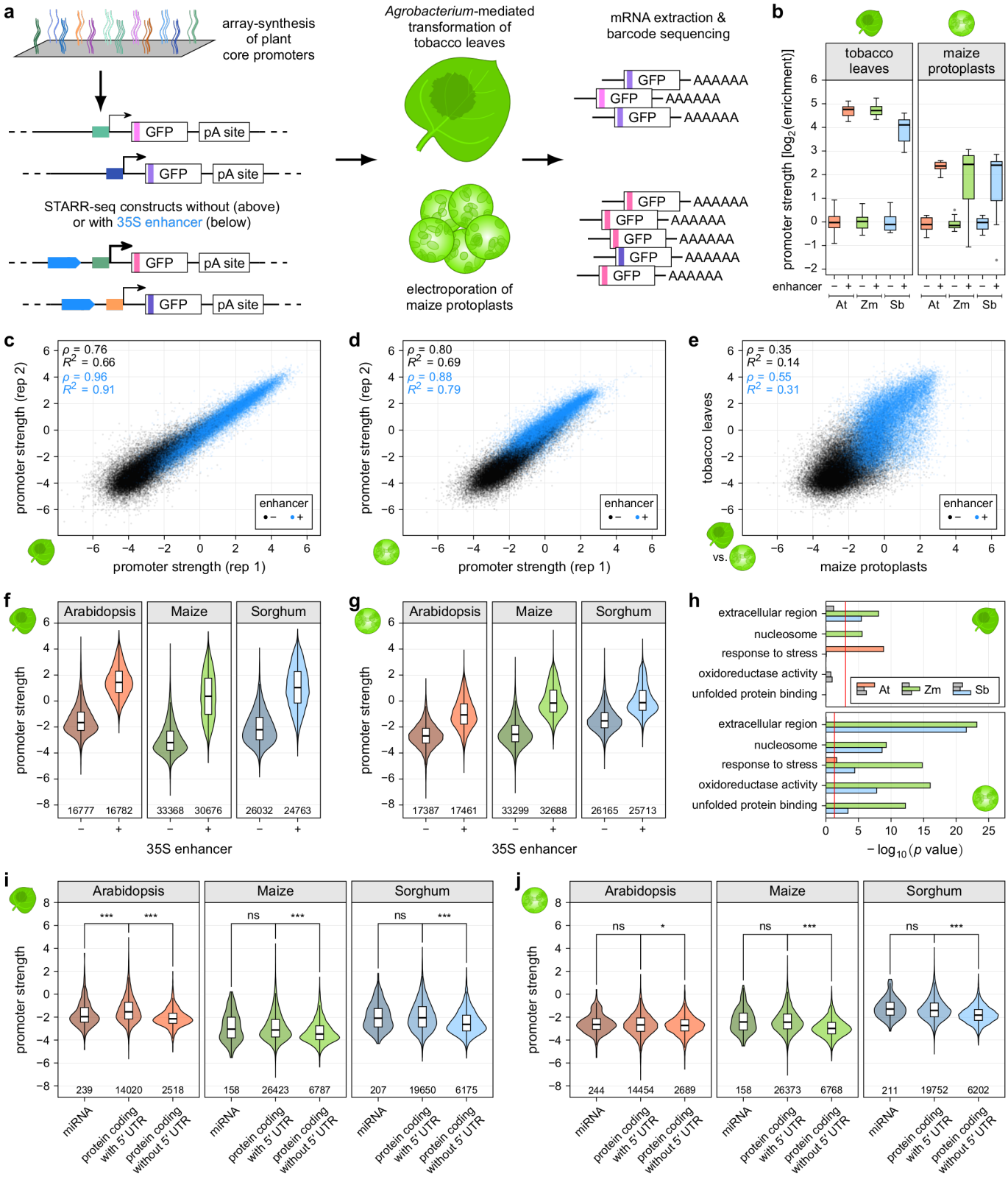


Jores *et al.*, Fig. 1

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, Arabiopsis; 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 (center 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 construct. **c**,**d**, Correlation 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. **f**,**g**, Violin plots of promoter strength as measured by STARR-seq in tobacco leaves (**f**) or maize protoplasts (**g**) for libraries without (−) or with (+) the 35S enhancer upstream of the promoter. **h**, Enrichment of selected GO terms for genes associated with the 1000 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 value ≤ 0.05). Non-significant bars are shown in gray. **i**,**j**, Violinplots of promoter strength (libraries without 35S enhancer) in tobacco leaves (**i**) or maize protoplasts (**j**). Promoters were grouped by gene type. In all figures, violinplots represent the kernel density distribution and the boxplots within represent the median (center 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 Wilcoxon rank-sum test and are indicated: *, $p \leq 0.01$; **, $p \leq 0.001$; ***, $p \leq 0.0001$; ns, not significant.

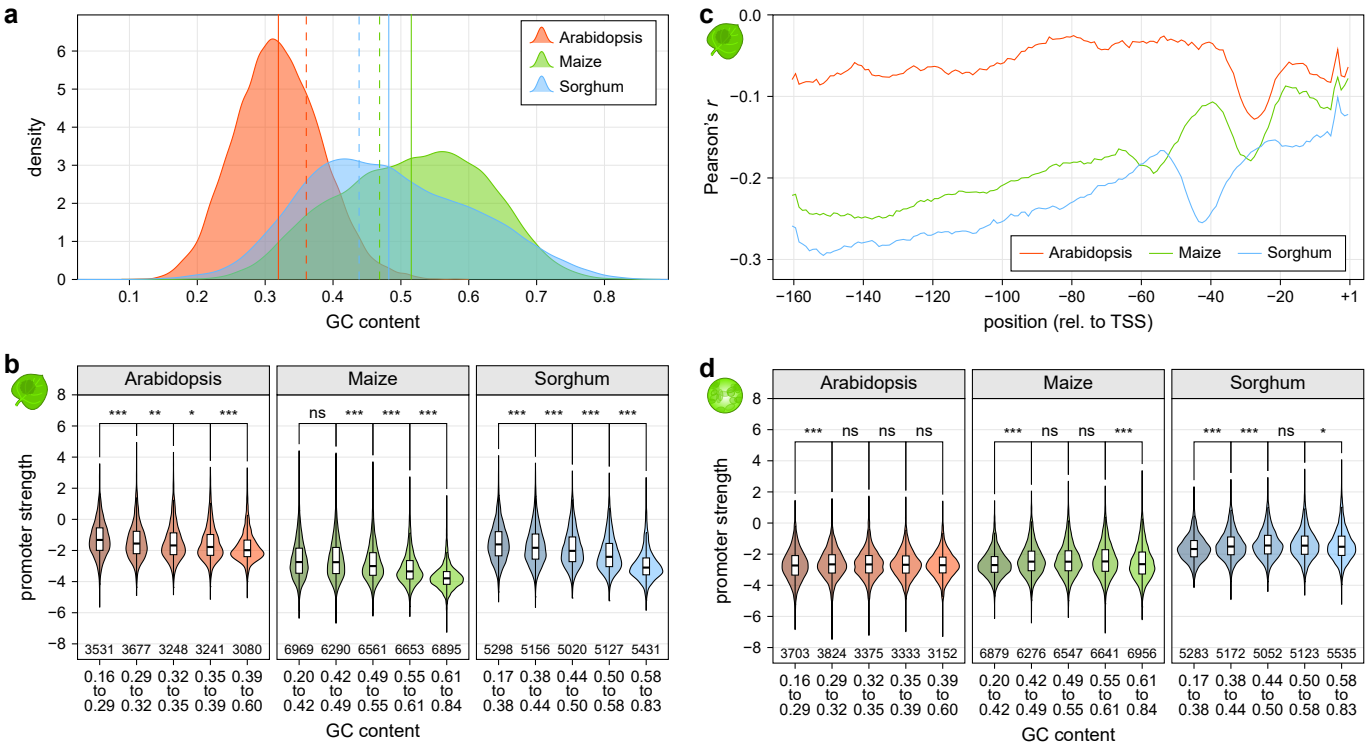


Fig. 2 | GC content affects promoter strength in tobacco leaves. a, Distribution of GC content for all promoters of the indicated species. Lines denote the mean GC content of promoters (solid line) and the whole genome (dashed line). **b**, Violin plots (as defined in Figure 1) of promoter strength for 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 10 base window around the indicated position in the plant promoters. **d**, As (**b**) but for promoter strength in maize protoplasts.

