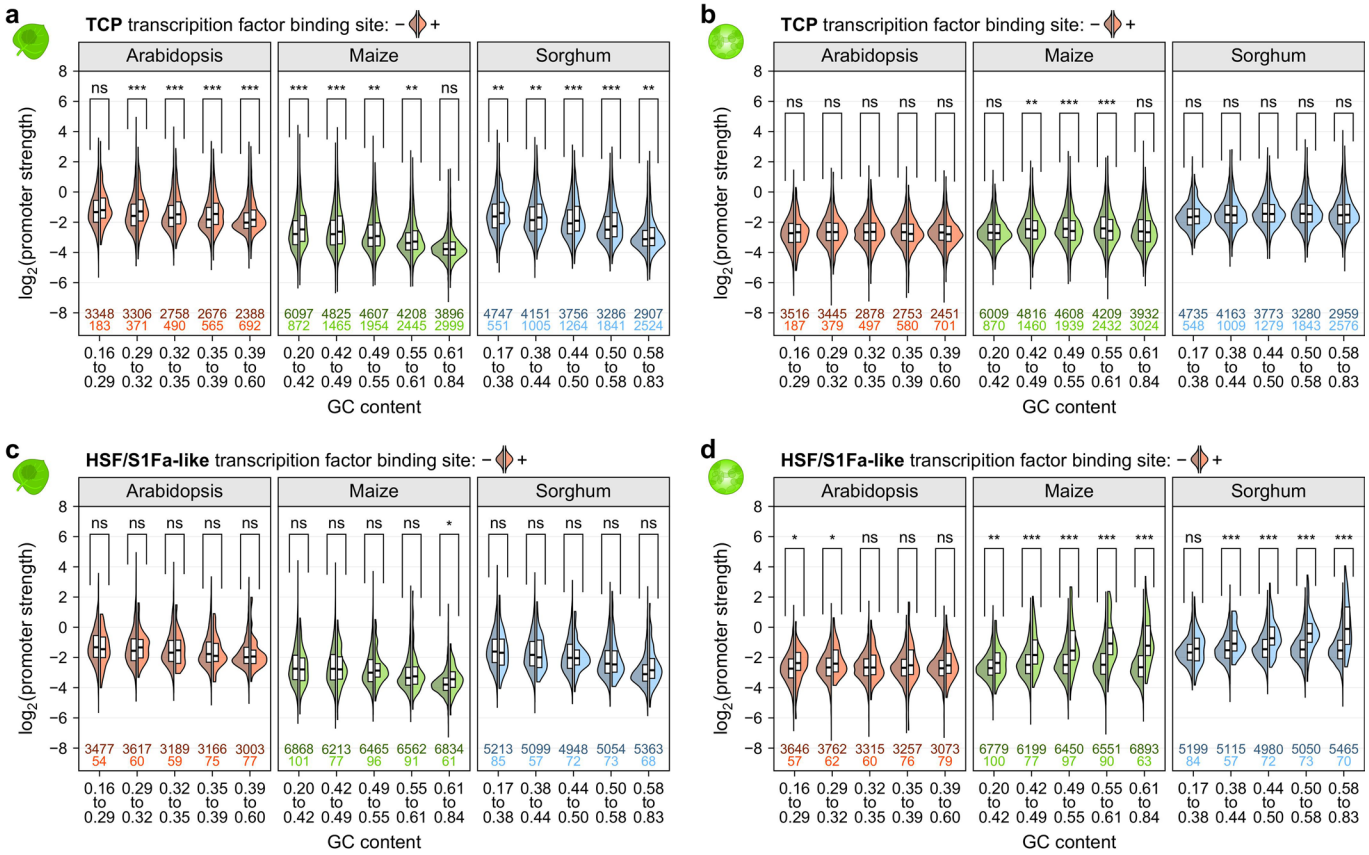
Extended Data Fig. 3 | The BREu element is most active in maize protoplasts. a-d, Violin plots of promoter strength in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters with a strong or intermediate TATA box (motif score ≥ 0.7; see Methods) were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a BRE^u (**a**,**b**), or BRE^d (c,d) element. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots. **e**,**f**, Logoplots for promoters with a BRE^u (**e**) or BRE^d (**f**) before (WT) and after (mut) introducing mutations that disrupt the elements. **g**, Logoplots for promoters without a BRE (WT) and with an inserted BRE^u (+ BRE^u) or BRE^d (+ BRE^d) element. **h**, Boxplots and significance levels (as defined in Fig. 4) for the relative strength of the promoter variants shown in (**e**-**g**). The corresponding WT promoter was set to 0 (horizontal black line).