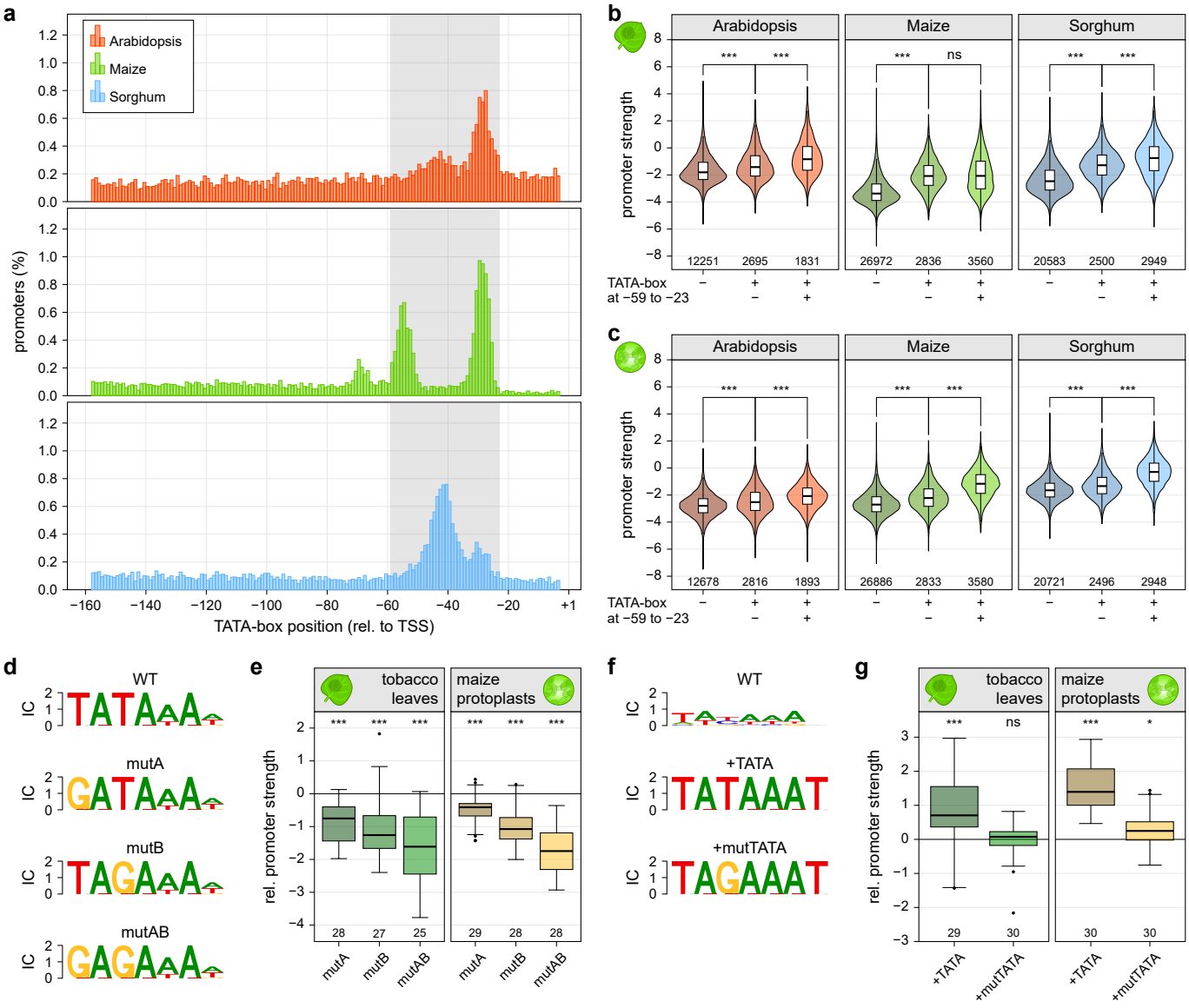


Fig. 3 | The TATA-box is a key determinant of promoter strength. a, Histograms showing the percentage of promoters with a TATA-box at the indicated position. The region between positions −59 and −23 in which most TATA-boxes reside is highlighted in gray. **b**,**c**, Violin plots (as defined in Figure 1) 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 (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 inserted into promoters with a 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 (center 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 Wilcoxon rank-sum test and are indicated: ⋆, $p \leq 0.01$; ⋆⋆, $p \leq 0.001$; ⋆⋆⋆, $p \leq 0.0001$; ns, not significant.

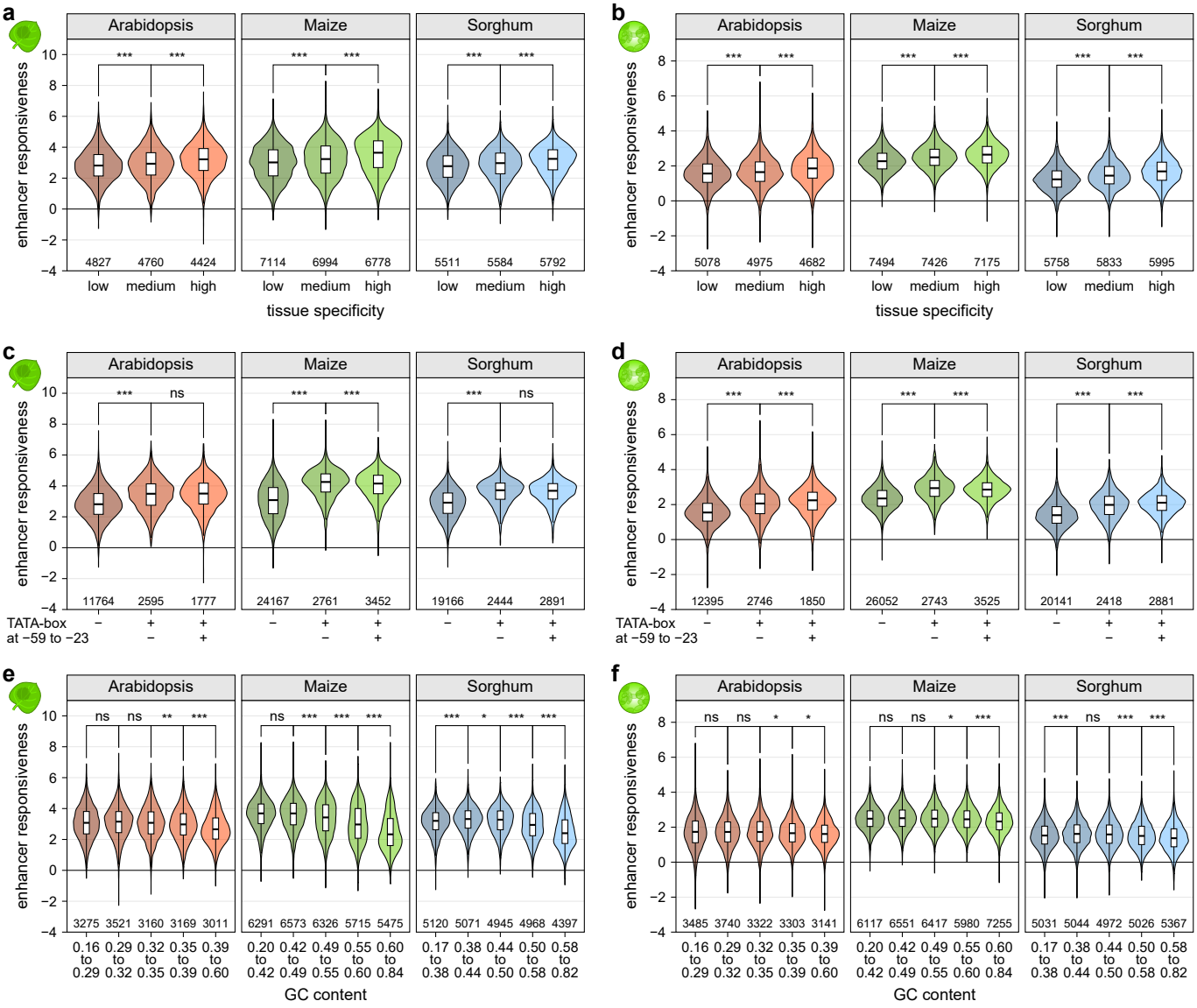


Fig. 4 | Enhancer responsiveness of promoters depends on the TATA-box and GC content. a,**b**, Violin plots (as defined in Figure 1) of enhancer responsiveness (promoter strength$^{\text{with enhancer}}$ − promoter strength$^{\text{without enhancer}}$) in tobacco leaves (**a**) or maize protoplasts (**b**). Promoters were grouped into three bins of approximately similar size according to the tissue-specificity *τ* (Yanai et al., 2005) 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.

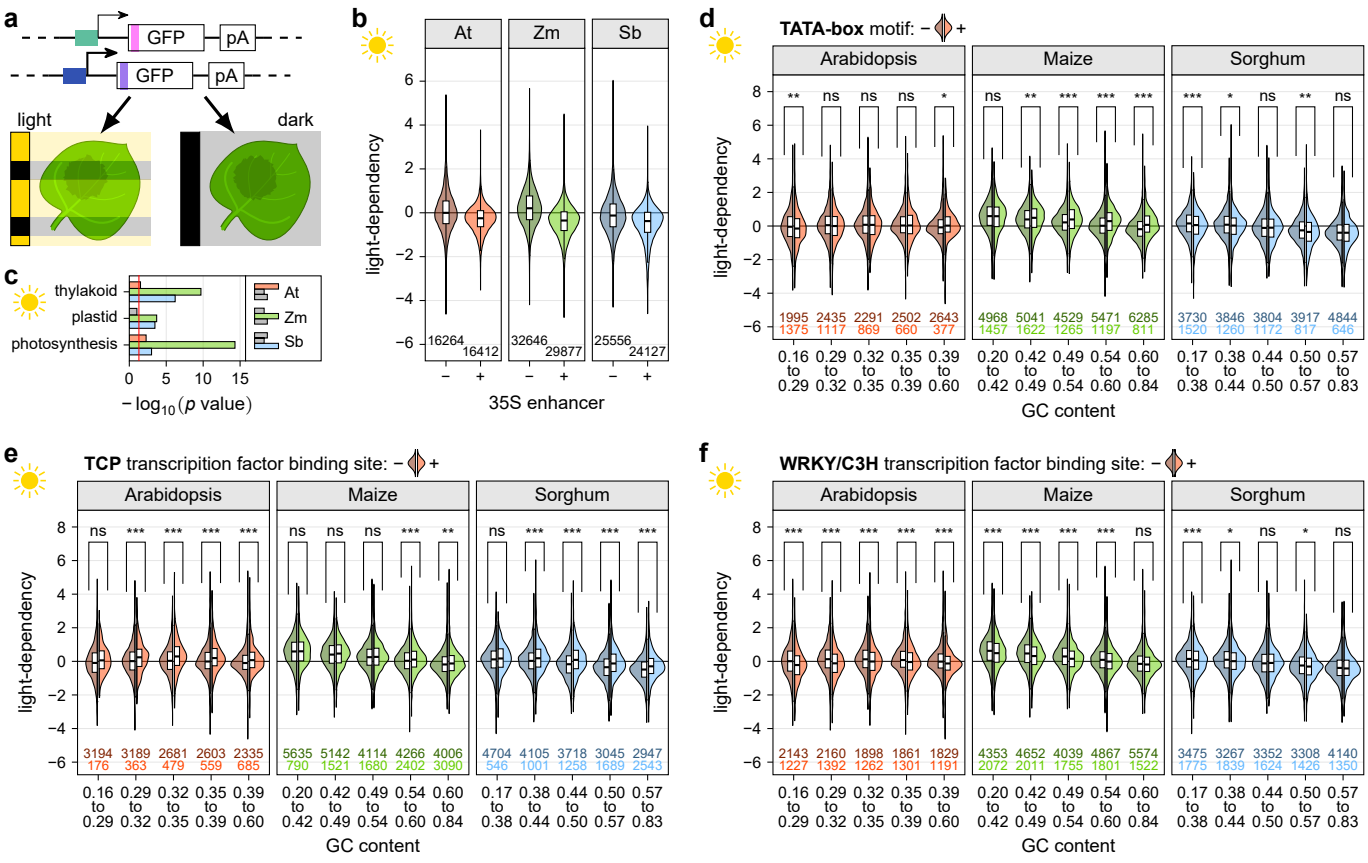


Fig. 5 | Promoter strength can be modulated by light. a, Tobacco leaves were transiently transformed with STARR-seq promoter libraries and the plants were kept for two days in 16h light/8h dark cycles (light) or completely in the dark (dark) prior to mRNA extraction. **b**, Violin plots (as defined in Figure 1) of light-dependency (promoter strength^{light} − 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 1000 most light-dependent promoters. The red line marks the significance threshold (adjusted *p* value ≤ 0.05). Non-significant bars are gray. **d**-**f**, Violin plots of light-dependency. Promoters are grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a TATA-box (**d**), or a binding site for TCP (**e**) or WRKY (**f**) transcription factors.

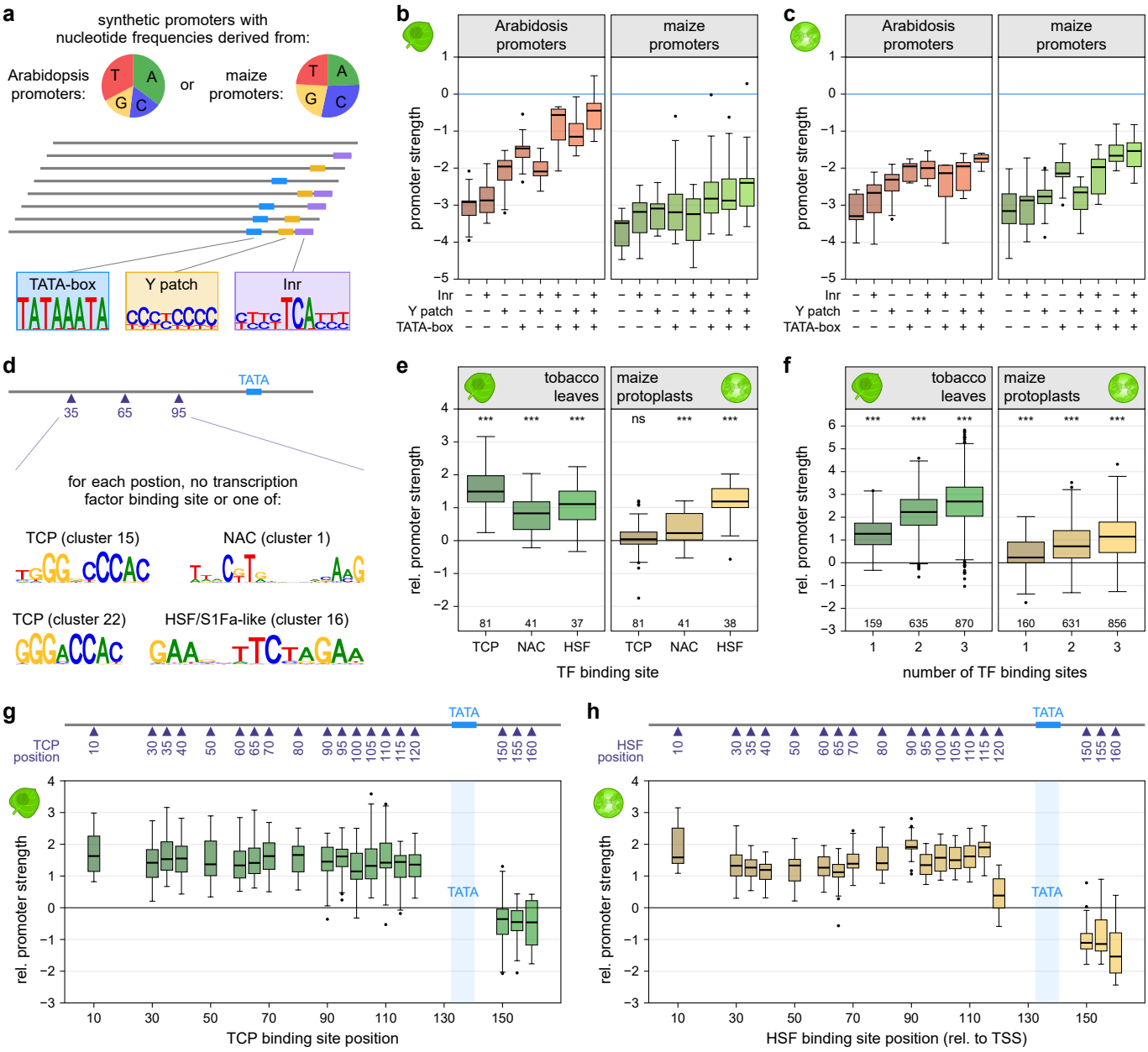


Fig. 6 | Design and validation of synthetic promoters. a-c, Synthetic promoters with nucleotide frequencies similar to an average Arabidopsis (35.2% A, 16.6% C, 15.3% G, 32.8% T) or maize (24.5% A, 29.0% C, 22.5% G, 23.9% T) promoter were created and modified by adding a TATA-box, Y patch, and/or Inr 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. **d-f**, Transcription factor 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 transcription factor (**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**) transcription factor 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 are as defined in Figure 3. In (**e-h**), the corresponding promoter without any transcription factor binding was set to 0 (horizontal black line).