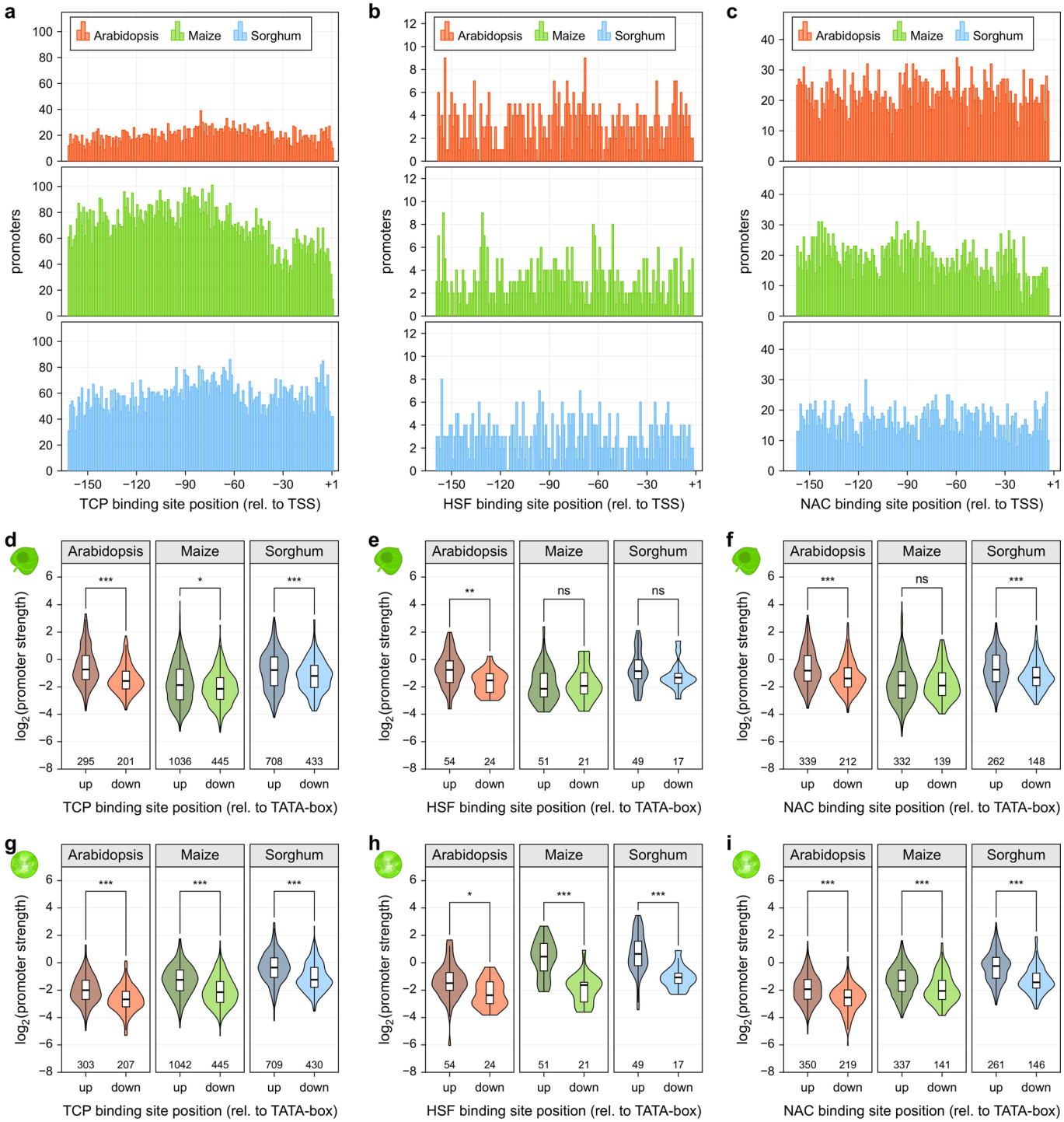
Extended Data Fig. 4 | The Y patch is a plant-specific core promoter element. a, Histogram showing the percentage of promoters with a TATA box at the indicated position. **b**,**c**, Violin plots of promoter strength in tobacco leaves (**b**) or maize protoplasts (**c**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a Y patch. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