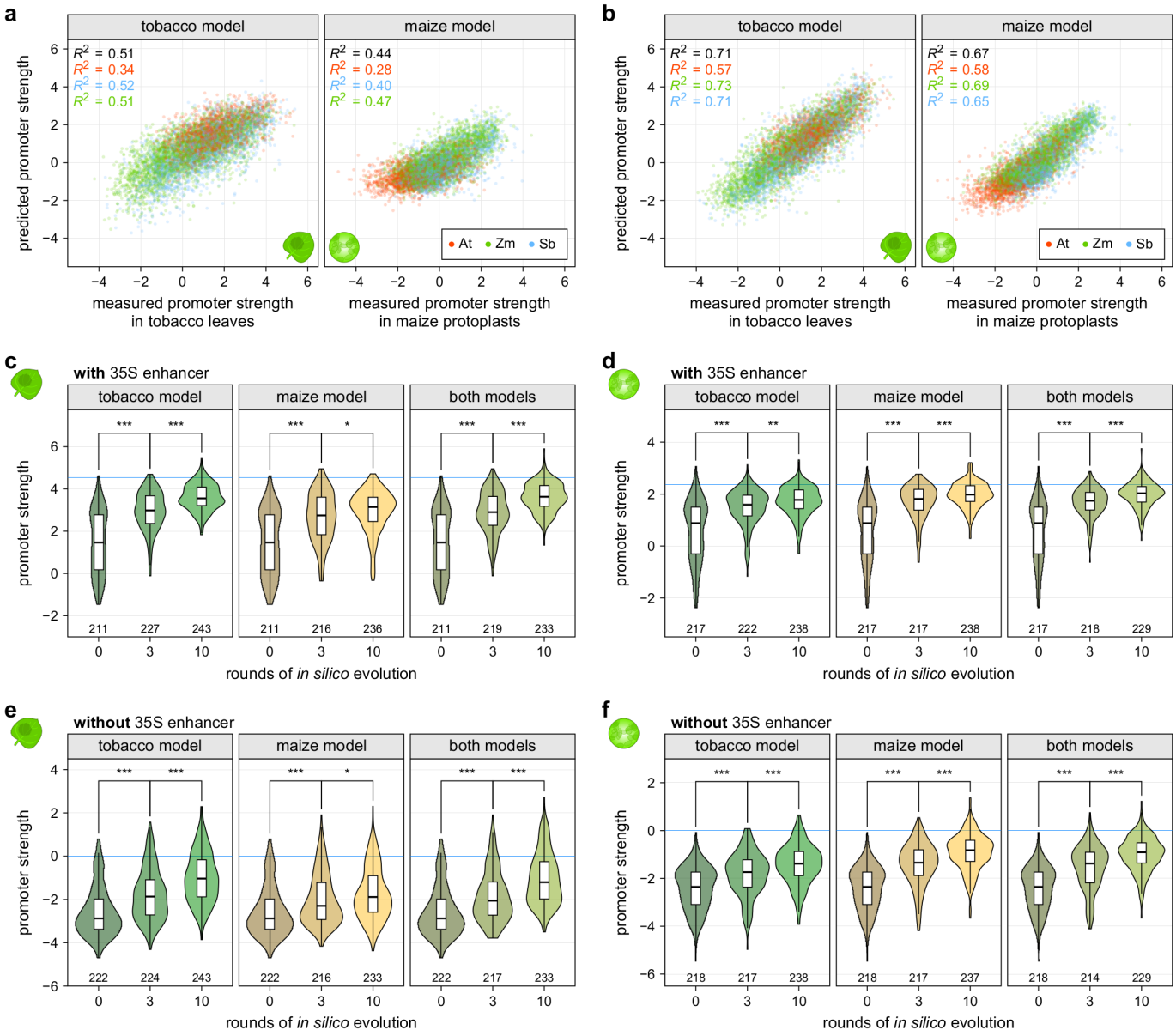
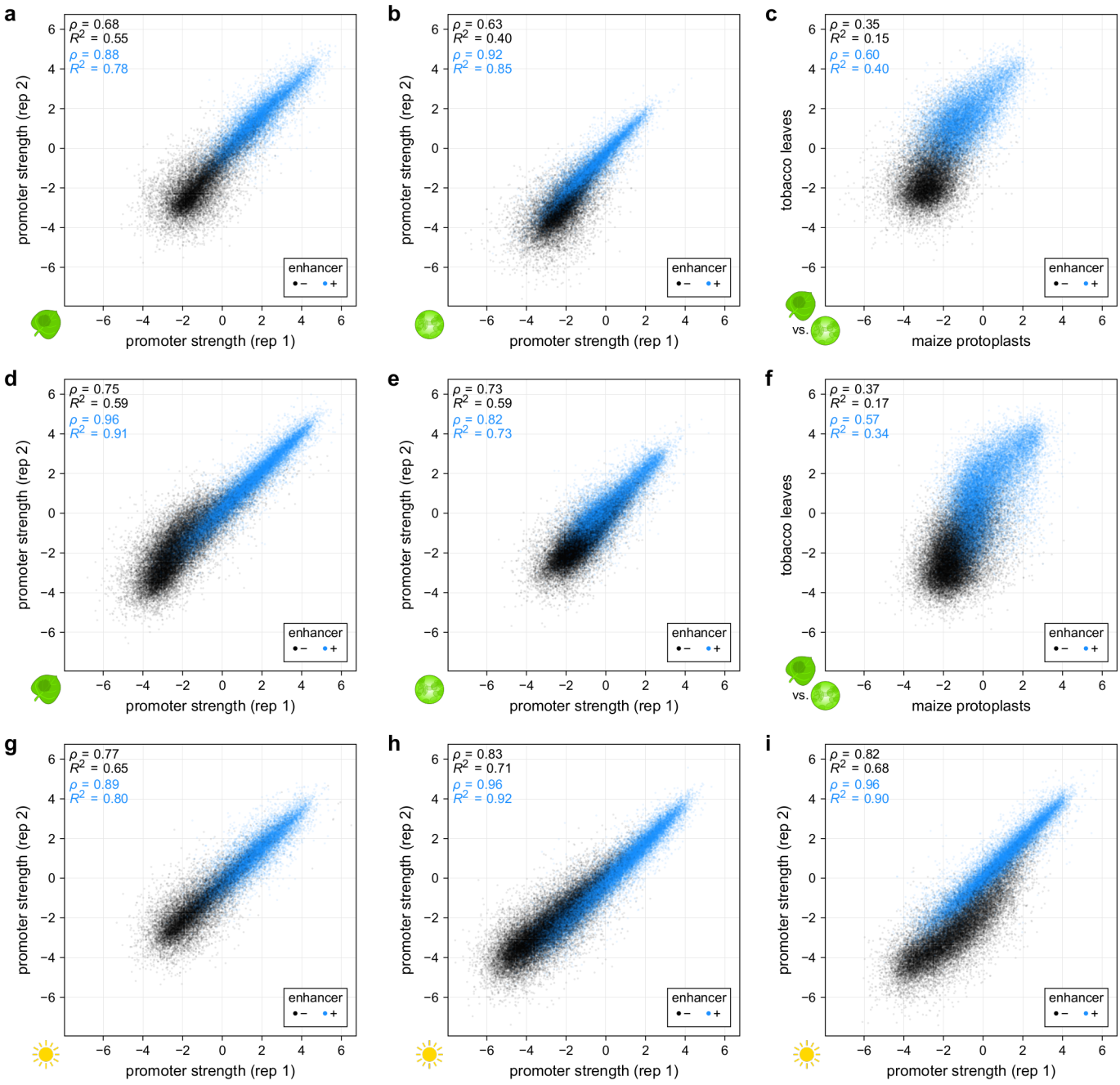
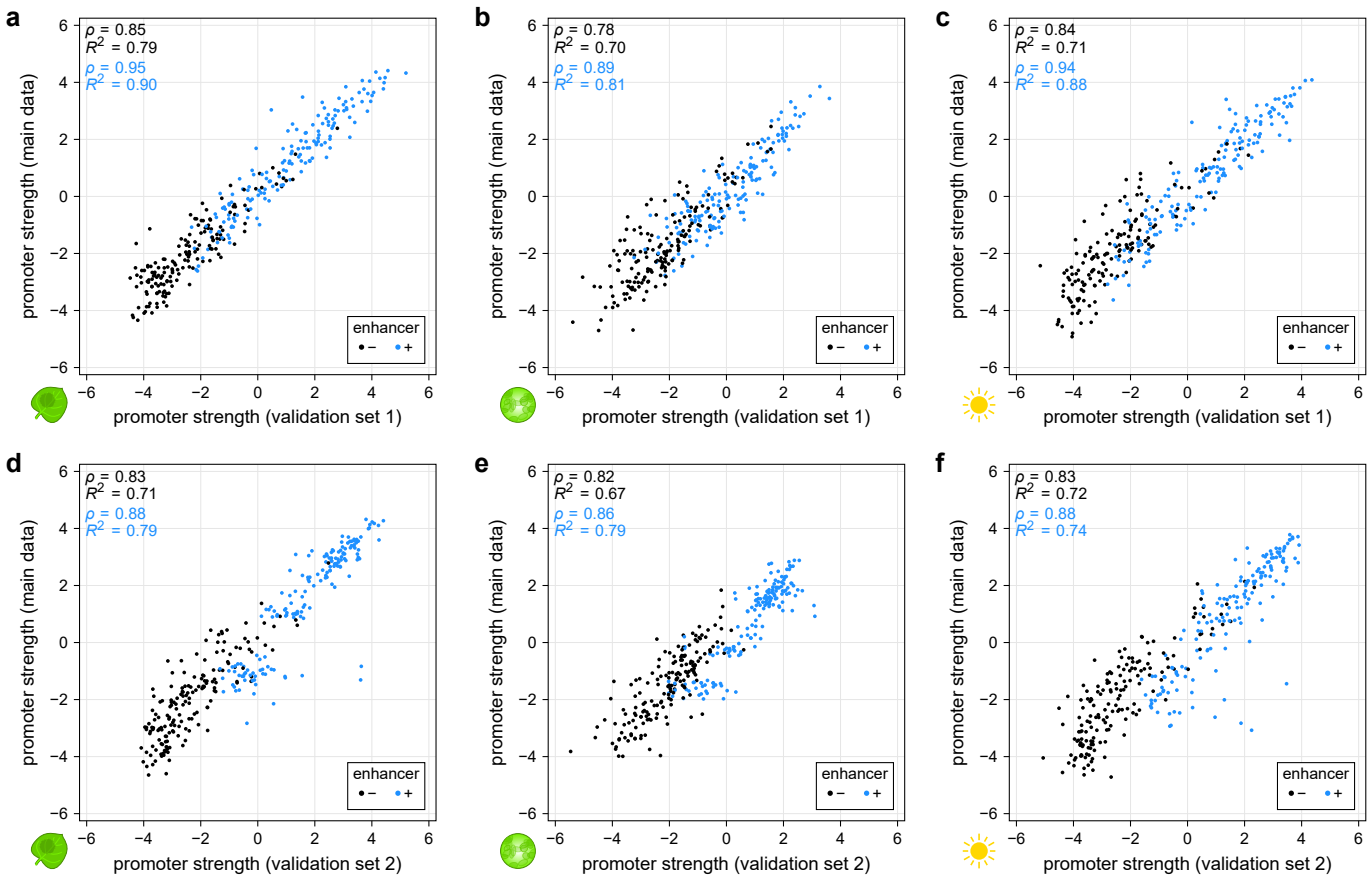


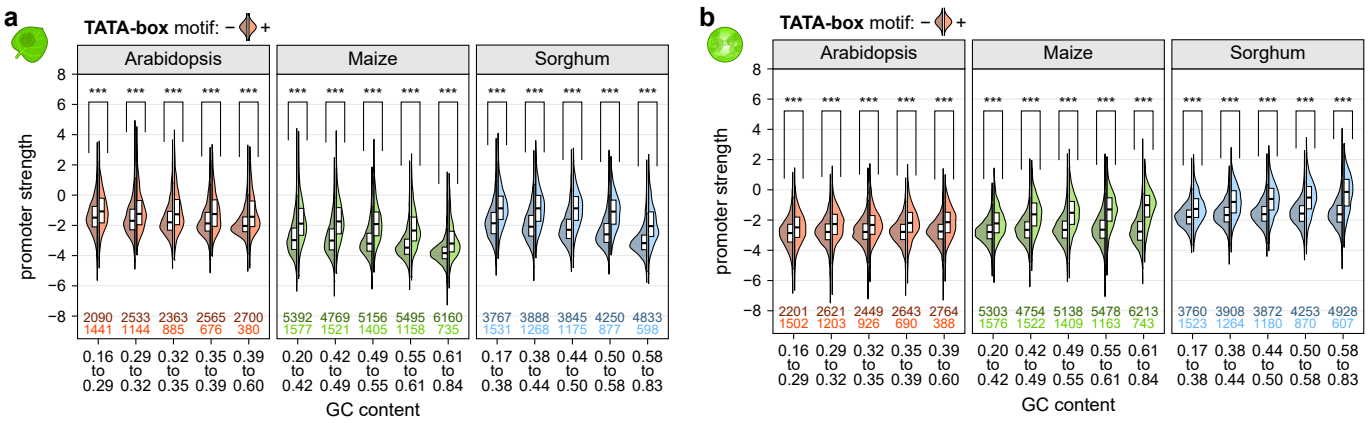
Fig. 7 | Computational models can predict promoter strength and enable *in silico* evolution of plant promoters. a, Correlation between the promoter strength as determined by STARR-seq using promoter libraries with the 35S enhancer and predictions from a linear model based on the GC content and motif scores for core promoter elements and transcription factors. 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 colored as indicated (inset). Correlations are shown for a test set of 10% of all promoters. **b**, Similar to (**a**) but the prediction is based on a convolutional neural network trained on promoter sequences. **c-f**, Violin plots (as defined in Figure 1) 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 model(s) used for the *in silico* evolution is indicated on each plot. The promoter strength of the 35S promoter is indicated by a horizontal blue line.



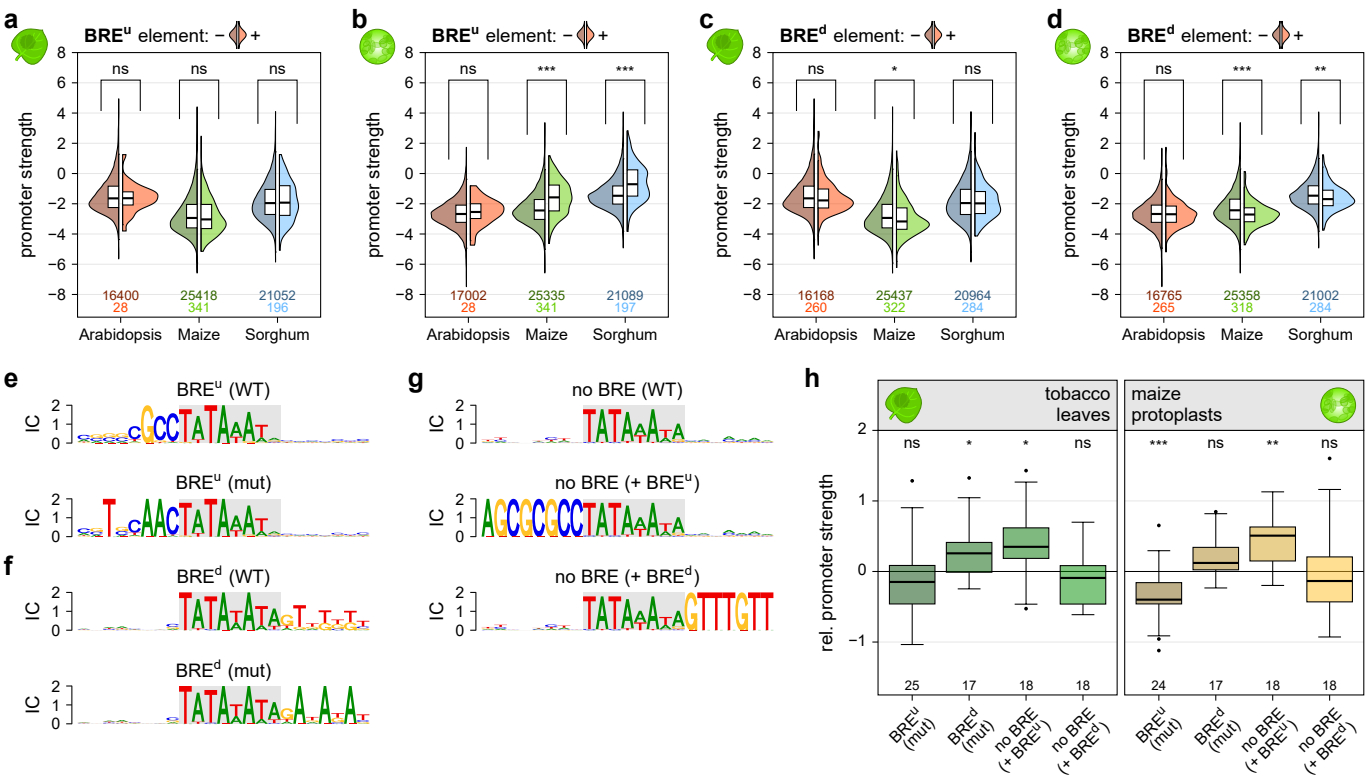
Supplementary Fig. 1 | The promoter STARR-seq assay is highly reproducible but promoter strength depends on the assay system. a,**b**, Correlation of two biological replicates of STARR-seq using the Arabidopsis promoter libraries in tobacco leaves (**a**) or in maize protoplasts (**b**). **c**, Comparison of the strength of Arabidopsis promoters in tobacco leaves and maize protoplasts. **d**,**e**, Correlation of two biological replicates of STARR-seq using the sorghum promoter libraries in tobacco leaves (**d**) or in maize protoplasts (**e**). **f**, Comparison of the strength of sorghum promoters in tobacco leaves and maize protoplasts. **g-i**, Correlation of two biological replicates of STARR-seq using the Arabidopsis (**g**), maize (**h**), or sorghum (**i**) promoter libraries in tobacco leaves that were kept for two days in 16h light/8h dark cycles prior to mRNA extraction.



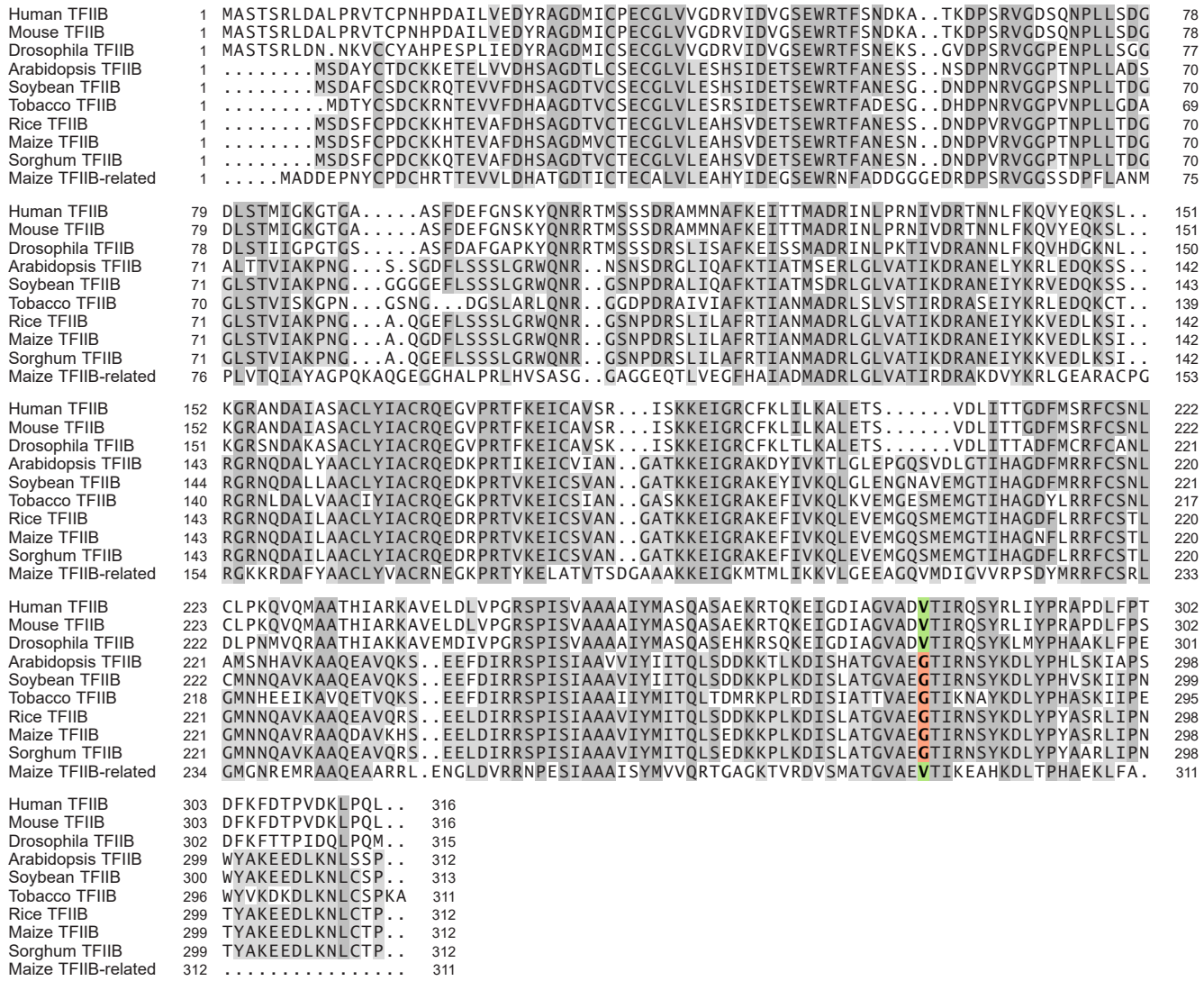
Supplementary Fig. 2 | Promoter strength in small validation libraries correlates highly with comprehensive data. a-c, Correlation between the strength of promoters present in the comprehensive promoter libraries (main data) and in a separate, smaller validation library. The promoter strength was determined in tobacco leaves (**a**) and maize protoplasts (**b**) that were kept in the dark prior to mRNA extraction. Additionally, promoter strength was measured in tobacco leaves that were kept for two days in 16h light/8h dark cycles prior to mRNA extraction (**c**). **d-f**, As in (**a-c**) but for a second validation library.