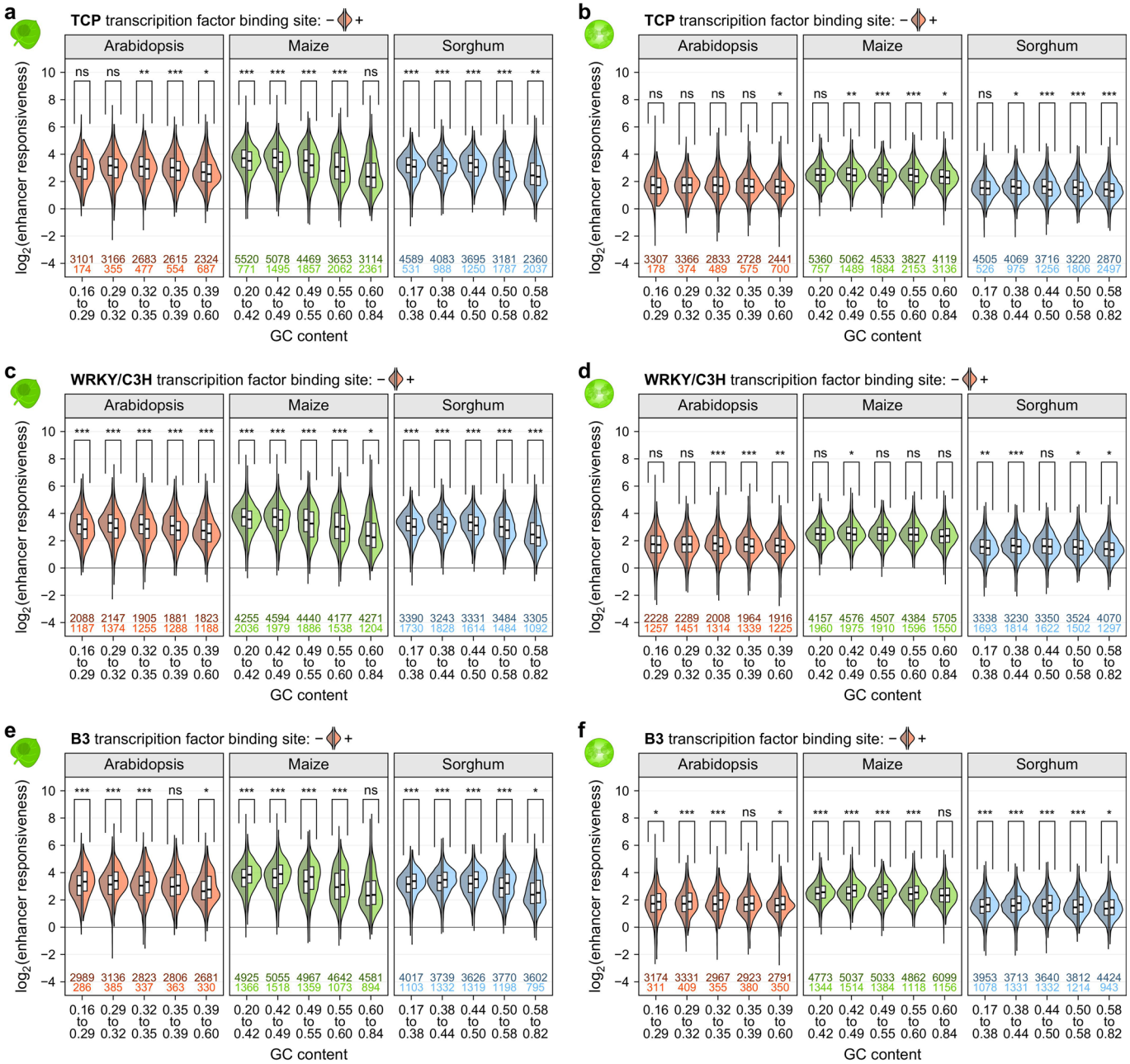
Extended Data Fig. 5 | Core promoter elements at the TSS influence promoter strength. a-d, Violin plots of promoter strength in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) an Inr (**a**,**b**), or TCT (**c**,**d**) element at the TSS. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