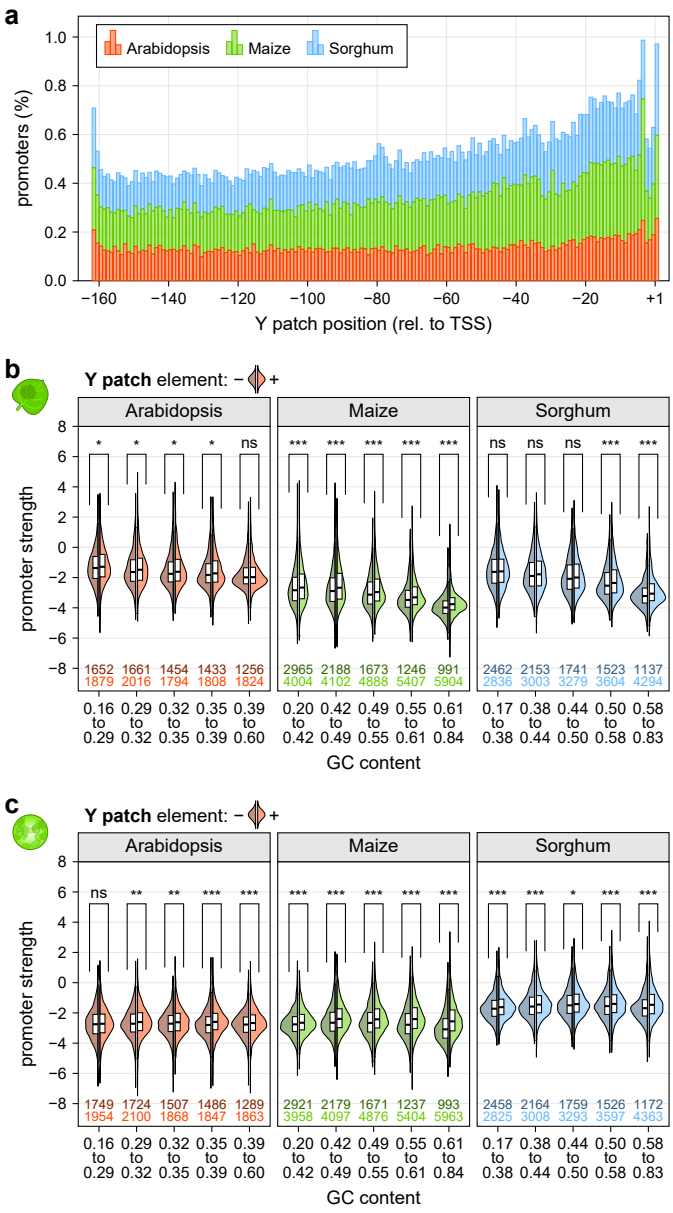
Supplementary Fig. 3 | The effect of the TATA-box on promoter strength is not a result of decreased GC content. a,**b**, Violin plots of promoter strength in tobacco leaves (**a**) or maize protoplasts (**b**). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a TATA-box. Violin plots are as defined in Figure 1, except only one half is shown.



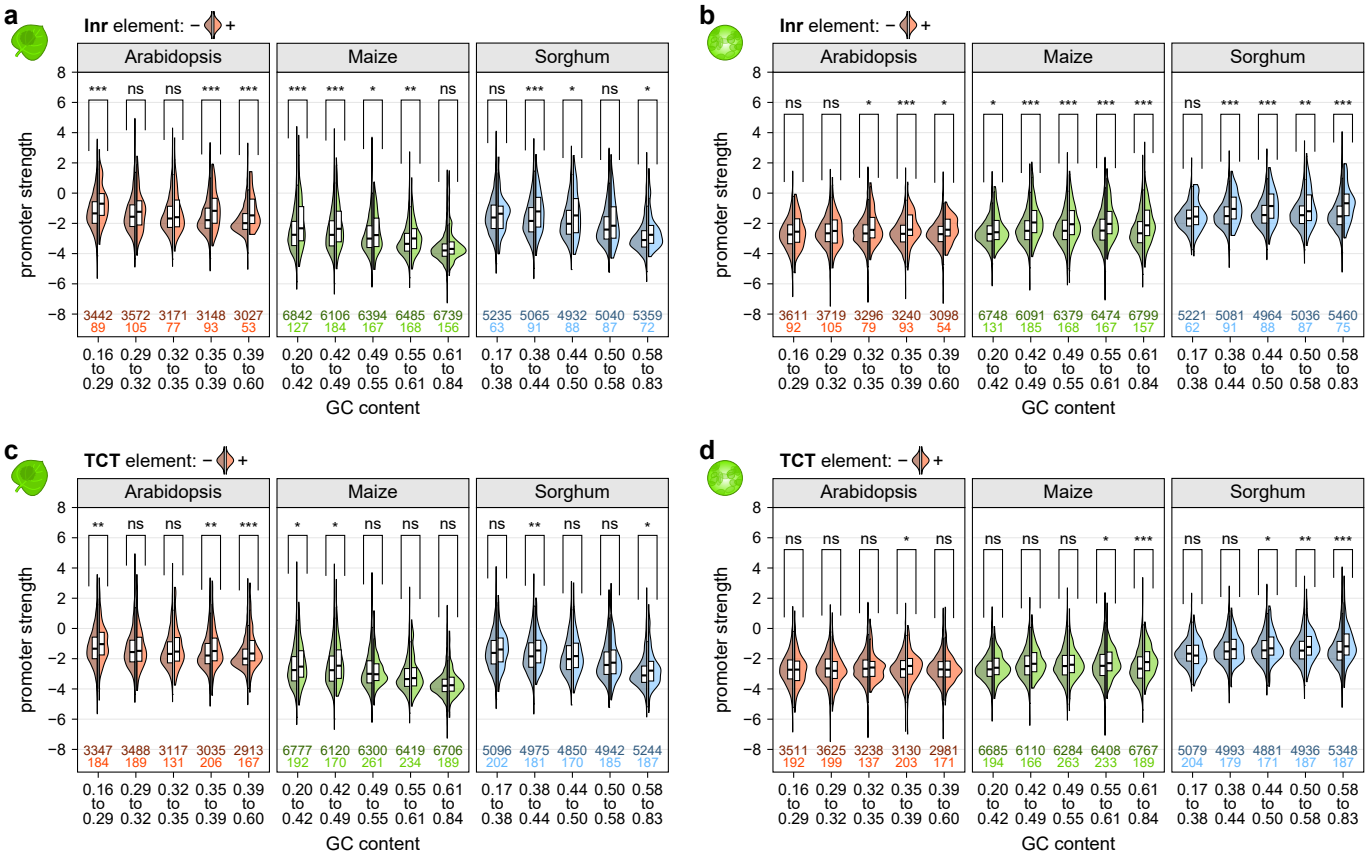
Supplementary Fig. 4 | The BRE^u element is most active in maize protoplasts. a-d, Violin plots of promoter strength in tobacco leaves (**a,c**) or maize protoplasts (**b,d**). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a BRE^u (**a,b**), or BRE^d (**c,d**) element. Violin plots are as defined in Figure 1, except only one half is shown. **e**,**g**, 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 (as defined in Figure 3) for the relative strength of the promoter variants shown in (**e-g**). The corresponding WT promoter was set to 0 (horizontal black line).



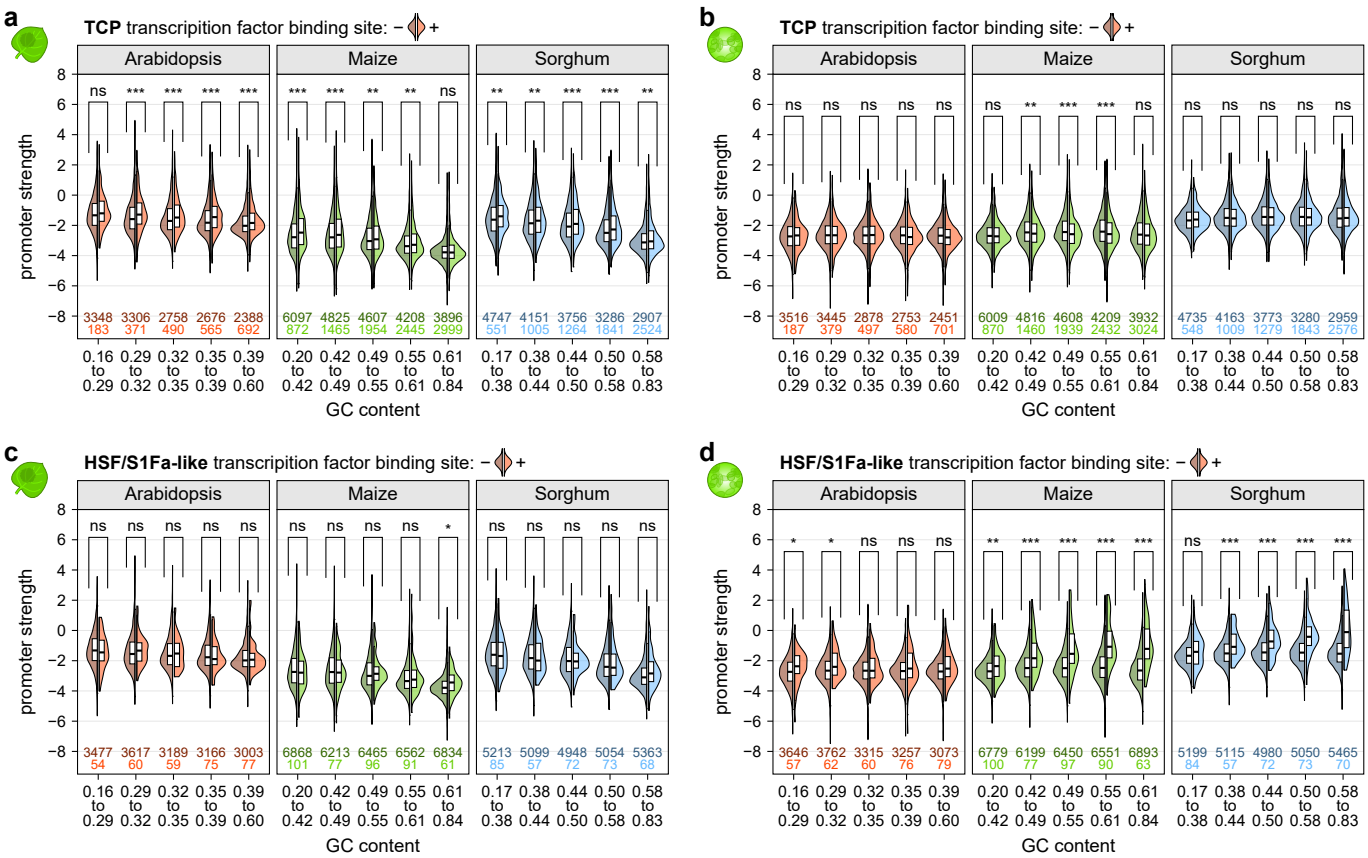
Supplementary Fig. 5 | The maize genome encodes a TFIIB-related protein with a conserved valine residue required for BRE^{u} recognition. Alignment of TFIIB and TFIIB-like protein sequences from indicated species. Residues conserved in 80 or 50% of the sequences are highlighted in dark or light gray, respectively. The valine residue required for recognition of BRE^{u} is highlighted in green.



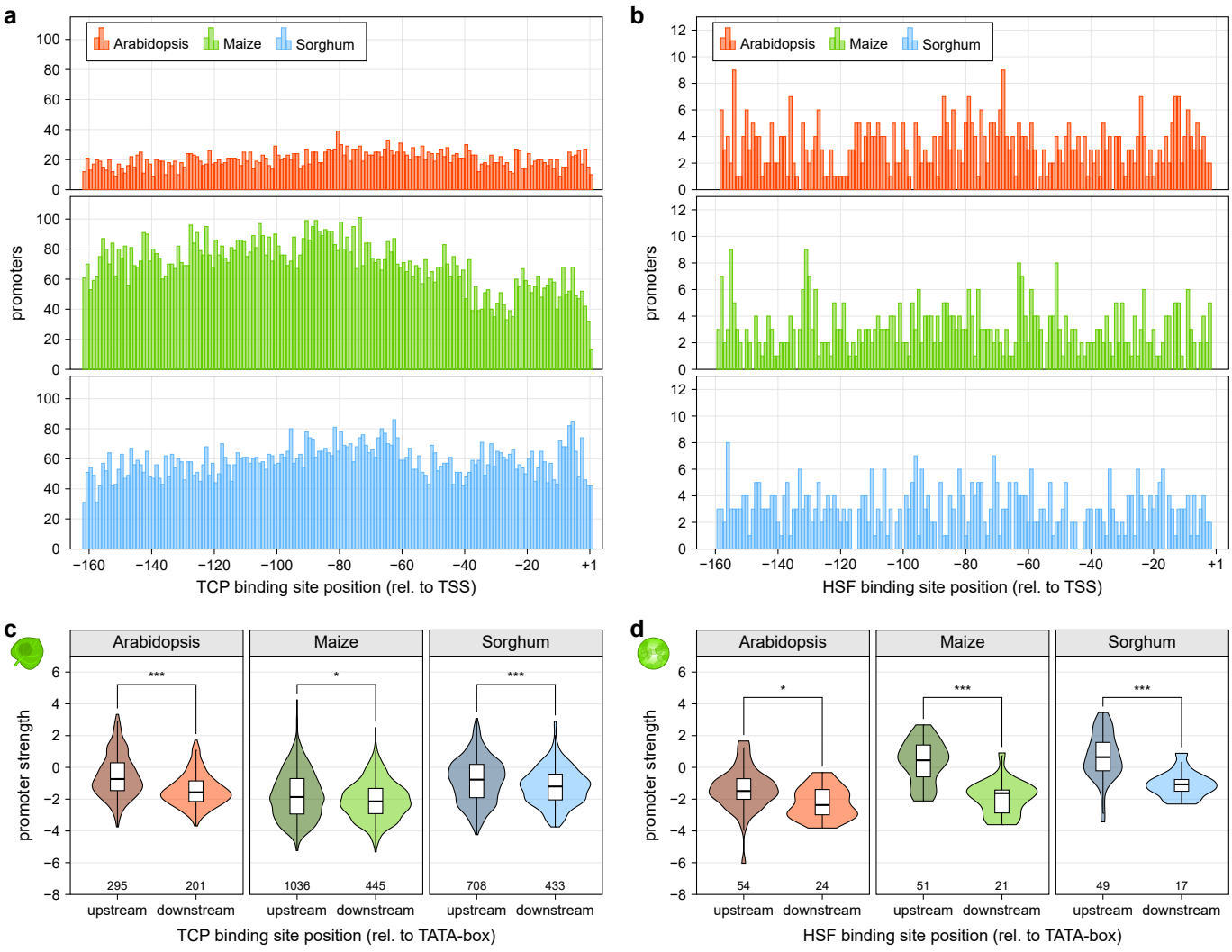
Supplementary Fig. 6 | The Y patch is a plant-specific core promoter element. a, Histogram showing the percentage of promoters with a TATA-box at the indicated position. **b**,**c**, Violin plots of promoter strength in tobacco leaves (**b**) or maize protoplasts (**c**). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a Y patch. Violin plots are as defined in Figure 1, except only one half is shown.



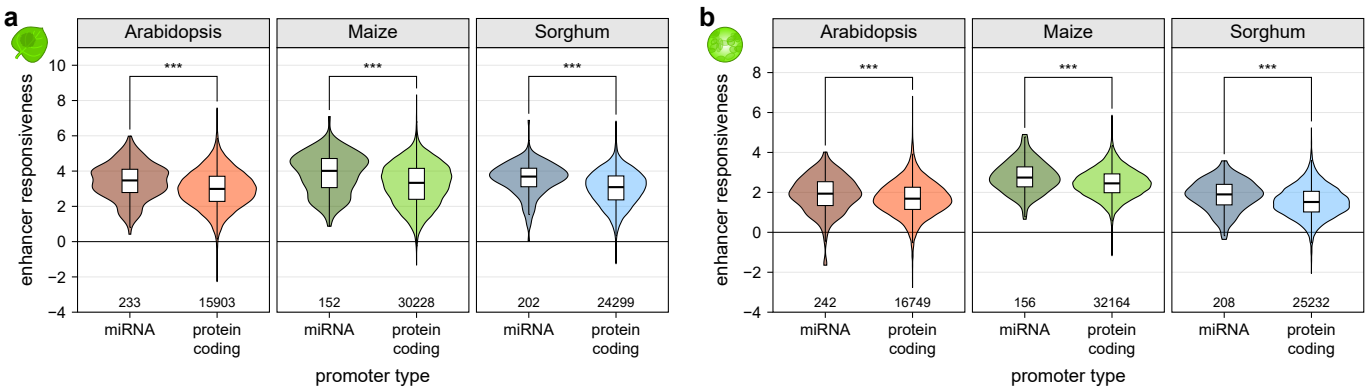
Supplementary Fig. 7 | Core promoter elements at the TSS influence promoter strength. a-d, Violin plots of promoter strength in tobacco leaves (**a,c**) or maize protoplasts (**b,d**). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) an Inr (**a,b**), or TCT (**c,d**) element at the TSS. Violin plots are as defined in Figure 1, except only one half is shown.