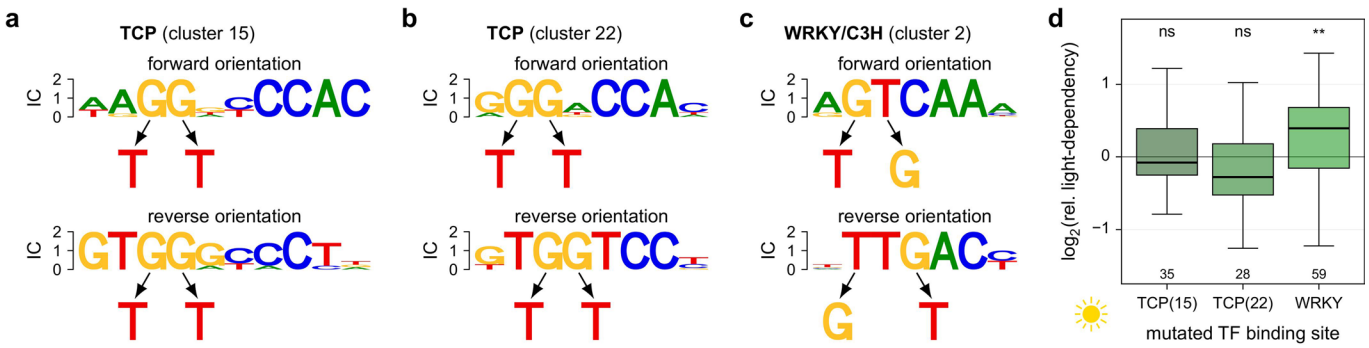
Extended Data Fig. 6 | Transcription factor binding sites contribute to promoter strength in an assay system-dependent manner. a-d, Violin plots of promoter strength for libraries without enhancer in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a binding site for TCP (**a**,**b**) or HSF (**c**,**d**) transcription factors. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



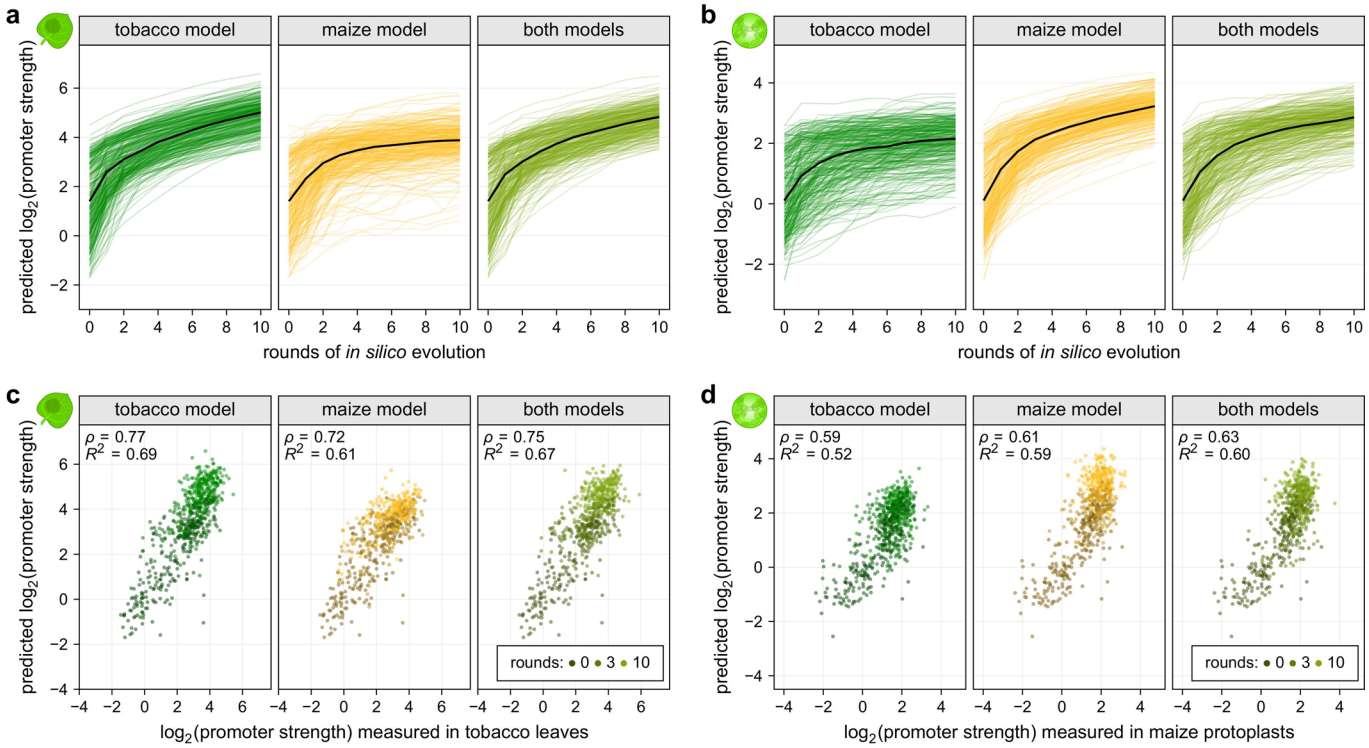
Extended Data Fig. 7 | Transcription factor binding sites are more active upstream of the TATA box. a-**c**, Histograms showing the number of promoters with a TCP (**a**), HSF (**b**), or NAC (**c**) transcription factor binding site at the indicated position. **d**-**i**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves (**d**-**f**) or maize protoplasts (**g**-**i**). Promoters were grouped by the position of their TCP (**d**,**g**), HSF (**e**,**h**), or NAC (**f**,**i**) transcription factor binding site relative to the TATA box: either upstream (up) or downstream (down).



Extended Data Fig. 8 | Promoter-proximal transcription factor binding sites influence enhancer responsiveness. a-**f**, Violin plots of enhancer responsiveness in tobacco leaves (**a**,**c**,**e**) or maize protoplasts (**b**,**d**,**f**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a TCP (**a**,**b**), WRKY (**c**,**d**), or B3 (**e**,**f**) transcription factor binding site. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



Extended Data Fig. 9 | Mutations in transcription factor binding sites alter light-dependency. a-**c**, One or two T > G mutations were introduced in binding sites for TCP (**a**,**b**) or WRKY (**c**) transcription factors. The orientation of a binding site in the wild type promoter determined the bases that were mutated. **d**, Boxplots and significance levels (as defined in Fig. 4) for the relative light-dependency of promoters harbouring mutations in the indicated transcription factor binding site as shown in (**a**-**c**). The corresponding wild type promoter was set to 0 (horizontal black line).



Extended Data Fig. 10 | The in silico evolution of promoters is most effective in early rounds. a,**b**, 150 native and 160 synthetic promoters were subjected to 10 rounds of *in silico* evolution and the strength of the evolved promoters was predicted with the tobacco model (**a**) or the maize model (**b**). The black line represents the median promoter strength after each round. **c**,**d**, Correlation (Pearson's *R*2 and Spearman's ρ) between the predicted and experimentally determined strength of promoters after 0, 3, or 10 rounds of *in silico* evolution. Promoter strengths measured in tobacco leaves were compared to predictions from the tobacco model (**c**) and the data from maize protoplasts was compared to the predictions from the maize model (**d**). The models used for the *in silico* evolution are indicated on each plot.

nature research

Corresponding author(s): Josh T Cuperus, Stanley Fields, Christine Queitsch

Last updated by author(s): Apr 16, 2021

Reporting Summary

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

Statistics

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

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection	Sequencing was performed on a Illumina NextSeq 550 with NextSeq Control Software (version 4.0.1.41) and reads were demultiplexed using the bcl2fastq2 Conversion Software (version 2.20)
Data analysis	bowtie2 (version 2.4.1), bioawk (https://github.com/lh3/bioawk; commit fd40150), cutadapt (version 2.5), trim_galore (version 0.6.6), PANDAseq (version 2.11), R (version 4.0.0), bedtools (version 2.29.2), bedops (version 2.4.35), python (version 3.7.9), tensorflow (version 2.2) The code used for analyses is available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing results are deposited in the NCBI Sequence Read Archive under the BioProject accession PRJNA714258 (http://www.ncbi.nlm.nih.gov/bioproject/714258). The code used in this study is available on Github (https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample sizes were not statistically predetermined, but instead set based on the available genomic annotation.
Data exclusions	No data were excluded from analysis.
Replication	Two replicates were performed for each STARR-seq experiment and correlation plots are shown for all experiments (Fig. 1c,d; Supp Fig. 1 and 2). All attempts at replication were successful.
Randomization	STARR-seq experiments were performed with multiple, randomly selected plants.
Blinding	Blinding was not relevant to this study, as the experiments were performed with DNA libraries in bulk.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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