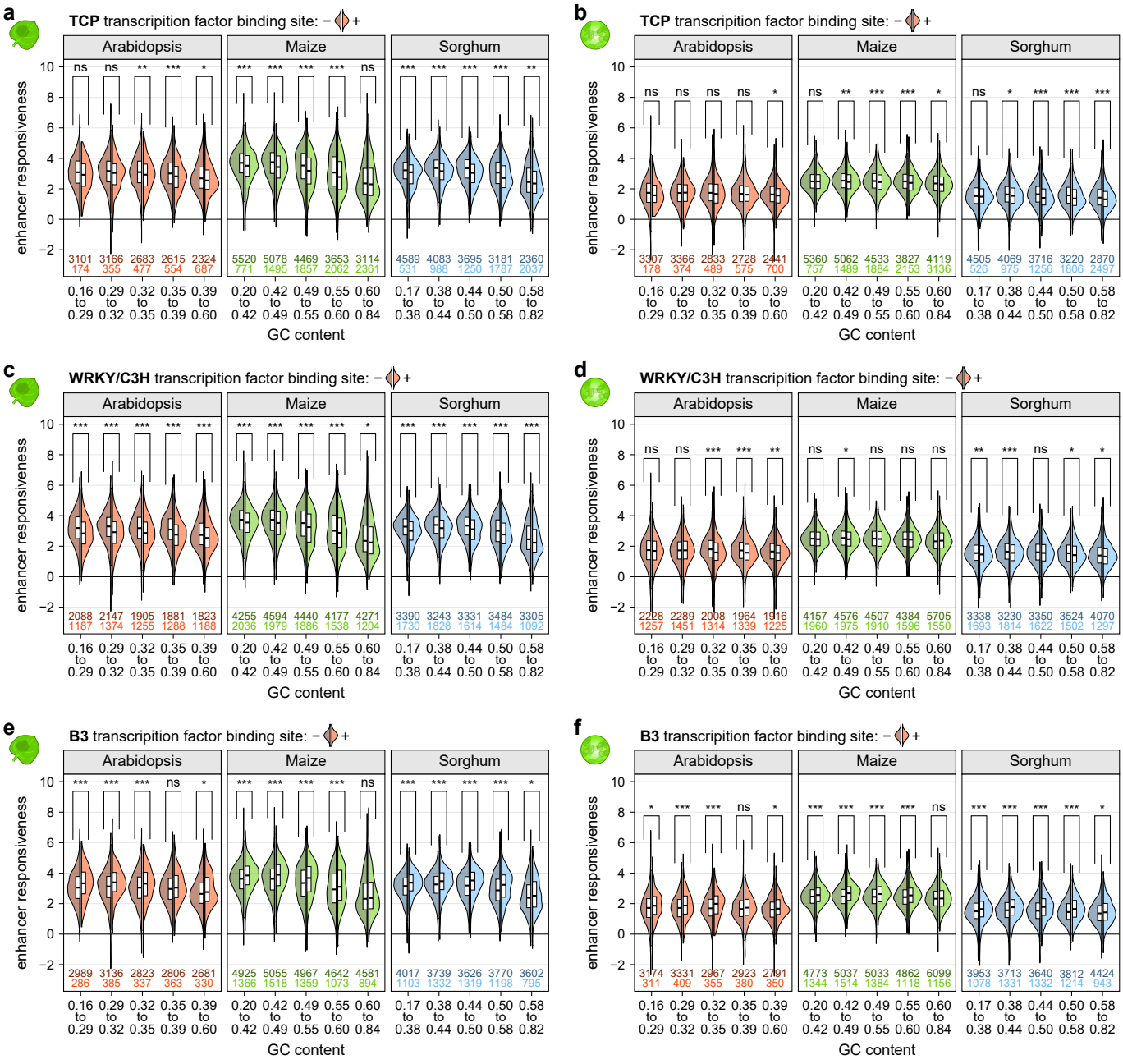
Supplementary Fig. 8 | Transcription factor binding sites contribute to promoter strength in an assay system-dependent manner. a-d, Violin plots of promoter strength for libraries without enhancer in tobacco leaves (**a,c**) or maize protoplasts (**b,d**). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a binding site for TCP (**a,b**) or HSF (**c,d**) transcription factors. Violin plots are as defined in Figure 1, except only one half is shown.



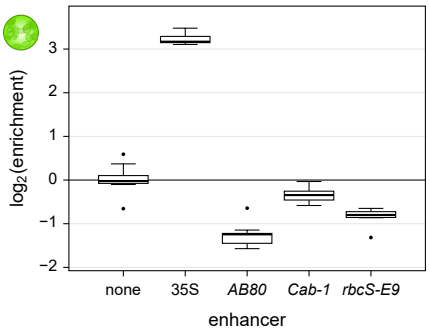
Supplementary Fig. 9 | Transcription factor binding sites are more active upstream of the TATA-box. a,**b**, Histograms showing the number of promoters with a TCP (**a**) or HSF (**b**) transcription factor binding site at the indicated position. **c-f**, Violin plots (as defined in Figure 1) of promoter strength for libraries without enhancer in tobacco leaves (**c**,**e**) or maize protoplasts (**d**,**f**). Promoters were grouped by the position of their TCP (**c**,**d**), or HSF (**e**,**f**) transcription factor binding site relative to the TATA-box.



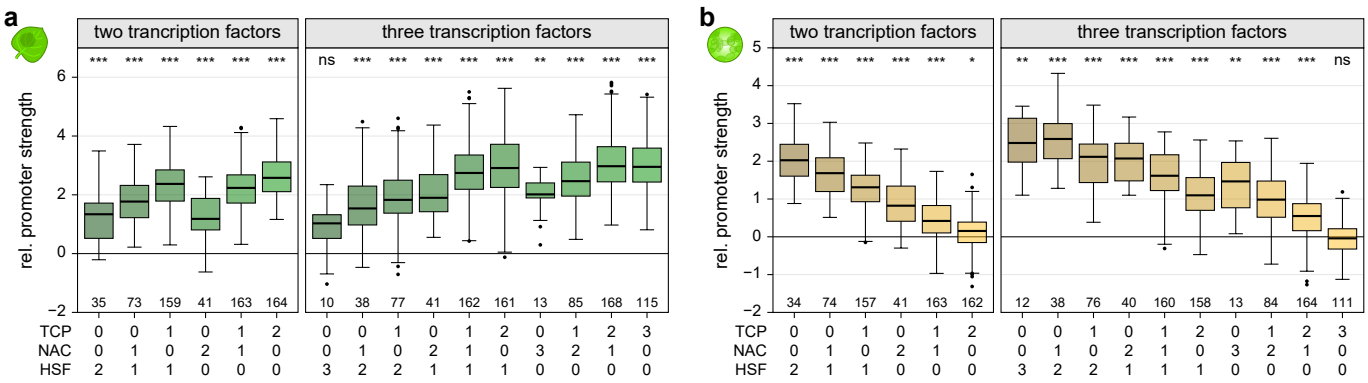
Supplementary Fig. 10 | Promoters of miRNA genes are more responsive to the 35S enhancer than those associated with protein-coding genes. a,**b**, Violin plots (as defined in Figure 1) of enhancer responsiveness (promoter strength$^{\text{with enhancer}}$ – promoter strength$^{\text{without enhancer}}$) in tobacco leaves (**a**) or maize protoplasts (**b**). Promoters associated with miRNA or protein-coding genes are compared.



Supplementary Fig. 11 | 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 color) or with (right half, lighter color) a TCP (**a,b**), WRKY (**c,d**), or B3 (**e,f**) transcription factor binding site. Violin plots are as defined in Figure 1, except only one half is shown.



Supplementary Fig. 12 | Light-responsive plant enhancers are not active in maize protoplasts. Constructs harboring no enhancer (none), a 35S enhancer, or one of three light-responsive plant enhancers (*AB80*, *Cab-1*, or *rbcS-E9*) upstream of the 35S minimal promoter were subjected to STARR-seq in maize protoplasts generated from dark-grown plants (Jores et al., 2020). Each boxplot (center line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; points, outliers) denotes the enrichment of all recovered mRNA barcodes over the DNA input. Only one experiment was performed.



Supplementary Fig. 13 | Transcription factor binding sites affect promoter strength additively. a,**b**, Boxplots (as defined in Figure 3) of promoter strength for libraries without enhancer in tobacco leaves (**a**) or maize protoplasts (**b**) for synthetic promoters with the indicated numbers of binding sites for TCP, NAC, and HSF transcription factors. The corresponding promoter without any transcription factor binding site was set to 0 (horizontal black line).